



UMA NOVA ABORDAGEM PARA A CRIAÇÃO DE PERFIS DE CLIENTES  
RENTÁVEIS UTILIZANDO *MACHINE LEARNING* EM AMBIENTE DE *CLOUD*  
*COMPUTING*

Leandro da Silva Carvalho

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Civil, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Civil.

Orientador: Nelson Francisco Favilla Ebecken

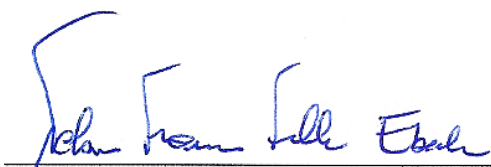
Rio de Janeiro  
Agosto de 2018


UMA NOVA ABORDAGEM PARA A CRIAÇÃO DE PERFIS DE CLIENTES  
RENTÁVEIS UTILIZANDO *MACHINE LEARNING* EM AMBIENTE DE *CLOUD*  
*COMPUTING*

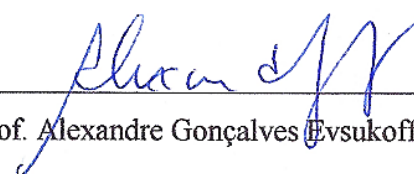
Leandro da Silva Carvalho

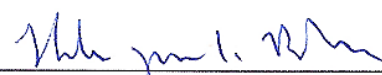
TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ  
COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA  
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS  
REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM  
CIÊNCIAS EM ENGENHARIA CIVIL.


Examinada por:

  
Prof. Nelson Francisco Favilla Ebecken, D.Sc.

  
Prof<sup>a</sup>. Beatriz de Souza Leite Pires de Lima, D.Sc.

  
Prof. Alexandre Gonçalves Evsukoff, Dr.

  
Prof. Helio José Corrêa Barbosa, D.Sc.

  
Prof. Elton Fernandes, Ph.D.

RIO DE JANEIRO, RJ – BRASIL

AGOSTO DE 2018

Carvalho, Leandro da Silva

Uma Nova Abordagem para a Criação de Perfis de Clientes Rentáveis Utilizando *Machine Learning* em Ambiente de *Cloud Computing* / Leandro da Silva Carvalho. – Rio de Janeiro: UFRJ/COPPE, 2018.

XVI, 194 p.: il., 29,7cm

Orientador: Nelson Francisco Favilla Ebecken

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia Civil, 2018.

Referências Bibliográficas: p. 190-194.

1. Modelos RFMP. 2. *Machine Learning*. 3. *Cloud Computing*. 4. Modelos RFM. 5. Comportamento do Consumidor. 6. Classificação não supervisionada. I. Ebecken, Nelson Francisco Favilla II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Título.

# Agradecimentos

Definitivamente não cheguei até aqui sozinho e nem haveria como. Muito menos deixei de ser abençoado pela sorte, apesar de existirem diversas maneiras para maximizar a probabilidade de se encontrá-la. Por isso, deixo registrado os meus mais profundos e verdadeiros agradecimentos àqueles que me acompanharam, de um jeito ou de outro, até aqui:

Aos meus avós Nedes e Atualpina (*in-memorian*), por terem fornecido o pilar central que nortearam a formação do caráter e a base familiar.

À minha mãe Neiva, por dedicar toda sua vida e amor à criação dos dois filhos.

Ao meu irmão Leonardo, por ser esse eterno amigo fiel e seguro, que com amor continua a me apoiar, acompanhar e dividir os desafios da vida.

À minha amada companheira Lilian, por toda dedicação, compreensão e amor.

Aos meus familiares e amigos que sempre estiveram comigo, ao meu lado ou em suas intenções e preces.

Ao prof. Nelson Ebecken, por toda sua sabedoria e simplicidade, que me motivou e acreditou, com toda liberdade e confiança, enquanto eu duvidava.

À NetCommerce, por ser essa ideia fixa.

Por fim, permaneço agradecendo ao que continua a permitir saúde e paz, não só a mim, mas em especial a todos os meus familiares e amigos. Afinal, hoje sou um homem da ciência; mas garanto, nunca deixei de ser um homem de fé.

*Tenho a vida doida  
Encabeço o mundo  
Sou ariano torto  
Vivo de amor profundo  
Sou perecível ao tempo  
Vivo por um segundo  
Perdoa meu amor  
Esse nobre vagabundo*

Marcelo Mello – Nobre Vagabundo

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

UMA NOVA ABORDAGEM PARA A CRIAÇÃO DE PERFIS DE CLIENTES  
RENTÁVEIS UTILIZANDO *MACHINE LEARNING* EM AMBIENTE DE *CLOUD  
COMPUTING*

Leandro da Silva Carvalho

Agosto/2018

Orientador: Nelson Francisco Favilla Ebecken

Programa: Engenharia Civil

Este trabalho tem por objetivo apresentar a criação de um novo modelo, denominado RFMP (*Recency – Frequency – Monetary – Profitability*), como alternativa ao tradicional modelo RFM. O propósito é avaliar se a inclusão de um novo parâmetro P (da palavra inglesa *Profitability*), associado à lucratividade dos consumidores, trará algum impacto na segmentação de uma base de clientes de um site de *e-commerce*.

Adicionalmente, uma nova metodologia de classificação foi proposta – em substituição à forma tradicional utilizada – para aumentar o grau de confiança dos modelos gerados e melhorar a assertividade entre os *clusters* e seus respectivos conteúdos. Além disso, foi recomendada também a criação de três novos índices de mensuração para suprir a ausência de indicadores capazes de determinar a consistência e a qualidade dos *clusters* e modelos produzidos. Tudo isso foi realizado através de um estudo empírico, feito com o uso de uma plataforma de aprendizado de máquina a partir de um ambiente computacional em nuvem. Por fim, com os resultados obtidos, foi possível evidenciar que houve um impacto direto na formação dos *clusters* gerados, pois os grupos de clientes foram diferenciados não só pelo valor monetário, mas como também a partir das suas respectivas rentabilidades, o que permitiu constatar que nem sempre os clientes com os maiores valores monetários eram de fato os mais lucrativos.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

A NEW APPROACH TO CREATING PROFITABLE CUSTOMER PROFILES  
USING MACHINE LEARNING IN THE CLOUD COMPUTING ENVIRONMENT

Leandro da Silva Carvalho

August/2018

Advisor: Nelson Francisco Favilla Ebecken

Department: Civil Engineer

This work aims to present the creation of a new model, called RFMP (Recency – Frequency – Monetary – Profitability), as an alternative to the traditional RFM model to evaluate whether the inclusion of a new parameter P – associated with customer profitability – will have any impact on targeting a client database from an e-commerce site.

Additionally, a new classification methodology was proposed to increase the confidence level of the generated models and improve the assertiveness between the clusters and their respective contents, instead of the traditional form used. Besides, it was also recommended to create three new measurement indices in order to overcome the lack of existing indicators capable of determining the consistency and quality of the clusters and models produced. All of this was accomplished through an empirical study, done using a machine learning platform from a cloud computing environment. Finally, with the results obtained, it was possible to demonstrate that there was a direct impact on the formation of the clusters produced, since the customer groups were differentiated not only by the monetary value, but also from their respective profitability, which allowed to determine that customers with the highest monetary values were not always the most profitable.

# Índice

Agradecimentos .....	iv
Índice .....	viii
Lista de Figuras .....	xi
Lista de Tabelas .....	xiv
1 Introdução.....	1
1.1 Objetivos.....	2
1.2 Metodologia.....	2
1.3 Organização do Trabalho.....	3
2 Fundamentação Teórica.....	4
2.1 Conceitos Básicos.....	4
2.1.1 <i>E-Commerce</i> .....	4
2.1.2 Modelo RFM ( <i>Recency – Frequency – Monetary</i> ).....	5
2.1.3 Computação em Nuvem .....	7
2.1.4 <i>Analytics-as-a-Services (AaaS)</i> .....	9
2.1.5 <i>Azure Machine Learning (Azure ML)</i> .....	10
2.2 Estado da Arte .....	11
2.2.1 Pesquisas Relacionadas .....	11
2.2.2 Revisão da Literatura.....	19
2.2.3 Contribuição .....	25
3 Conjunto de Dados .....	28
3.1 Apresentação .....	28
3.2 Estatísticas Básicas – Dados Originais .....	28
4 Desenvolvimento da Tese.....	36
4.1 Processo de Desenvolvimento .....	36
4.2 Processo de Avaliações dos Modelos .....	44
4.3 Métricas de Avaliações dos Modelos .....	47
5 Desenvolvimento dos Modelos .....	49
5.1 Análise RFM.....	49
5.1.1 <i>Simplified Silhouette</i> .....	49
5.1.2 <i>Davies-Bouldin</i> .....	53
5.1.3 <i>Dunn</i> .....	59

5.1.4	<i>Average Deviation</i> .....	60
5.1.5	Síntese dos Resultados RFM Encontrados .....	67
5.1.6	Tabelas Auxiliares – RFM.....	70
5.2	Análise RFMP .....	72
5.2.1	<i>Simplified Silhouette</i> .....	72
5.2.2	<i>Davies-Bouldin</i> .....	74
5.2.3	<i>Dunn</i> .....	79
5.2.4	<i>Average Deviation</i> .....	83
5.2.5	Síntese dos Resultados – RFMP .....	90
5.2.6	Tabelas Auxiliares – RFMP .....	92
5.3	Análise RFM (Z-SCORE) .....	95
5.3.1	<i>Simplified Silhouette</i> .....	95
5.3.2	<i>Davies-Bouldin</i> .....	97
5.3.3	<i>Dunn</i> .....	98
5.3.4	<i>Average Deviation</i> .....	99
5.3.5	Síntese dos Resultados – RFM ( <i>Z-Score</i> ).....	106
5.3.6	Tabelas Auxiliares – RFM ( <i>Z-Score</i> ) .....	108
5.4	Análise RFMP (Z-SCORE) .....	110
5.4.1	<i>Simplified Silhouette</i> .....	110
5.4.2	<i>Davies-Bouldin</i> .....	112
5.4.3	<i>Dunn</i> .....	113
5.4.4	<i>Average Deviation</i> .....	115
5.4.5	Síntese dos Resultados.....	123
5.4.6	Tabelas Auxiliares – RFMP ( <i>Z-Score</i> ) .....	125
5.5	Resumo .....	127
5.5.1	Da Análise .....	127
5.5.2	Dos Resultados .....	129
6	Desafios Encontrados .....	133
6.1	Problemas Identificados .....	133
6.2	Soluções Recomendadas.....	134
6.3	Reorganização dos Dados.....	136
6.4	Definição das Faixas RFM/P .....	138
7	Modelos Recomendados.....	143
7.1	Modelo Sugerido – RFM.....	143

7.1.1	Análise .....	143
7.1.2	Síntese dos Resultados e Melhorias Recomendadas .....	152
7.1.3	Tabela Auxiliar – Modelo Sugerido RFM.....	154
7.2	Modelo Final – RFM .....	155
7.2.1	Análise .....	155
7.2.2	Tabela Auxiliar – Modelo Final RFM.....	156
7.3	Modelo Sugerido – RFMP.....	157
7.3.1	Análise .....	157
7.3.2	Síntese dos Resultados e Melhorias Recomendadas .....	164
7.3.3	Tabela Auxiliar – Modelo Sugerido RFMP .....	166
7.4	Modelo Final – RFMP.....	167
7.4.1	Análise .....	167
7.4.2	Tabela Auxiliar – Modelo Final RFMP.....	172
7.5	Resumo .....	173
7.5.1	Da Análise .....	173
7.5.2	Dos Resultados .....	174
8	Conclusões.....	178
8.1	Avaliação das Análises e dos Resultados .....	178
8.2	Sobre o Ambiente Tecnológico .....	184
8.3	Estudos Futuros .....	186
8.4	Considerações Finais .....	187
	Referências .....	190

# Lista de Figuras

Figura 2.1: Distribuição anual do número de publicações referentes aos termos “ <i>RFM Model</i> ” e “ <i>RFM Analysis</i> ” até dezembro/2017 – Fonte: <i>Scopus</i> . .....	11
Figura 2.2: Distribuição anual do número de publicações referentes aos termos “ <i>RFM Model</i> ” e “ <i>RFM Analysis</i> ” até dezembro/2017 – Fonte: <i>Web of Science</i> . .....	12
Figura 2.3: Número de publicações por área de domínio referentes aos termos “ <i>RFM Model</i> ” e “ <i>RFM Analysis</i> ” até dezembro/2017 – Fonte: <i>Scopus</i> . .....	12
Figura 2.4: Número de publicações por área de domínio referentes aos termos “ <i>RFM Model</i> ” e “ <i>RFM Analysis</i> ” até dezembro/2017 – Fonte: <i>Web of Science</i> . .....	12
Figura 2.5: Distribuição anual do número de publicações referentes aos termos “ <i>Machine Learning</i> ” e “ <i>Data Mining</i> ” até dezembro/2017 – Fonte: <i>Scopus</i> . .....	13
Figura 2.6: Distribuição anual do número de publicações referentes aos termos “ <i>Machine Learning</i> ” e “ <i>Data Mining</i> ” até dezembro/2017 – Fonte: <i>Web of Science</i> . .....	13
Figura 2.7: Distribuição anual relativa do número de consultas referentes aos termos “ <i>Machine Learning</i> ” e “ <i>Data Mining</i> ” nos últimos 5 anos – Fonte: <i>Google Trends</i> . ....	14
Figura 2.8: Distribuição anual do número de publicações referentes ao termo “ <i>Cloud Computing</i> ” até dezembro/2017 – Fonte: <i>Scopus</i> . .....	14
Figura 2.9: Distribuição anual do número de publicações referentes ao termo “ <i>Cloud Computing</i> ” até dezembro/2017 – Fonte: <i>Web of Science</i> . .....	15
Figura 2.10: Distribuição anual do número de publicações referentes aos termos “ <i>Cloud Computing</i> ” e “ <i>Data Mining</i> ” ou “ <i>Machine Learning</i> ” até dezembro/2017 – Fonte: <i>Scopus</i> . .....	15
Figura 2.11: Distribuição anual do número de publicações referentes aos termos “ <i>Cloud Computing</i> ” e “ <i>Data Mining</i> ” ou “ <i>Machine Learning</i> ” até dezembro/2017 – Fonte: <i>Web of Science</i> . .....	16
Figura 2.12: Distribuição anual do número de publicações referentes aos termos “ <i>RFM Model</i> ” ou “ <i>RFM Analysis</i> ” e “ <i>Data Mining</i> ” ou “ <i>Machine Learning</i> ” até dezembro/2017 – Fonte: <i>Scopus</i> . .....	16
Figura 2.13: Distribuição anual do número de publicações referentes aos termos “ <i>RFM Model</i> ” ou “ <i>RFM Analysis</i> ” e “ <i>Data Mining</i> ” ou “ <i>Machine Learning</i> ” até dezembro/2017 – Fonte: <i>Web of Science</i> . .....	17
Figura 2.14: Gráfico de tendências tecnológicas do <i>Gartner Group</i> 2014. ....	17
Figura 2.15: Gráfico de tendências tecnológicas do <i>Gartner Group</i> 2015. ....	18

Figura 2.16: Gráfico de tendências tecnológicas do <i>Gartner Group</i> 2016. ....	18
Figura 2.17: Gráfico de tendências tecnológicas do <i>Gartner Group</i> 2017. ....	19
Figura 3.1: Gráfico que contém a matriz de correlação entre as variáveis de entrada utilizadas pelo estudo.....	34
Figura 3.2: Gráfico de projeção entre as variáveis de entrada utilizadas pelo estudo. ....	35
Figura 4.1: Parâmetros de customização do algoritmo <i>k-means</i> no ambiente <i>Azure ML</i> . .....	40
Figura 4.2: Parâmetros de customização do recurso <i>Sweep Clustering</i> . ....	42
Figura 4.3: Processo de desenvolvimento e execução dos modelos.....	43
Figura 5.1: Gráfico da visualização dos <i>clusters</i> da 1º análise RFM com a métrica <i>Simplified Silhouette</i> . ....	50
Figura 5.2: Gráfico da visualização dos <i>clusters</i> da 1º análise RFM com a métrica <i>Davies-Bouldin</i> .....	55
Figura 5.3: Quantidade de registros por <i>cluster</i> da 1º análise RFM com a métrica <i>Average Deviation</i> . ....	61
Figura 5.4: Histograma dos <i>clusters</i> da 1º análise RFM com a métrica <i>Average Deviation</i> . .....	61
Figura 5.5: Gráfico da visualização dos <i>clusters</i> da 1º análise RFM com a métrica <i>Average Deviation</i> . ....	62
Figura 5.6: Gráfico da visualização dos <i>clusters</i> da 1º análise RFMP com a métrica <i>Davies-Bouldin</i> . ....	75
Figura 5.7: Gráfico da visualização dos <i>clusters</i> da 1º análise RFMP com a métrica <i>Dunn</i> . .....	81
Figura 5.8: Quantidade de registros por <i>cluster</i> da 1º análise RFMP com a métrica <i>Average Deviation</i> . ....	85
Figura 5.9: Histograma dos <i>clusters</i> da 1º análise RFMP com a métrica <i>Average Deviation</i> . ....	86
Figura 5.10: Gráfico da visualização dos <i>clusters</i> da 1º análise RFMP com a métrica <i>Average Deviation</i> . ....	86
Figura 5.11: Gráfico da visualização dos <i>clusters</i> da análise RFM ( <i>Z-Score</i> ) com a métrica <i>Simplified Silhouette</i> . ....	96
Figura 5.12: Quantidade de registros por <i>cluster</i> da análise RFM ( <i>Z-Score</i> ) com a métrica <i>Average Deviation</i> . ....	100

Figura 5.13: Histograma dos <i>clusters</i> da análise RFM ( <i>Z-Score</i> ) com a métrica <i>Average Deviation</i> . .....	100
Figura 5.14: Gráfico da visualização dos <i>clusters</i> da análise RFM ( <i>Z-Score</i> ) com a métrica <i>Average Deviation</i> . .....	101
Figura 5.15: Gráfico da visualização dos <i>clusters</i> da análise RFMP ( <i>Z-Score</i> ) com a métrica <i>Simplified Silhouette</i> . .....	111
Figura 5.16: Gráfico da visualização dos <i>clusters</i> da análise RFMP ( <i>Z-Score</i> ) com a métrica <i>Simplified Silhouette</i> . .....	114
Figura 5.17: Quantidade de registros por <i>cluster</i> da análise RFMP ( <i>Z-Score</i> ) com a métrica <i>Average Deviation</i> . .....	116
Figura 5.18: Histograma dos <i>clusters</i> da análise RFMP ( <i>Z-Score</i> ) com a métrica <i>Average Deviation</i> . .....	117
Figura 5.19: Gráfico da visualização dos <i>clusters</i> da análise RFMP ( <i>Z-Score</i> ) com a métrica <i>Average Deviation</i> . .....	117
Figura 7.1: Quantidade de registros por <i>cluster</i> da análise RFM do Modelo Sugerido. ....	144
Figura 7.2: Histograma dos <i>clusters</i> da análise RFM do Modelo Sugerido. ....	145
Figura 7.3: Gráfico da visualização dos <i>clusters</i> da análise RFM do Modelo Sugerido. ....	145
Figura 7.4: Quantidade de registros por <i>cluster</i> da análise RFMP do Modelo Sugerido. ....	158
Figura 7.5: Histograma dos <i>clusters</i> da análise RFMP do Modelo Sugerido. ....	158
Figura 7.6: Gráfico da visualização dos <i>clusters</i> da análise RFMP do Modelo Sugerido. ....	159

# Lista de Tabelas

Tabela 2.1: Principais artigos utilizados pela tese ordenados por ano de publicação. ...	22
Tabela 3.1: Dados sobre o número de clientes e pedidos. ....	29
Tabela 3.2: Gráficos com os histogramas de cada variável R, F, M e P. ....	31
Tabela 3.3: Gráficos <i>box-plot</i> das variáveis R, F, M e P. ....	31
Tabela 3.4: Correlação entre as variáveis de entrada utilizadas pelo estudo. ....	34
Tabela 4.1: Relação da frequência de compras e clientes recorrentes. ....	36
Tabela 5.1: Pontuação da métrica <i>Simplified Silhouette</i> da análise RFM. ....	49
Tabela 5.2: Gráficos dos resultados da métrica <i>Simplified Silhouette</i> da análise RFM. ....	50
Tabela 5.3: Pontuação da métrica <i>Davies-Bouldin</i> da análise RFM. ....	54
Tabela 5.4: Gráficos dos resultados da métrica <i>Davies-Bouldin</i> da análise RFM. ....	55
Tabela 5.5: Pontuação da métrica <i>Dunn</i> da análise RFM. ....	59
Tabela 5.6: Pontuação da métrica <i>Average Deviation</i> da análise RFM. ....	60
Tabela 5.7: Tabela auxiliar com os resultados da métrica <i>Simplified Silhouette</i> da análise RFM. ....	70
Tabela 5.8: Tabela auxiliar com os resultados das métricas <i>Davies-Bouldin</i> e <i>Dunn</i> da análise RFM. ....	70
Tabela 5.9: Tabela auxiliar com os resultados da métrica <i>Average Deviation</i> da análise RFM. ....	71
Tabela 5.10: Pontuação da métrica <i>Simplified Silhouette</i> da análise RFMP. ....	72
Tabela 5.11: Pontuação da métrica <i>Davies-Bouldin</i> da análise RFMP. ....	74
Tabela 5.12: Gráficos dos resultados da métrica <i>Davies-Bouldin</i> da RFMP. ....	75
Tabela 5.13: Pontuação da métrica <i>Dunn</i> da análise RFMP. ....	80
Tabela 5.14: Gráficos dos resultados da métrica <i>Dunn</i> da análise RFMP. ....	80
Tabela 5.15: Pontuação da métrica <i>Average Deviation</i> da análise RFMP. ....	84
Tabela 5.16: Tabela auxiliar com os resultados da métrica <i>Simplified Silhouette</i> da análise RFMP. ....	92
Tabela 5.17: Tabela auxiliar com os resultados da métrica <i>Davies-Bouldin</i> da análise RFMP. ....	92
Tabela 5.18: Tabela auxiliar com os resultados da métrica <i>Dunn</i> da análise RFMP. ....	93
Tabela 5.19: Tabela auxiliar com os resultados da métrica <i>Dunn</i> da análise RFMP <i>Average Deviation</i> . ....	93
Tabela 5.20: Pontuação da métrica <i>Simplified Silhouette</i> da análise RFM ( <i>Z-Score</i> ). ....	95

Tabela 5.21: Gráficos dos resultados da métrica <i>Simplified Silhouette</i> da análise RFM ( <i>Z-Score</i> ).....	96
Tabela 5.22: Pontuação da métrica <i>Davies-Bouldin</i> da análise RFM ( <i>Z-Score</i> ).....	98
Tabela 5.23: Pontuação da métrica <i>Dunn</i> da análise RFM ( <i>Z-Score</i> ).....	98
Tabela 5.24: Pontuação da métrica <i>Average Deviation</i> da análise RFM ( <i>Z-Score</i> ).....	99
Tabela 5.25: Tabela auxiliar com os resultados das métricas <i>Simplified Silhouette</i> , <i>Davies Bouldin</i> e <i>Dunn</i> da análise RFM ( <i>Z-Score</i> ).....	108
Tabela 5.26: Tabela auxiliar com os resultados da métrica <i>Average Deviation</i> da análise RFM ( <i>Z-Score</i> ).....	108
Tabela 5.27: Pontuação da métrica <i>Simplified Silhouette</i> da análise RFMP ( <i>Z-Score</i> ).....	110
Tabela 5.28: Gráficos dos resultados da métrica <i>Simplified Silhouette</i> da análise RFMP ( <i>Z-Score</i> ).....	111
Tabela 5.29: Pontuação da métrica <i>Davies-Bouldin</i> da análise RFMP ( <i>Z-Score</i> ).....	113
Tabela 5.30: Pontuação da métrica <i>Dunn</i> da análise RFMP ( <i>Z-Score</i> ).....	113
Tabela 5.31: Gráficos dos resultados da métrica <i>Dunn</i> da análise RFMP ( <i>Z-Score</i> )... ..	114
Tabela 5.32: Pontuação da métrica <i>Average Deviation</i> da análise RFMP ( <i>Z-Score</i> )... ..	116
Tabela 5.33: Tabela auxiliar com os resultados das métricas <i>Simplified Silhouette</i> e <i>Davies Bouldin</i> da análise RFMP ( <i>Z-Score</i> ).....	125
Tabela 5.34: Tabela auxiliar com os resultados da métrica <i>Dunn</i> da análise RFMP ( <i>Z-Score</i> ).....	125
Tabela 5.35: Tabela auxiliar com os resultados da métrica <i>Average Deviation</i> da análise RFMP ( <i>Z-Score</i> ).....	126
Tabela 5.36: Resumo dos resultados da primeira avaliação dos modelos.....	129
Tabela 6.1: Divisão das classes do atributo R em trimestres e seus comportamentos.....	136
Tabela 6.2: Divisão das classes do atributo F e seus comportamentos.....	137
Tabela 6.3: Divisão das classes do atributo M e seus comportamentos.....	137
Tabela 6.4: Divisão das classes do atributo P e seus comportamentos.....	138
Tabela 6.5: Faixas de valores para classificação RFM/P do atributo R.....	139
Tabela 6.6: Faixas de valores para classificação RFM/P do atributo F.....	140
Tabela 6.7: Faixas de valores para classificação RFM/P do atributo M.....	141
Tabela 6.8: Faixas de valores para classificação RFMP do atributo P.....	141
Tabela 7.1: Pontuação da Taxa de Acertos RFM do Modelo Sugerido.....	143
Tabela 7.2: Tabela auxiliar com os resultados do Modelo Sugerido RFM.....	154
Tabela 7.3: Tabela auxiliar com os resultados do Modelo Final RFM.....	156

Tabela 7.4: Pontuação da Taxa de Acertos RFMP do Modelo Sugerido. ....	157
Tabela 7.5: Tabela auxiliar com os resultados do Modelo Sugerido RFMP. ....	166
Tabela 7.6: Tabela auxiliar com os resultados do Modelo Final RFMP. ....	172
Tabela 7.7: Resumo dos resultados da avaliação final dos modelos. ....	174

# 1 Introdução

Com a utilização das técnicas de *Machine Learning* as empresas passaram a contar com um poderoso recurso para gerar conhecimento através da exploração dos seus conjuntos de dados. Contudo, a grande maioria das aplicações potenciais ainda está à espera de suas implementações, pois o maior benefício não virá pelo simples uso dos algoritmos – afinal, tais métodos já existem há décadas – mas sim da utilização pelas empresas para a geração de valor através das soluções de problemas reais.

Deste modo, similar a qualquer outro projeto, o uso dos métodos de *Machine Learning* deve ser precedido por uma profunda análise sobre a definição de um problema significativo que a empresa deseja resolver. Em seguida, é preciso avaliar se com os dados existentes há a possibilidade para responder ao problema levantado, e se tais dados estão disponíveis para análise.

Na sequência, uma vez identificado o problema de negócio, e todas as considerações sobre a coleta de dados e seus respectivos processamentos, deve-se definir quem utilizará a solução proposta e de que maneira criará valor para as companhias, o que envolve um processo humano de interação. Assim, é necessário tirar o foco dos detalhes técnicos e ter a ciência de dados mais orientada às soluções de negócio, conectando diretamente os modelos desenvolvidos às ações estratégicas das empresas.

Por isso, é preciso contexto para que as soluções propostas com o uso de tais técnicas façam sentido e tragam resultados positivos às organizações. Pois do contrário, do que adiantará criar novas soluções se não houver quem siga adiante com as suas respectivas execuções e resultados.

E foi justamente pela busca de contexto que a tese conseguiu encontrar o principal incentivo para o seu desenvolvimento. Primeiro, pela criação de perfis de clientes a partir de uma das principais motivações das empresas: o lucro, ao aplicar uma nova abordagem para melhorar os modelos RFM, a partir da criação de modelos RFMP. Segundo, pelo uso de um serviço de aprendizado de máquina com base nos recursos da computação em nuvem, o que possibilitou focar naquilo que era mais importante: a solução de um problema real, e não em um tecnológico.

## 1.1 Objetivos

Este trabalho tem por objetivo propor a criação de um novo modelo, denominado RFMP (*Recency – Frequency – Monetary – Profitability*), como alternativa à criação do tradicional modelo RFM desenvolvido por (Hughes, 1994). O plano é avaliar se a inclusão de um novo parâmetro P (da palavra inglesa *Profitability*) – relacionado à lucratividade dos consumidores – trará algum impacto no processo de segmentação de clientes. Todo este desenvolvimento será feito a partir de uma plataforma de aprendizado de máquina na nuvem, através da segmentação de uma base de consumidores de um site de *e-commerce*.

A pesquisa propõe também a criação de um novo método para a classificação dos modelos RFM e RFMP, em substituição à forma tradicional utilizada pela metodologia padrão do modelo RFM (quando realizado em conjunto com os métodos de *Machine Learning*). Esta nova proposta tem por objetivo aumentar o grau de confiança das classificações dos modelos gerados e melhorar a assertividade entre os *clusters* e seus respectivos conteúdos.

A tese recomenda também a criação de três novos índices de mensuração para avaliar os resultados obtidos por cada modelo. Estes índices têm por objetivo suprir a ausência de indicadores existentes no processo de criação de modelos RFM e RFMP, com o intuito de ajudar na comparação dos resultados e mensurar a consistência e a qualidade dos *clusters* e modelos produzidos.

Por fim, será apresentada uma análise sobre o uso da plataforma de aprendizado de máquina utilizada para o desenvolvimento desta pesquisa, realizada a partir do ambiente computacional em nuvem denominado *Azure Machine Learning*.

## 1.2 Metodologia

A metodologia utilizada pelo estudo seguiu uma abordagem exploratória descritiva, que teve por objetivo caracterizar o processo existente e identificar as variáveis e fluxos pertinentes que foram abordados ao longo da tese.

Quanto ao desenvolvimento, ele foi iniciado a partir de uma breve revisão literária. Em seguida, foi feito um estudo experimental para a implementação e avaliação de uma pesquisa empírica – com um conjunto de dados transacionais de um portal de

comércio eletrônico – para a criação de modelos RFM e RFMP, através de uma plataforma de aprendizado de máquina na nuvem.

Na sequência foram avaliados os problemas identificados e as recomendações sugeridas. Porém, a cada etapa do trabalho foram descritas detalhadamente as fundamentações teóricas e os métodos empregados para alcançar os objetivos da tese. No final, é apresentada uma conclusão sobre os resultados obtidos e seus impactos futuros.

### **1.3 Organização do Trabalho**

O conteúdo desta tese está organizado em oito capítulos, divididos da seguinte forma:

- ✓ **Capítulo 1** – expõe as motivações por detrás deste trabalho, bem como os objetivos e a metodologia que foi utilizada ao longo da tese;
- ✓ **Capítulo 2** – fundamenta a base teórica decorrente de toda pesquisa adotada, como os conceitos básicos, a revisão da literatura, o estado da arte e as propostas de contribuição;
- ✓ **Capítulo 3** – apresenta o conjunto de dados e o processo das estatísticas básicas;
- ✓ **Capítulo 4** – descreve sobre o processo inicial utilizado para o desenvolvimento dos modelos, assim como o processo de avaliação e métricas que foram utilizadas;
- ✓ **Capítulo 5** – detalha o desenvolvimento dos modelos RFM e RFMP através do processo de classificação original;
- ✓ **Capítulo 6** – identifica os obstáculos encontrados pela análise inicial e propõe as recomendações necessárias para as soluções dos desafios;
- ✓ **Capítulo 7** – desenvolve e avalia os novos modelos produzidos a partir das recomendações sugeridas, bem como da nova proposta de classificação dos modelos RFM e RFMP; e
- ✓ **Capítulo 8** – apresenta as considerações finais da tese e sugere as principais direções a serem seguidas na continuidade desta pesquisa.

Desta forma, o estudo finaliza sua apresentação e segue adiante com a execução das tarefas necessárias ao seu desenvolvimento.

## 2 Fundamentação Teórica

### 2.1 Conceitos Básicos

O presente estudo combina diferentes campos de pesquisa e possui um potencial de alcance interdisciplinar. Sua proposta pode ser aplicada em diversas áreas do conhecimento, como: na tecnologia da informação, na gestão empresarial, no desenvolvimento de estratégias de marketing, no setor de vendas, no segmento de *e-commerce* e no avanço das pesquisas científicas a partir do uso das técnicas de *Machine Learning* e *Cloud Computing*.

Por isso, a tese apresenta a seguir alguns conceitos básicos a respeito dos principais temas abordados durante a pesquisa.

#### 2.1.1 *E-Commerce*

Segundo dados do último relatório *Webshoppers* (Ebit, 2018), o faturamento do comércio eletrônico brasileiro foi de R\$ 47,7 bilhões em 2017, o que representa um crescimento nominal de 7,5% em relação a 2016. Este é um mercado que continua em forte expansão, crescendo rapidamente, ano após ano, tanto pelo aumento do número de usuários com acesso à Internet (em especial a partir dos acessos feitos através dos dispositivos móveis, os chamados *smartphones*) quanto pelo excelente apelo no que diz respeito à praticidade e conveniência no processo de compra dos produtos e na contratação de serviços.

Contudo, um ponto de destaque, e de grande valor para as companhias que fazem uso do *e-commerce*, está na facilidade que estas ferramentas proporcionam às empresas para coletar informações relacionadas aos dados transacionais e comportamentais dos seus consumidores. Entretanto, do que adianta gerar e manter um grande volume de dados se as empresas continuam sem saber como processá-los e o que fazer a partir deste processamento.

Por isso, muitas empresas falham neste objetivo por uma razão simples: o principal desafio não está associado à construção de sistemas ou banco de dados, mas sim nas estratégias que deveriam ser adotadas através do conhecimento extraído a partir das suas bases de dados. Ou seja, não é uma questão de tecnologia, mas sim do uso das

informações a favor do desenvolvimento de estratégias que permitam às empresas alavancarem os seus negócios.

Este é um problema antigo e que há anos vem sendo tratado de diversas maneiras. Inclusive, através da utilização do modelo RFM e que será discutido logo em seguida.

### **2.1.2 Modelo RFM (*Recency – Frequency – Monetary*)**

O modelo *RFM* (*Recency – Frequency – Monetary*), criado por (Hughes, 1994), é um método de classificação de registros de clientes que permite saber quem são os compradores mais recentes, os mais frequentes e os mais rentáveis, ou seja, é uma forma que possibilita a criação de perfis através do comportamento de compra de cada consumidor.

Este conceito permite a segmentação dos clientes – e o desenvolvimento de campanhas de marketing personalizadas – através da identificação dos diferentes tipos de consumidores a partir do histórico de compras de cada um deles e de seus respectivos relacionamentos com as empresas. Deste modo, ele diferencia os clientes em três atributos: R (Recente), F (Frequente) e M (Monetário), e mensura de fato quando as pessoas compraram, quantas vezes compraram e quanto elas compraram em um determinado período (Wei *et al.*, 2010).

Especificamente, o R (de Recente) denota a duração do período de tempo desde a última compra; já o F (de Frequente) significa o número de compras dentro de um determinado período; e o M (de Monetário) representa a quantidade de dinheiro gasto durante um intervalo de tempo (Wang, 2010). Sendo assim, o atributo R pode ser definido como uma métrica temporal que indica o quão recente é um cliente, o atributo F pode ser considerado como uma métrica da força do relacionamento do cliente com a empresa, e o atributo M significa o valor gasto (Miguéis *et al.*, 2012), seja ele a partir do valor total ou pelo tíquete médio das compras de um determinado cliente.

Desta forma, a maneira tradicional para a segmentação dos modelos RFM é realizada através da classificação dos dados de cada cliente a partir dos valores contidos em cada um dos atributos R, F e M. E isto é feito pela divisão do conjunto de dados em cinco partes iguais, de modo que os dados de R são classificados em ordem crescente, ou seja, do mais recente para o mais antigo; enquanto que os dados de F e M são classificados

em ordem decrescente. Neste caso, F é ordenado pelos registros de compras da maior frequência para a menor; e M é ordenado a partir do maior valor monetário para o menor.

Assim, para cada atributo, os 20% dos registros iniciais do topo do ordenamento são codificados como 5, o próximo segmento de 20% como 4, e assim por diante. Deste modo, todos os clientes podem ser apresentados por rótulos como 5-5-5, 5-5-4, ..., 1-1-1; o que dá um total de 125 possibilidades de classificação ( $5 \times 5 \times 5$ ) (Wei *et al.*, 2013).

Porém, esta abordagem tradicional também traz algumas desvantagens. Isto porque o procedimento de codificação de 1 a 5 é arbitrário e faz com que cada grupo contenha quase sempre o mesmo número de registros. Deste modo, há um balanceamento do número de clientes em cada segmento, o que torna o modelo mais sensível à divisão por percentil do que ao comportamento de compras dos clientes. E isto acaba por gerar um problema: pois se os recursos são escassos, que tipo de cliente ou segmento de cliente deverá ser trabalhado?

É que para uma empresa pequena, uma segmentação com 125 grupos pode ser um excesso (até pela quantidade de clientes contidos em cada um destes grupos). Mas em compensação, para uma grande corporação, como é o caso da varejista de *e-commerce* americana Amazon, por exemplo, uma classificação com 125 grupos é muito pouco. Além disso, há a possibilidade de um determinado segmento conter clientes com padrões de compras muito distintos, da mesma forma que pode permitir que clientes com comportamentos idênticos estejam em grupos separados.

Outra desvantagem é que o procedimento de discretização traz como consequência uma perda de informação (Coussement *et al.*, 2014), ao rotular todos os clientes com atributos de 1 a 5, abrindo mão dos verdadeiros valores contidos em cada domínio RFM. Por isso, para que as classificações do conjunto de dados fossem feitas sem os rótulos de 1 a 5 (mantendo então os valores originais contidos nas propriedades dos atributos R, F e M), o modelo RFM passou a ser integrado com sucesso no processo de mineração de dados, já que possuem métodos muito eficientes e escaláveis na tarefa de agrupar os dados. Como foi o caso da aplicação do algoritmo *k-means* (Macqueen, 1967) pelos estudos de (Wu *et al.*, 2009) e (Chang *et al.*, 2010), por exemplo.

Deste modo, os registros são segmentados a partir de um método de classificação não supervisionada usando os valores originais. Em seguida, para todo *cluster*

encontrado, a média de cada atributo RFM é computada; depois, a média global de cada atributo é calculada a partir de todos os registros do conjunto de dados; e por último, as médias dos *clusters* são comparadas com as médias globais. Portanto, a comparação entre a média do *cluster* e a média global é que irá definir a classificação do modelo RFM. Desta maneira, se a média de um *cluster* para o atributo R for menor ou igual que a média global, ele será classificado como positivo (+), pois quanto mais recente, melhor; mas se a média do *cluster* ficar acima da média global, o *cluster* será classificado como negativo (-). O mesmo ocorre com os atributos F e M. Porém, nestes casos a regra se inverte, pois quanto maior o valor, mais importante ele é, ou seja, se um *cluster* tiver a média de F ou M maior ou igual do que a média global, este *cluster* será classificado como positivo (+), isto é, quanto mais recente ou mais monetário, melhor; porém, se a média do *cluster* ficar abaixo da média global, este *cluster* será classificado como negativo (-).

No entanto, as técnicas de mineração de dados estão praticamente disponíveis aos pesquisadores e cientistas de dados, o que dificulta a adoção pelas companhias. Em especial, pelos obstáculos impostos quanto às necessidades de vultosos investimentos na compra de equipamentos, *softwares* e contratação de equipe especializada. Contudo, graças aos recursos existentes na computação em nuvem, que torna as soluções mais robustas, integradas e escaláveis, tal cenário começa a mudar.

### **2.1.3 Computação em Nuvem**

O processamento e a análise de grandes quantidades de dados requerem uma combinação de *hardware* e *software* de alto desempenho, o que para muitas organizações ainda é caro e injustificável (Demirkan e Delen, 2013). Porém, para minimizar estes impactos, grandes empresas fornecedoras de TI passaram a disponibilizar seus recursos através de um modelo de negócio em nuvem, a partir de um formato de plataforma de serviço – o que permite uma contratação de acordo com a necessidade e a capacidade de investimento de cada companhia. Já que segundo (Zhang *et al.*, 2010), uma das principais características da computação em nuvem é a capacidade de adquirir e liberar recursos sob demanda.

Este é um modelo em que os serviços de computação (tanto de *hardware* como de *software*) são entregues em uma rede de maneira auto atendida, independente do dispositivo e do local. De modo que os recursos são compartilhados, escaláveis

dinamicamente, rapidamente provisionados, virtualizados e lançados com uma interação mínima. Assim, os usuários pagam o serviço como uma despesa operacional, sem incorrer – inicialmente – em qualquer despesa significativa de capital (Marston *et al.*, 2011).

Desta forma, com este modelo de computação em nuvem as empresas e os usuários são capazes de acessar aplicativos de qualquer lugar do mundo sob demanda, transformando o desenvolvimento de *software* em um serviço, em vez de serem executados em computadores individuais (Buyya *et al.*, 2009). Entretanto, conforme apresentado por (Al-Aqrabi *et al.*, 2015), existem três principais modelos de entrega de serviços de computação em nuvem, sendo: *Software* como Serviço (*SaaS*), Plataforma como Serviço (*PaaS*) e Infraestrutura como Serviço (*IaaS*).

De modo que cada um destes modelos é melhor definido por (Sinha e Khreisat, 2014) da seguinte maneira:

- *SaaS* – neste modelo, os usuários não técnicos podem usar um aplicativo na nuvem sem ter que comprar e instalar os *softwares* em suas próprias máquinas.
- *PaaS* – fornece uma plataforma na qual os usuários não precisam gerenciar a infraestrutura de nuvem subjacente.
- *IaaS* – fornece uma infraestrutura completa com servidores, armazenamento, rede, bancos de dados, etc.

Por tudo isso, uma arquitetura orientada a serviços, combinada a uma infraestrutura em nuvem, pode fornecer os recursos e a flexibilidade que os sistemas de análise de negócios tanto precisam para cumprir suas promessas (Delen e Demirkan, 2013). Assim, as empresas fornecedoras de tecnologia não só conseguem oferecer um ambiente de desenvolvimento flexível, mas também garantem uma utilização eficaz e eficiente dos recursos computacionais para produzir resultados mais precisos, completos e oportunos. Além disso, a computação em nuvem pode levar a um mercado mais competitivo e, portanto, a preços mais baixos (Buyya *et al.*, 2009).

Desta forma, o uso de sistemas de suporte à decisão orientados a serviços na nuvem é uma das principais tendências para muitas organizações na esperança de se tornarem mais ágeis (Demirkan e Delen, 2013). No entanto, é bom ter em mente que

apesar dos benefícios significativos oferecidos pela computação em nuvem, as tecnologias atuais não estão maduras o suficiente para realizar todo o seu potencial (Zhang *et al.*, 2010). E um ponto muito crítico e bastante importante sobre este tema é justamente a segurança, assunto que já foi tratado por (Sun *et al.*, 2011), (Schroepfer *et al.*, 2013) e (Rao e Selvamani, 2015).

De todo modo, esta nova abordagem de análise de dados orientada a serviços, combinada com os recursos existentes da computação em nuvem, irá criar grandes oportunidades para os negócios das empresas, assim como trará também muitos desafios para serem resolvidos.

#### **2.1.4 *Analytics-as-a-Services (AaaS)***

Este conceito de orientação a serviço está ganhando popularidade na aplicação de sistemas de apoio à tomada de decisão, já que permite a configuração rápida dos complexos programas (Delen e Demirkan, 2013), além de oferecer uma abordagem genérica para a integração com diversos sistemas (Olejnik *et al.*, 2009).

Por estas razões, muitos fornecedores de tecnologia, em especial os que fazem uso da inteligência computacional, estão migrando seus sistemas para a nuvem. Assim, estas empresas conseguem entregar de uma só maneira um produto de análise de dados para as pessoas não especialistas, através de uma plataforma de serviço (*PaaS*) que muitos passaram a definir pelo termo: “Análise como Serviço” (*AaaS – Analytics as a Services*), além de oferecer às empresas uma economia de custos, melhor desempenho e acesso mais rápido a novas aplicações (Zorrilla e García-Saiz, 2013).

Desta forma, é importante avaliar o uso das plataformas de aprendizado de máquina com base nos recursos da computação em nuvem a partir de um modelo de negócio orientado a serviços. Mas não só pela redução dos custos envolvidos e na facilidade da sua aplicação, como também por focar naquilo que é mais importante: a solução dos problemas reais, e não nos tecnológicos.

### 2.1.5 *Azure Machine Learning (Azure ML)*

*Azure* é a plataforma de serviços de computação em nuvem da Microsoft. Do ponto de vista técnico, ela fornece um modelo de programação projetado para criar aplicativos escaláveis e disponíveis (Tajadod *et al.*, 2012). Já o ambiente *Azure Machine Learning (Azure ML)* é um serviço em nuvem, desenvolvido para a criação de soluções de aprendizado de máquina em um modelo de plataforma como serviço. E uma das suas principais vantagens está na capacidade de integrar facilmente o desenvolvimento de modelos em um padrão de fluxo de trabalho repetitivo para criar soluções de análises preditivas. Isto o torna acessível tanto para um principiante como para um cientista de dados experiente (Barnes, 2015).

Assim, não é preciso a instalação e execução de qualquer aplicativo nos computadores dos usuários, além de não haver a necessidade de se preocupar com as tarefas de manutenção e suporte do próprio ambiente. Quanto ao seu uso, ele é feito a partir de um navegador na Internet que se conecta com o ambiente de desenvolvimento, onde é possível criar modelos de análise e predição de acordo com a necessidade de cada empresa.

Um ponto interessante desta plataforma fica por conta da capacidade de criação de *Web Services* para acesso aos modelos gerados. Este é uma grande benefício, pois facilita a integração com outros sistemas, já que os *Web Services* são implementados por um conjunto de tecnologias que fornecem os mecanismos para comunicação, descrição e descoberta de serviços (Devi *et al.*, 2012).

Já em relação aos custos, a cobrança pode ser feita tanto pelo uso dos recursos computacionais a partir da quantidade de horas utilizadas quanto pelo volume de processamento efetuado. Enquanto que a criação das soluções é feita dentro do ambiente *Microsoft Azure Machine Learning Studio*, que não tem qualquer tipo de cobrança. Por ser gratuito (e não se sabe até quando) ele possui algumas limitações, como o volume de dados a ser processado e a quantidade de predições que podem ser realizadas. Já para os casos mais complexos, que precisem de um processamento mais robusto ou um volume de dados maior, é necessário usar os modelos com preços definidos pela própria empresa, que variam de acordo com diversos requisitos. De qualquer modo, para a realização desta tese, o estudo conseguiu utilizar os recursos deste ambiente sem qualquer custo.

## 2.2 Estado da Arte

### 2.2.1 Pesquisas Relacionadas

Para reforçar a relevância da tese, este estudo fez uma revisão sobre as pesquisas científicas produzidas e publicadas a partir dos temas relacionados com os assuntos abordados. Para isso, foram utilizadas as bases de conhecimento da *Web of Science* (Thomson Reuters / Clarivate) e *Scopus* (Elsevier). Além disso, consultas sobre determinados assuntos também foram analisadas a partir da ferramenta *Google Trends*. A seguir, uma compilação do que foi encontrado:

A primeira pesquisa foi realizada com os termos “*RFM Model*” ou “*RFM Analysis*” como critério de busca. A Figura 2.1 e Figura 2.2 representam todos os trabalhos revisados e agrupados por ano de publicação. Contudo, apesar da relevância do tema para as empresas, não foi encontrado um número muito expressivo de publicações (principalmente em relação aos demais assuntos da tese). De todo modo, foi possível perceber que em 2014 (ano em que a tese começou as suas primeiras pesquisas) houve um salto nas publicações, que continuou até 2016 pela base da *Web of Science* e até 2017 pela base da *Scopus* (não acompanhado em 2015). Já pela Figura 2.3 e Figura 2.4 são observadas as áreas temáticas empregadas pelos artigos publicados, onde é possível notar que o tema se aplica a diversas áreas do conhecimento, com um potencial de alcance interdisciplinar, como: Ciência da Computação, Engenharia, Negócio e Gestão, Inteligência Artificial, Matemática, Ciências Sociais, entre outros.

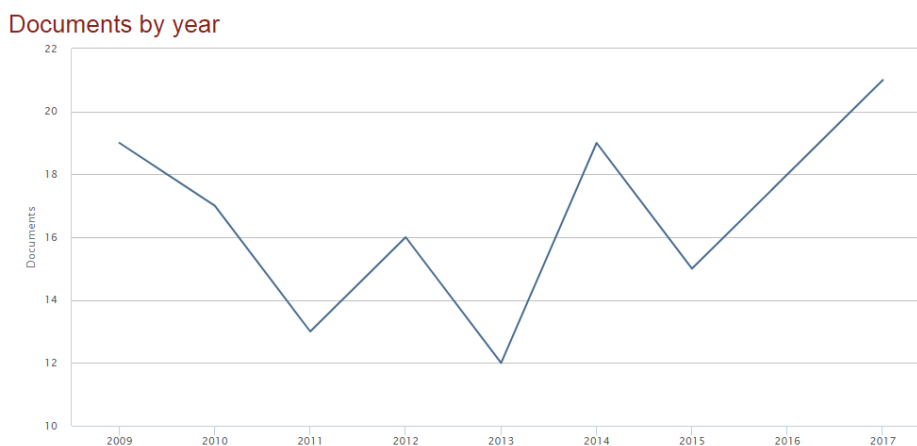


Figura 2.1: Distribuição anual do número de publicações referentes aos termos “*RFM Model*” e “*RFM Analysis*” até dezembro/2017 – Fonte: *Scopus*.

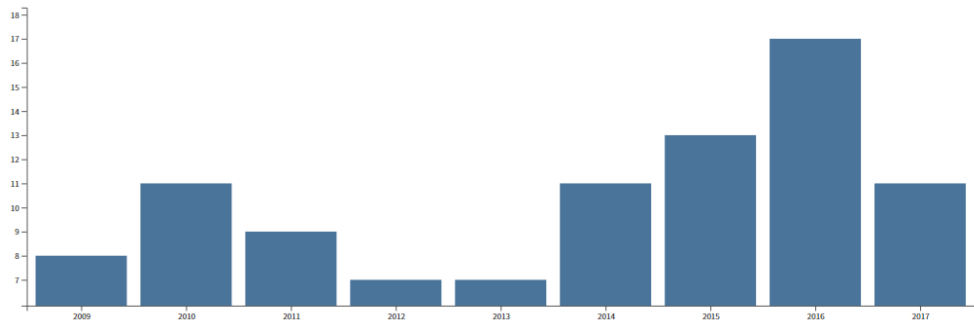


Figura 2.2: Distribuição anual do número de publicações referentes aos termos “*RFM Model*” e “*RFM Analysis*” até dezembro/2017 – Fonte: *Web of Science*.

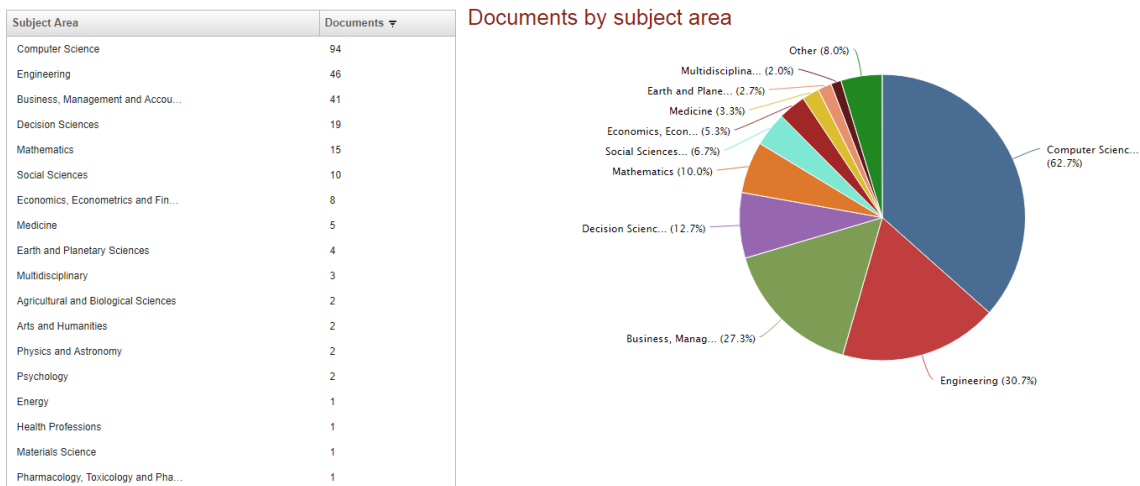


Figura 2.3: Número de publicações por área de domínio referentes aos termos “*RFM Model*” e “*RFM Analysis*” até dezembro/2017 – Fonte: *Scopus*.

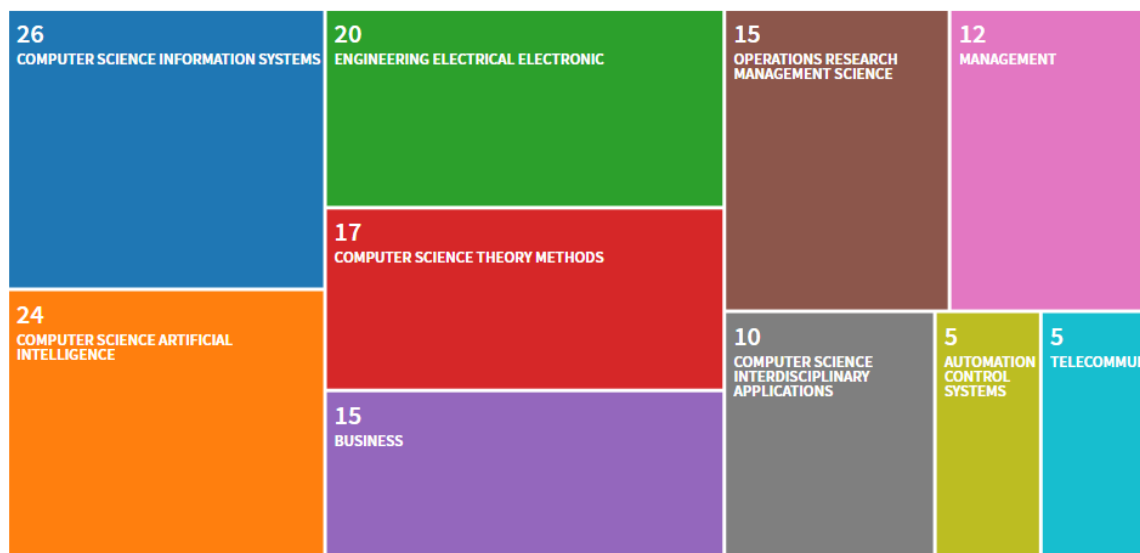


Figura 2.4: Número de publicações por área de domínio referentes aos termos “*RFM Model*” e “*RFM Analysis*” até dezembro/2017 – Fonte: *Web of Science*.

Já as buscas produzidas pelos termos “*Machine Learning*” e “*Data Mining*” foram as campeãs, sem dúvida, conforme observado na Figura 2.5 e Figura 2.6. Tanto que a base da *Web of Science* sequer conseguiu produzir um gráfico de distribuição. Neste caso, para o enriquecimento da pesquisa, o estudo tabulou os dados e produziu um gráfico para mostrar a evolução do uso de cada termo. O mesmo foi feito com os resultados obtidos pela base *Scopus*.

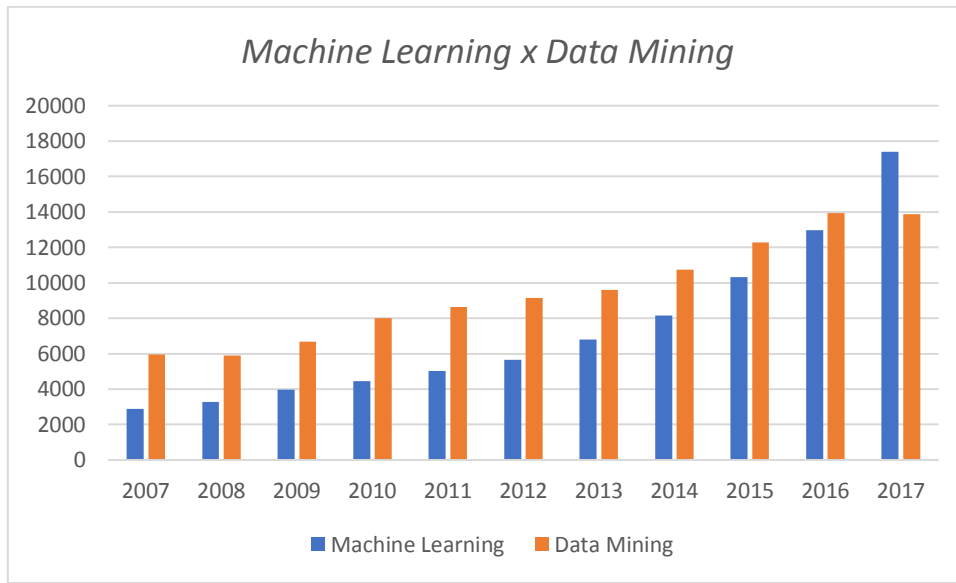


Figura 2.5: Distribuição anual do número de publicações referentes aos termos “*Machine Learning*” e “*Data Mining*” até dezembro/2017 – Fonte: *Scopus*.

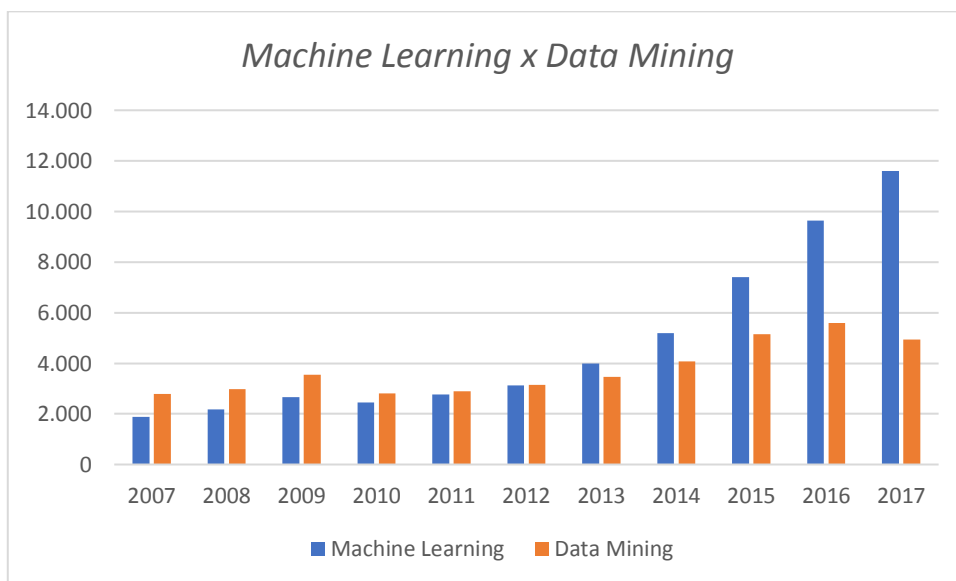


Figura 2.6: Distribuição anual do número de publicações referentes aos termos “*Machine Learning*” e “*Data Mining*” até dezembro/2017 – Fonte: *Web of Science*.

Uma curiosidade interessante foi que o termo “*Machine Learning*” passou a avançar sobre o termo “*Data Mining*” a partir de um determinado período. No caso da base *Web of Science*, este avanço se deu a partir de 2013; já pela base *Scopus*, foi a partir de 2017; enquanto que pelo *Google Trends* (Figura 2.7) foi a partir de 2015.

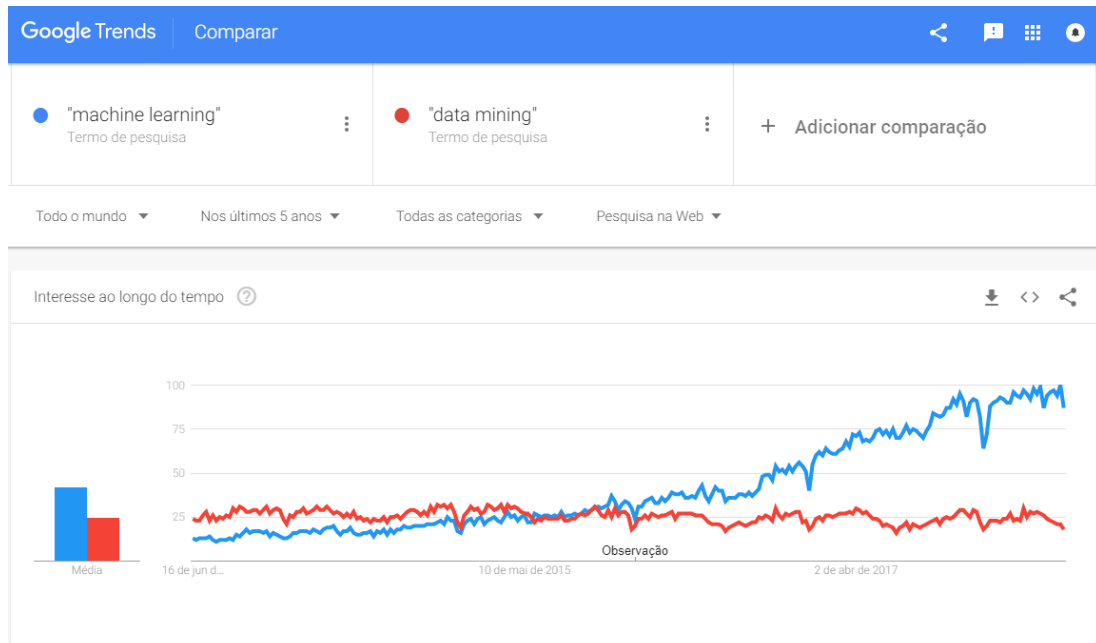


Figura 2.7: Distribuição anual relativa do número de consultas referentes aos termos “*Machine Learning*” e “*Data Mining*” nos últimos 5 anos – Fonte: *Google Trends*.

O termo “*Cloud Computing*” também possui bastante apelo na comunidade científica, quando a partir de 2009 começou a ganhar uma grande relevância, conforme observado na Figura 2.8 e Figura 2.9.

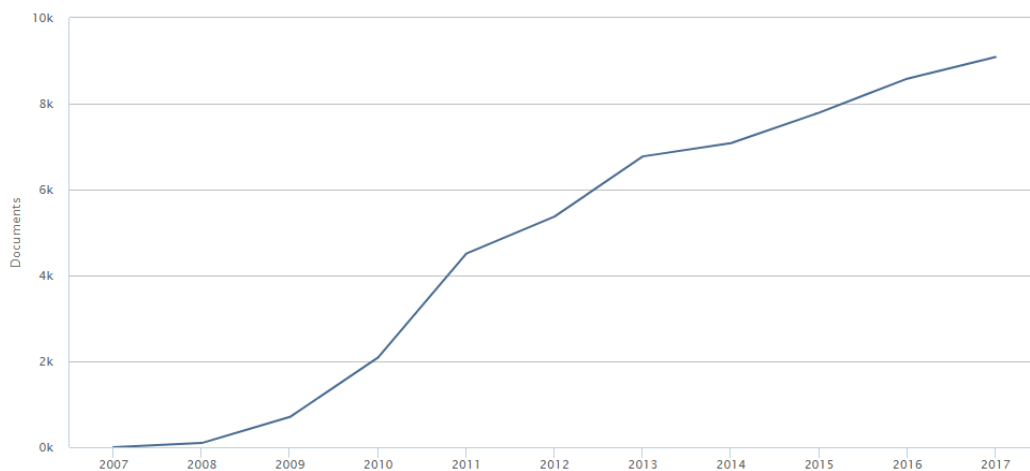


Figura 2.8: Distribuição anual do número de publicações referentes ao termo “*Cloud Computing*” até dezembro/2017 – Fonte: *Scopus*.

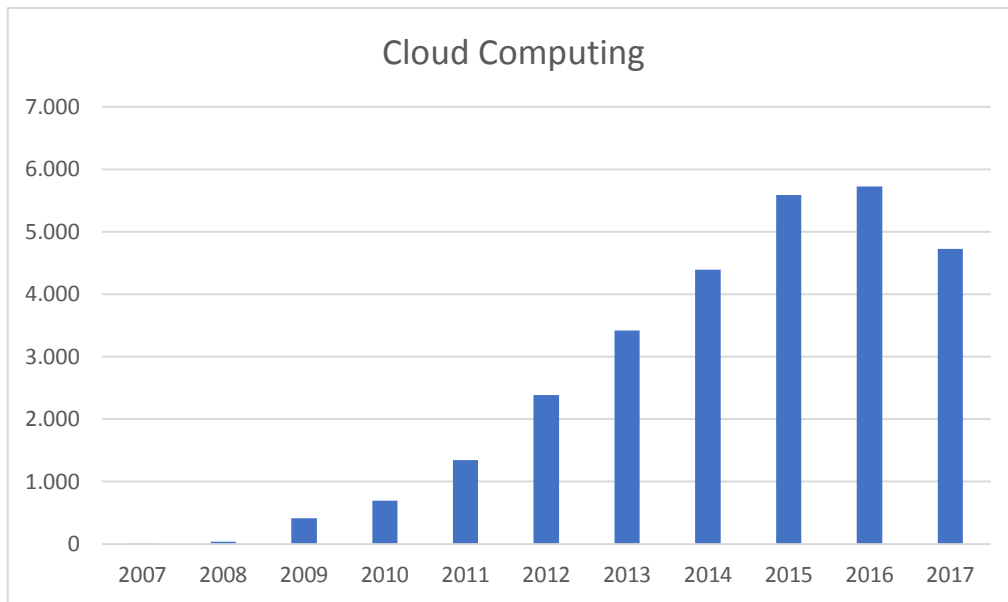


Figura 2.9: Distribuição anual do número de publicações referentes ao termo “*Cloud Computing*” até dezembro/2017 – Fonte: *Web of Science*.

Além disso, também foi possível encontrar os termos “*Cloud Computing*” e “*Machine Learning*” ou “*Data Mining*” vinculados em uma mesma publicação, conforme as buscas realizadas nas bases pesquisadas e demonstradas pela Figura 2.10 e Figura 2.11, o que confirma a relevância entre os temas abordados pela tese.

#### Documents by year

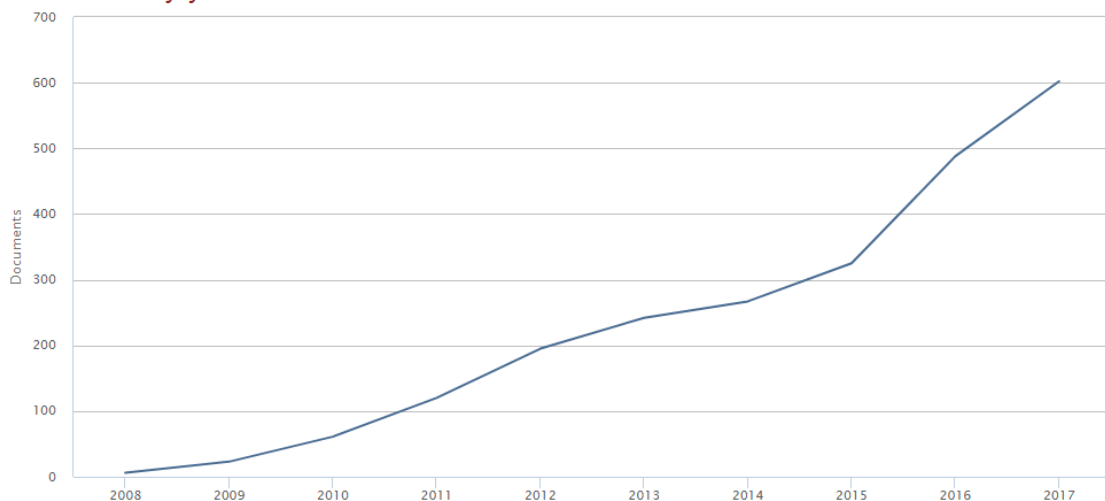


Figura 2.10: Distribuição anual do número de publicações referentes aos termos “*Cloud Computing*” e “*Data Mining*” ou “*Machine Learning*” até dezembro/2017 – Fonte: *Scopus*.

Total de publicações

968

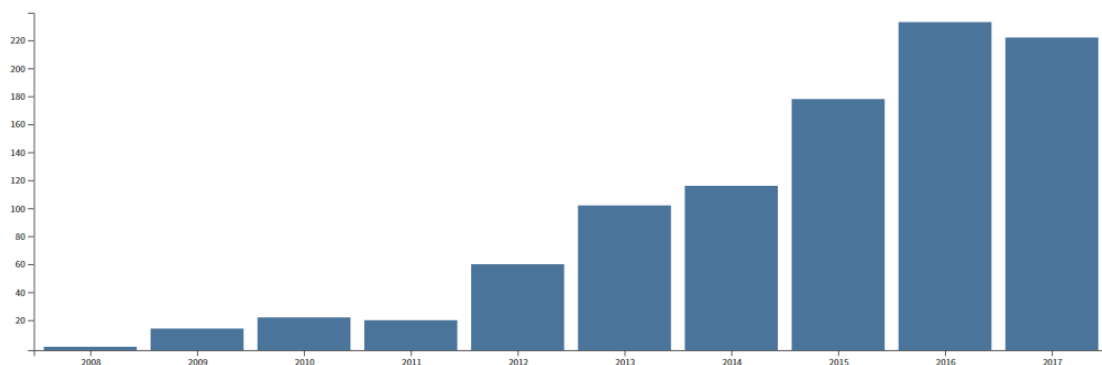


Figura 2.11: Distribuição anual do número de publicações referentes aos termos “*Cloud Computing*” e “*Data Mining*” ou “*Machine Learning*” até dezembro/2017 – Fonte: *Web of Science*.

Já a busca pelos termos “*RFM Model*” ou “*RFM Analysis*” e “*Cloud Computing*” não trouxe qualquer referência nas buscas realizadas. Entretanto, as buscas com os termos “*RFM Model*” ou “*RFM Analysis*” e “*Machine Learning*” ou “*Data Mining*” trouxeram algumas listagens, conforme apresentado na Figura 2.12 e Figura 2.13.

#### Documents by year

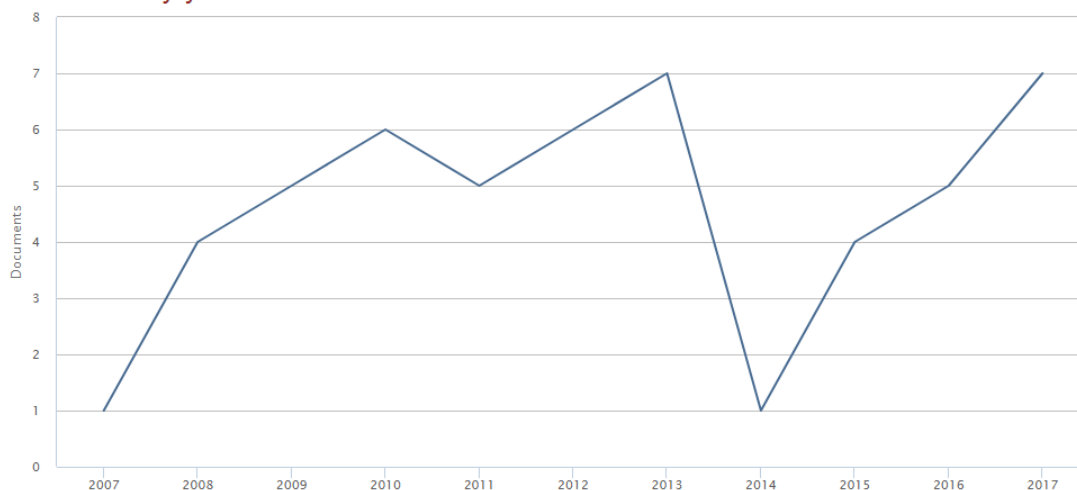


Figura 2.12: Distribuição anual do número de publicações referentes aos termos “*RFM Model*” ou “*RFM Analysis*” e “*Data Mining*” ou “*Machine Learning*” até dezembro/2017 – Fonte: *Scopus*.

Total de publicações

42

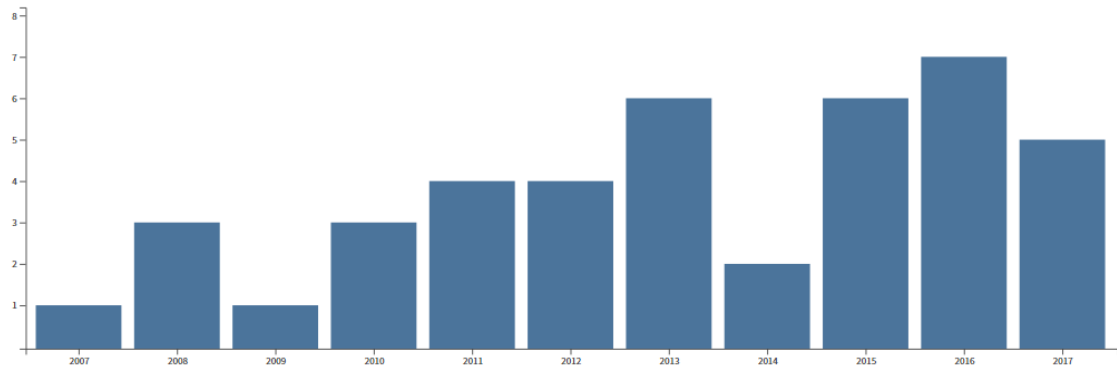


Figura 2.13: Distribuição anual do número de publicações referentes aos termos “RFM Model” ou “RFM Analysis” e “Data Mining” ou “Machine Learning” até dezembro/2017 – Fonte: Web of Science.

Por fim, o estudo reforça a importância dos temas trazidos pela tese a partir dos gráficos de tendências tecnológicas do *Gartner Group* entre os anos de 2014 e 2017, conforme exibidos na Figura 2.14, Figura 2.15, Figura 2.16 e Figura 2.17. De modo que em 2014 (Figura 2.14) os termos “cloud Computing”, “hybrid cloud Computing”, “big data”, “Content Analytics”, “In-Memory Database Management Systems”, “data science” e “Prescriptive Analytics” foram encontrados. Todos, de certa forma, possuem uma relação direta ou indireta com os assuntos tratados pela pesquisa.

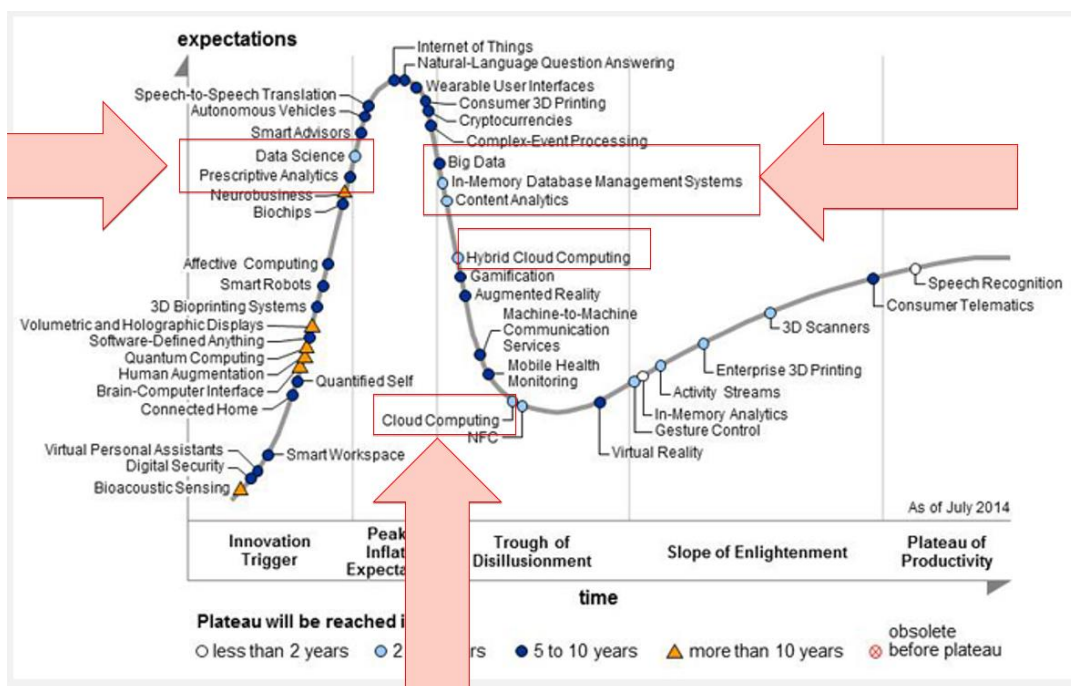


Figura 2.14: Gráfico de tendências tecnológicas do *Gartner Group* 2014.

Já no gráfico de tendências de 2015 (Figura 2.15) foram encontrados os termos: “Customer Journey Analytics” e “Predictive Analytics”.

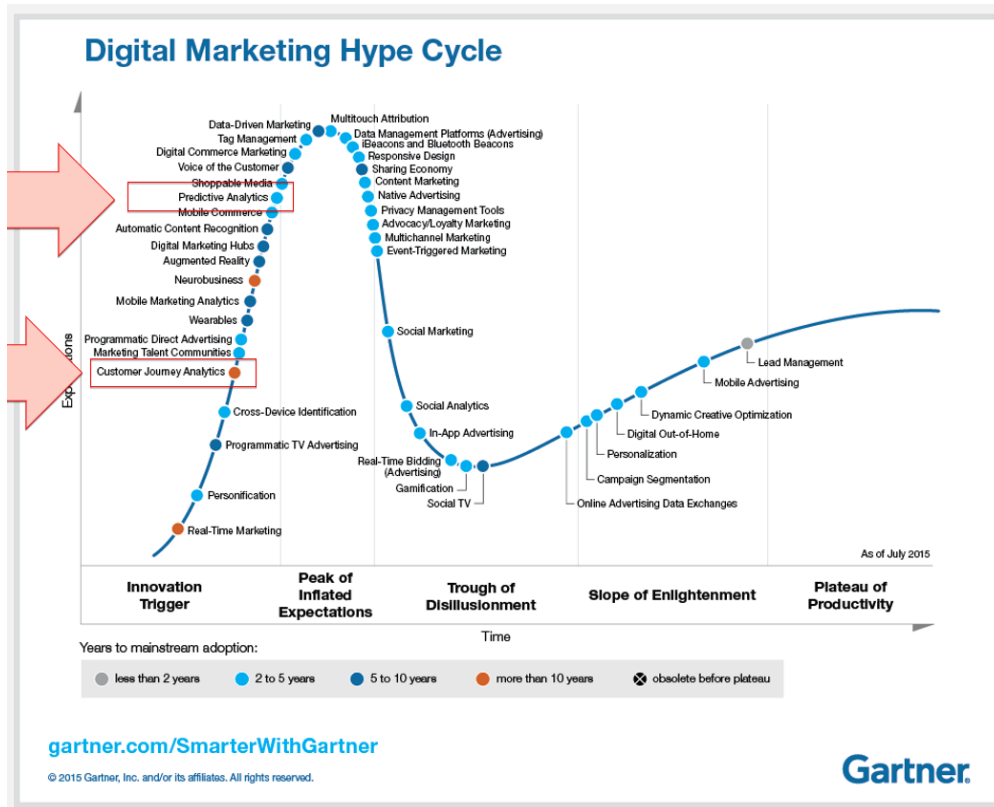


Figura 2.15: Gráfico de tendências tecnológicas do *Gartner Group* 2015.

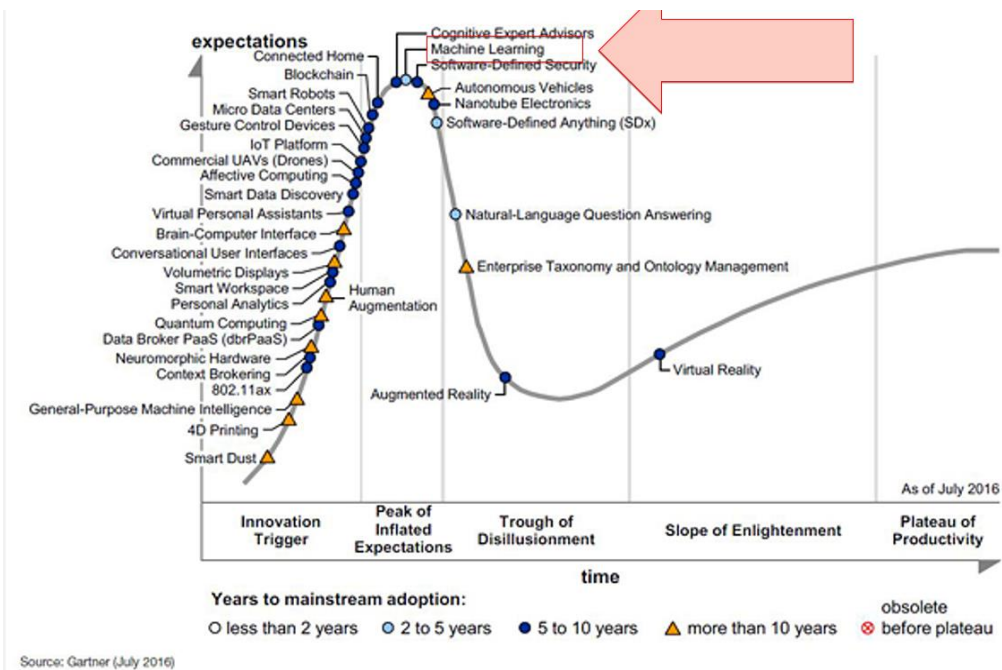


Figura 2.16: Gráfico de tendências tecnológicas do *Gartner Group* 2016.

## Gartner Hype Cycle for Emerging Technologies, 2017

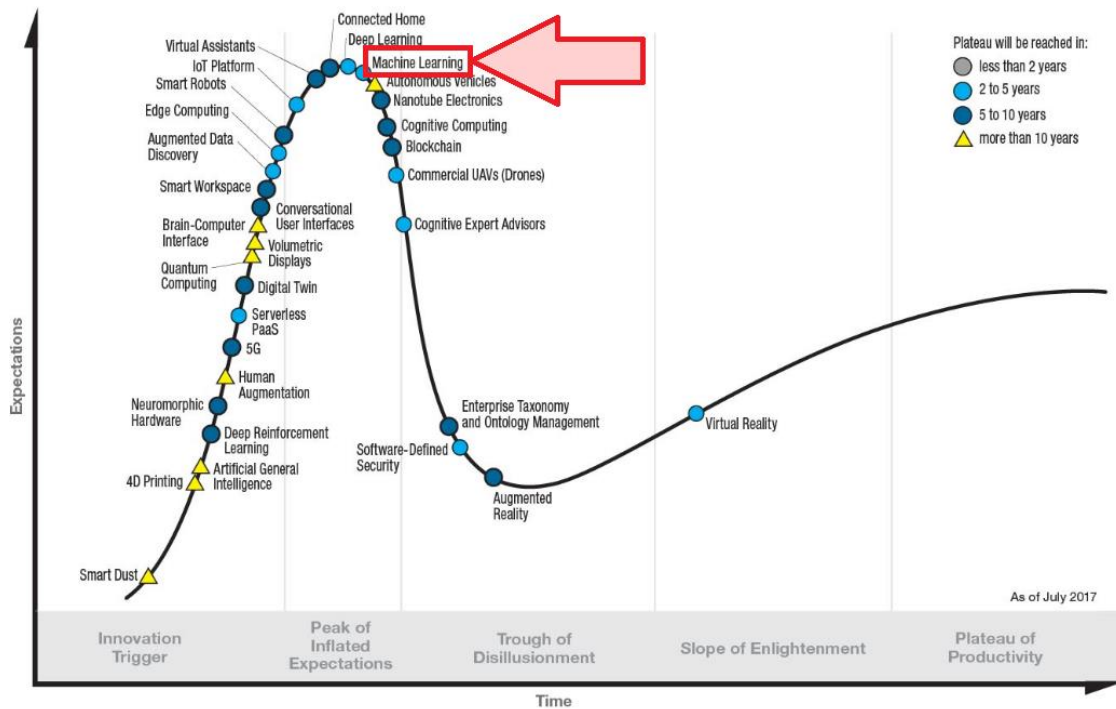


Figura 2.17: Gráfico de tendências tecnológicas do *Gartner Group* 2017.

Em compensação, nos anos de 2016 (Figura 2.16) e 2017 (Figura 2.17) foi o termo “*Machine Learning*” que passou a ter destaque na curva de tendências. Inclusive, o que é mais relevante, por dois anos consecutivos.

Desta forma, por tudo o que foi apresentado, o estudo reafirma a relevância dos temas abordados no desenvolvimento da pesquisa, tanto para a comunidade acadêmica e científica quanto para as empresas e seus estrategistas.

### 2.2.2 Revisão da Literatura

A despeito de todas as referências e citações já realizadas até o momento, e das demais que serão inseridas ao longo da tese, vale destacar uma revisão da literatura a partir dos artigos e publicações que serviram de base para motivar e fundamentar o desenvolvimento deste estudo, conforme é descrito a seguir:

A história por detrás desta pesquisa não poderia começar sem a citação do artigo que deu origem à busca pelo assunto base da tese, o modelo RFM. Este artigo foi encontrado por acaso, muito mais pelo apelo do tema do caso de estudo do que pelo conteúdo encontrado em seu texto, já que ele não fazia qualquer menção ao assunto RFM

em seu título. É que na época, este pesquisador estava envolvido em um projeto específico do setor de *e-commerce* com a indústria de cosméticos, que possui uma forte ligação com o segmento de salões de beleza.

Por isso, o artigo escrito por (Wei *et al.*, 2013), cujo título era: *Customer relationship management in the hairdressing industry: An application of data mining techniques*, despertou tamanha curiosidade, já que descrevia o uso das técnicas de mineração de dados para a gestão de relacionamentos com os clientes (os chamados CRM – *Customer Relationship Management*) aplicadas ao segmento de salões de beleza. E para surpresa e conhecimento, o artigo introduziu a este autor o tema RFM – que em conjunto com as técnicas de agrupamentos baseadas em redes neurais artificiais *Self-Organization Maps – SOM* (Kohonen, 1995) (também conhecidas como redes de *Kohonen* ou mapas auto ajustáveis) e a técnica de agrupamento *k-means* (Macqueen, 1967) – produziu uma segmentação de clientes (com base no comportamento de compras dos consumidores) para desenvolver estratégias de marketing em um salão de beleza de Taiwan. Este foi o início de uma busca literária que culminou em todo o embasamento teórico que orientou o desenvolvimento da tese.

Além disso, com o intuito de aprofundar ainda mais o conhecimento foi analisado o artigo de (Shim *et al.*, 2012), cujo tema tratava das estratégias de CRM para um shopping on-line de pequeno porte com base em regras de associação e padrões sequenciais obtidos pela análise dos dados das transações de compras. Neste artigo, o autor fez uso dos atributos R, F e M em conjunto com outros atributos, como idade, frete, canal de registro (busca orgânica ou recomendação), entre outros, para desenvolver um modelo classificador que fosse capaz de identificar os clientes VIPs a partir de um conjunto de técnicas de mineração de dados. E por fim, propor um conjunto de estratégias de marketing baseadas na identificação das regras e padrões obtidos pela análise.

Em seguida, o artigo de (Khajvand *et al.*, 2011) também aplicou um estudo de caso em uma empresa de saúde e beleza, com o intuito de estimar o valor da vida útil dos clientes a partir de uma análise RFM. Neste caso, o estudo propôs estender o modelo RFM a partir de um parâmetro adicional CI (denominado Contador de Itens – referente à variedade de produtos adquiridos pelos clientes), que pelas conclusões do autor não produziu qualquer diferença nos resultados das segmentações. Entretanto, a afirmação do autor não corresponde aos resultados obtidos. Afinal, apesar do estudo ter encontrado a

mesma quantidade de *clusters*, foi possível observar que a quantidade de registros por agrupamento variou, o que por si só já era motivo para considerar a alteração nos resultados.

De todo modo, dando continuidade, a publicação de (Hu e Yeh, 2014) sugeriu desenvolver um novo algoritmo para descobrir os padrões de frequências, com base em análises RFM, onde não houvesse qualquer informação sobre a identificação dos clientes. Enquanto que a publicação de (Hosseini *et al.*, 2010) recomendou a criação de um novo modelo, unindo o método WRFM (Liu e Shih, 2005) – que determina os pesos relativos das variáveis R, F e M na avaliação do valor da vida útil do cliente – ao algoritmo *k-means*, cujo resultado foi capaz de avaliar a lealdade dos cliente nas estratégias de marketing projetadas.

Já o artigo de (Olson e Chae, 2012) fez um estudo sobre algumas variações do modelo RFM para demonstrar que esta análise pode ser melhorada a partir do uso das técnicas de mineração de dados. Ao mesmo tempo, (Miguéis *et al.*, 2012) propôs dois modelos para detecção de *Churn* (taxa de evasão de clientes) a partir das categorias dos primeiros produtos adquiridos pelos clientes. Neste caso, segundo o autor, ambos os modelos superaram o padrão do modelo RFM para realizar a predição da taxa de evasão dos clientes. Abordagem parecida foi apresentada um pouco antes por (Chen *et al.*, 2009), que sugeriu uma estrutura de segmentação de padrões para gerar informações valiosas sobre o comportamento de compra dos clientes, propondo um novo algoritmo denominado RFM-*Apriori*.

Outra publicação interessante foi produzida por (Coussement *et al.*, 2014), que fez um comparativo entre os modelos RFM, regressão logística e o algoritmo de árvore de decisão CHAID (Kass, 1980), para avaliar a performance de cada um no processo de segmentação de clientes. Estudo semelhante também foi realizado anteriormente por (Mccarty e Hastak, 2007) com os mesmos algoritmos. Já os estudos de (Wu *et al.*, 2009) e (Chang *et al.*, 2010) avaliaram o algoritmo *k-means* em conjunto com o modelo RFM; assim como o artigo (Cheng e Chen, 2009), que neste caso, foi aperfeiçoado com o uso do algoritmo LEM2 (Grzymala-Busse, 1997). E para completar, o artigo de (Wang, 2010) que utilizou técnicas de agrupamento *fuzzy*, com o algoritmo *Fuzzy c-means* (Bezdek, 1981), para desenvolver uma abordagem híbrida que fosse capaz de detectar rapidamente os *outliers* e segmentar os clientes de maneira mais eficiente.

Sendo assim, é exibida a seguir na Tabela 2.1 uma listagem com os principais artigos utilizados pela tese (ordenados por ano de publicação em ordem decrescente) e suas respectivas técnicas ou algoritmos utilizados.

Tabela 2.1: Principais artigos utilizados pela tese ordenados por ano de publicação.

<b>Autor(es)</b>	<b>Ano</b>	<b>Técnica(s) / Algoritmo(s)</b>	<b>Título</b>
(Coussement <i>et al.</i> )	2014	<i>Logistic Regression e Decision Trees</i>	<i>Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees</i>
(Hu e Yeh)	2014	<i>Frequent Pattern e Association Rules</i>	<i>Discovering valuable frequent patterns based on RFM analysis without customer identification information</i>
(Wei <i>et al.</i> )	2013	<i>k-means e SOM</i>	<i>Customer relationship management in the hairdressing industry: An application of data mining techniques</i>
(Miguéis <i>et al.</i> )	2012	<i>Logistic Regression e Sequence Mining</i>	<i>Modeling partial customer churn: On the value of first product-category purchase sequences</i>
(Olson e Chae)	2012	<i>Decision Tree, Logistic Regression e Neural Networks</i>	<i>Direct marketing decision support through predictive customer response modeling</i>
(Shim <i>et al.</i> )	2012	<i>Association Rules e Sequential Patterns</i>	<i>CRM strategies for a small-sized online shopping mall based on association rules and sequential patterns</i>
(Khajvand <i>et al.</i> )	2011	<i>k-means</i>	<i>Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study</i>

(Chang <i>et al.</i> )	2010	<i>k-means e Spectral Clustering</i>	<i>Using K-means method and spectral clustering technique in an outfitter's value analysis</i>
(Hosseini <i>et al.</i> )	2010	<i>k-means</i>	<i>Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty</i>
(Wang)	2010	<i>Fuzzy c-means</i>	<i>Apply robust segmentation to the service industry using kernel induced fuzzy clustering techniques</i>
(Chen <i>et al.</i> )	2009	<i>Sequential Patterns e Apriori</i>	<i>Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data</i>
(Cheng e Chen)	2009	<i>k-means</i>	<i>Classifying the segmentation of customer value via RFM model and RS theory</i>
(Wu <i>et al.</i> )	2009	<i>k-means</i>	<i>Applying RFM Model and K-Means Method in Customer Value Analysis of an Outfitter</i>
(Mccarty e Hastak)	2007	<i>Logistic Regression e Decision Trees</i>	<i>Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression</i>

Estas foram as principais publicações utilizadas e que fundamentaram o conhecimento necessário entre a relação do modelo RFM e as técnicas de mineração de dados. Entretanto, duas outras publicações merecem ser citadas pela revisão do tema RFM, sendo elas: (Wei *et al.*, 2010) e (Kahan, 1998). Não obstante, outros artigos também auxiliaram no complemento das pesquisas, como os descritos por (Zorrilla e García-Saiz, 2013), (Demirkan e Delen, 2013), (Olejnik *et al.*, 2009) e (Delen e Demirkan, 2013), já que abordaram a aplicação das técnicas de mineração de dados a partir de um modelo de negócio como serviço.

Quanto ao tema da computação em nuvem, vale destaque para os seguintes artigos: (Marston *et al.*, 2011) que tratou do assunto sobre uma perspectiva de negócio, identificando os pontos fortes, os pontos fracos, as oportunidades e as ameaças do setor; (Buyya *et al.*, 2009) e (Zhang *et al.*, 2010) que passaram uma visão geral a respeito do ambiente computacional em nuvem; (Sinha e Khreisat, 2014), (Sun *et al.*, 2011),

(Schroepfer *et al.*, 2013) e (Rao e Selvamani, 2015) que abordaram dois dos principais desafios desta tecnologia, que são a segurança e a privacidade dos dados; (Tajadod *et al.*, 2012), (Gannon *et al.*, 2014) e (Barnes, 2015) que discutiram o ambiente a partir das plataformas oferecidas por grandes fornecedores de tecnologia; e (Al-Aqrabi *et al.*, 2015) que examinou o futuro do BI (*business intelligence*) a partir da perspectiva de uso do ambiente de computação em nuvem.

Por fim, vale destacar que as pesquisas mais recentes apontam para o uso prático do modelo RFM, ao descreverem a sua utilização em diversos setores a partir de casos de estudos, como: na avaliação de clientes de companhias aéreas (Wong e Wei, 2018); na avaliação de usuários de aplicativos móveis (Liu *et al.*, 2017); na previsão de avaliações de hotéis on-line (Hu *et al.*, 2017); no setor de saúde (Mohammadzadeh *et al.*, 2017); na indústria varejista de alimentos (Peker *et al.*, 2017); e até mesmo em um hospital veterinário (Wei *et al.*, 2016). Tudo isso só reforça a abrangência interdisciplinar que possui o tema da tese, assim como a importância pela busca de contexto, ou seja, a aplicação prática do assunto tratado.

Já em relação às novas abordagens, chamam atenção os estudos propostos por (Song *et al.*, 2017), que recomenda uma análise de série temporal para explorar as relações dos clientes a partir de um intervalo de tempo, com o intuito de segmentá-los para descobrir as relações quantitativas no modelo RFM; e por (Khodabandehlou e Rahman, 2017), que fornece uma nova abordagem para a segmentação de clientes com base nas mudanças feitas pelos consumidores ao longo do tempo em relação aos seus respectivos comportamentos de compra. Ambos sugerem que acompanhar o histórico temporal de cada grupo de clientes, e seus respectivos comportamentos de compras, possui uma grande relevância para a melhoria do modelo RFM.

Contudo, nenhum artigo abordou de forma clara uma das principais motivações das empresas, o lucro. E quando tentam, o fazem de forma equivocada ao tratar o parâmetro M (de monetário) como se fosse a rentabilidade, conforme visto nos artigos de (Bhensdadia e Kosta, 2010), (Ait Daoud *et al.*, 2016) e (Dursun e Caber, 2016), que tratam como rentável os clientes com a maior pontuação da classificação RFM tradicional (ex.: 5-5-5), ou seja, não levam em consideração a própria lucratividade do consumidor, apenas o valor monetário, o que fortalece a justificativa principal da proposta apresentada por esta pesquisa.

### 2.2.3 Contribuição

O processo de contribuição desta tese começou bem antes da produção de todo este conteúdo. Pois inicialmente, a primeira pesquisa relacionada ao que seria o ponto de partida para este estudo foi avaliar a eficácia de um serviço de aprendizado de máquina com base nos recursos da computação em nuvem. Entretanto, para fazer esta análise era preciso encontrar e definir um problema significativo para que a performance da plataforma e dos resultados encontrados pudessem ser mensurados.

Porém, na busca por um problema real e que pudesse ser resolvido com o uso destes métodos, surgiu – de forma despreziosa – um artigo muito interessante (Wei *et al.*, 2013) que descrevia o uso das técnicas de mineração de dados em conjunto com o modelo RFM para a gestão de relacionamento com os clientes. Esta foi a chave para todo conteúdo e contexto que daria origem a este estudo. Assim, foi desenvolvido o primeiro trabalho (Carvalho, 2015), apresentado no congresso *CILAMCE 2015: XXXVI – Ibero-Latin American Congress on Computational Methods in Engineering*, que orientou todo este processo de análise, ao aplicar o desenvolvimento do modelo RFM com ajuda de uma plataforma de aprendizado de máquina na nuvem.

Em seguida, na tentativa de aprofundar o conhecimento e aumentar a contribuição desta pesquisa, a tese passou a focar em uma das principais motivações para as empresas: a lucratividade; já que pela proposta original do modelo RFM este atributo é ignorado por completo. Assim, o processo de análise passou a ser enriquecido pela criação de modelos RFMP, com o claro objetivo de apresentar uma alternativa à criação dos modelos tradicionais RFM, para avaliar se a inclusão de um novo parâmetro traria algum impacto no processo de segmentação dos clientes. Tal parâmetro ficou diretamente associado à lucratividade dos consumidores, e foi definido pela letra P – da palavra inglesa *Profitability*.

Deste modo, o estudo deu início às avaliações que iriam compor o domínio da tese ao desenvolver diversas análises para compreender a geração dos modelos e distinguir o comportamento de separação entre os *clusters*. Porém, antes disso foi preciso definir uma forma de identificação do melhor valor para o parâmetro  $K$  do algoritmo *k-means*. Para isso, foi criada uma premissa para auxiliar na busca por este valor: encontrar dentro do conjunto de dados a maior quantidade possível de segmentos, mas desde que o

número de registros contidos em um determinado agrupamento não fosse menor do que 1% dos registros totais.

Então, o estudo passou a executar a pesquisa empírica para avaliar a criação de diversos modelos a partir da plataforma adotada pela tese. Porém, logo em seguida emergiu um novo desafio: que era encontrar uma maneira para mensurar a qualidade dos modelos e suas respectivas assertividades em relação aos modelos RFM e RFMP. Pois do contrário, o estudo não teria como comparar os modelos entre si, e muito menos mensurar a consistência e a qualidade dos *clusters* e modelos produzidos.

Assim – como contribuição da tese para o desenvolvimento de modelos RFM e RFMP mais apurados – foi necessário a criação de três novos índices de mensuração que pudessem avaliar os resultados obtidos, sendo: o primeiro atrelado à qualidade individual de cada *cluster* produzido (em correspondência com a classificação RFM ou RFMP associada); o segundo, para mensurar a qualidade média dos *clusters* gerados pelos modelos; e por último, e mais importante, para determinar a qualidade e a assertividade geral do modelo.

E desta forma, ao utilizar estes índices em seu processo de análise, o estudo conseguiu dar sequência a sua execução para avaliar e comparar os modelos entre si através dos seguintes critérios: o melhor valor de  $K$  produzido; se o menor *cluster* gerado atendia à premissa estabelecida; e se os resultados encontrados pelas taxas dos novos índices de mensuração sugeridos estavam coerentes com as classificações RFM ou RFMP atribuídas. Além disso, as análises propiciaram avaliar o comportamento dos *clusters*, observando não só os seus respectivos conteúdos, mas como também os critérios adotados para a separação entre eles. Neste caso, o estudo complementou os resultados com a criação de uma tabela auxiliar para ajudar no entendimento final de cada modelo.

Em consequência, a pesquisa conseguiu identificar novos desafios relacionados não só ao conjunto de dados adotado, como também à forma tradicional pela qual os modelos RFM eram classificados (quando utilizados em conjunto com os métodos de *Machine Learning*). Vale destacar que todos estes desafios foram fundamentados pelos valores dos índices de mensuração propostos, o que demonstrou a importância da criação e utilização das métricas sugeridas.

Por conta disso, a tese propôs um conjunto de recomendações para tentar reduzir os problemas identificados, aumentar a qualidade dos modelos e melhorar a assertividade dos *clusters* em relação às classificações RFM ou RFMP, recomendando assim as seguintes mudanças: variar o valor do parâmetro  $K$  do algoritmo *k-means* (a partir do índice Taxa de Acertos RFM ou RFMP) para encontrar aquele que iria produzir o melhor resultado; e agrupar os dados em intervalos de classes para melhor representar a estrutura do negócio. Além disso, a pesquisa também precisou propor um novo método para a classificação dos modelos RFM e RFMP – em substituição à forma tradicional utilizada pela metodologia padrão dos modelos RFM.

Com esta nova proposta, a tese objetivou aumentar o grau de confiança das classificações RFM e RFMP, e melhorar a consistência entre os *clusters* e seus respectivos conteúdos. Deste modo, o estudo recomendou uma nova metodologia para que tanto os *clusters* quanto os clientes não fossem mais classificados apenas como positivos ou negativos a partir de um determinado valor absoluto (no caso o valor da média global dos atributos), mas sim pela classificação a partir da criação de três faixas de valores, contendo: os Clientes Neutros (°); os Clientes Negativos (-); e os Clientes Positivos (+).

Por fim, a partir desta nova metodologia de classificação, a pesquisa conseguiu examinar se as recomendações sugeridas reduziriam os efeitos causados por cada um dos desafios identificados e se trariam alguma melhoria nos resultados dos modelos produzidos. Além disso, concluiu se a criação dos novos modelos RFMP, em contrapartida aos modelos RFM, enriqueceu o processo de análise ao explorar se a inclusão do parâmetro  $P$  (*Profitability*) – relacionado à lucratividade – traria algum impacto na segmentação dos clientes. Tudo isso, através do uso de uma plataforma de *Machine Learning* em um ambiente computacional em nuvem.

## 3 Conjunto de Dados

### 3.1 Apresentação

A base de dados utilizada por esta pesquisa foi extraída a partir do histórico de transações de um pequeno shopping virtual – mais conhecido como *marketplace* (portal de comércio eletrônico que agrupa vários vendedores em uma só plataforma). Este *e-commerce* está no ar há mais de 4 anos e possui mais de 18 mil registros transacionais de pedidos e mais de 13 mil clientes cadastrados. Porém, neste estudo foram avaliados somente os registros referentes aos pedidos efetivados entre julho de 2016 a junho de 2017, e específicos de uma única empresa que opera neste portal de vendas. Assim, a base inicial analisada ficou com um total de 2.336 clientes e 2.676 pedidos efetivados.

Para a criação dos modelos iniciais o estudo definiu no processo de extração de dados os respectivos valores de R, F, M e P da seguinte forma:

- R – definido pela diferença em dias do registro da data de compra mais recente de cada cliente;
- F – medido pela quantidade de compras realizadas pelo cliente dentro do período de análise do estudo;
- M – calculado pelo valor do tíquete médio de cada cliente;
- P – definido pelo valor médio percentual da lucratividade de cada cliente.

### 3.2 Estatísticas Básicas – Dados Originais

Nesta primeira fase do estudo foi elaborada uma breve avaliação das estatísticas básicas do conjunto de dados. A finalidade deste processo foi fazer uma verificação inicial sobre a qualidade das informações para saber se de fato elas seriam representativas para a pesquisa. Sendo assim, foi realizada uma avaliação exploratória que permitiu um maior entendimento sobre o conjunto de dados que foi usado no processo de criação e interpretação dos modelos.

De início, a primeira tarefa deste processo verificou a existência de valores incompletos ou ausentes. Porém, como os dados foram compilados a partir do histórico

de transações dos pedidos realizados pelos clientes, todos os registros continham as informações necessárias, e por isso não foi preciso aplicar nenhum tratamento aos dados nesta etapa.

Em seguida, foram criados gráficos e tabelas para melhor descrever as informações que seriam submetidas ao processo de análise. Neste ponto vale ressaltar que o ambiente *Azure ML* possui mecanismos tanto para o tratamento quanto para a limpeza dos dados, além de recursos para a criação das estatísticas básicas. Neste sentido, todos os gráficos e tabelas produzidos nesta etapa foram criados a partir da própria plataforma, o que permitiu uma compreensão direta dos dados dentro do ambiente utilizado para o desenvolvimento do estudo, sem a necessidade de uso de qualquer outro programa computacional para a realização deste processo (um ponto positivo sobre a plataforma utilizada). De todo modo, a Tabela 3.1 a seguir descreve as estatísticas básicas do conjunto de dados analisado.

Tabela 3.1: Dados sobre o número de clientes e pedidos.

	<b>R (Recency)</b>	<b>F (Frequency)</b>	<b>M (Monetary)</b>	<b>P (Profitable)</b>
<b>Valor Mínimo</b>	1	1	9,10	7,28%
<b>Valor Máximo</b>	365	11	2.480,00	100,00%
<b>Média</b>	157,33	1,14	237,60	44,49%
<b>Mediana</b>	149,5	1	206	43,69%
<b>Moda</b>	12	1	329,99	42,86%
<b>Desvio Padrão</b>	101,57	0,61	199,42	3,98

Como pode ser visto, os valores mínimo e máximo do atributo R ficaram entre 1 e 365 dias respectivamente, ou seja, existem clientes que fizeram sua última compra há 1 dia, e outros que fizeram sua última compra há 1 ano. Já o valor médio foi de 157,33 dias (o que representa um pouco mais de 5 meses), e a mediana ficou em 149,5 dias, o que representa um valor bem próximo da média. O que chamou a atenção na análise deste atributo foi o valor do desvio padrão, que foi de 101,57, o que indica uma grande dispersão dos dados.

Em relação ao atributo F, foi possível observar que os valores mínimo e máximo ficaram entre 1 e 11 respectivamente, ou seja, existem clientes que fizeram apenas uma única compra dentro do período de análise e outros que chegaram a comprar 11 vezes no período (quase 1 compra por mês). Já o valor médio deste atributo foi de 1,14, um número

muito baixo e que deve ser melhor explorado mais à frente, já que a mediana foi de apenas 1. Em relação ao desvio padrão o valor encontrado foi de 0,61. Este valor indica uma amostragem mais concentrada neste atributo.

Quanto ao atributo M, os valores mínimo e máximo ficaram entre R\$ 9,10 e R\$ 2.480,00 respectivamente. Já o valor médio dos gastos realizados pelos clientes foi de R\$ 237,60, com mediana em R\$ 206,00. O desvio padrão foi de 199,42, o que também pode indicar uma grande dispersão dos dados que serão analisados.

Por último, a análise do atributo P. Os valores mínimo e máximo ficaram entre 7,28% e 100%. O valor médio da rentabilidade foi de 44,49%, com mediana em 43,69%, ou seja, muito próximo ao valor da média. Já o desvio padrão de 3,98% indica uma relação mais uniforme entre os dados.

A etapa seguinte avaliou a distribuição de frequência dos valores de cada variável a partir da criação dos seus respectivos histogramas. O intuito aqui foi analisar a variação dos valores encontrados e com qual frequência eles ocorriam. A representação gráfica do histograma de cada variável pode ser vista na Tabela 3.2.

Cabe destacar que estes gráficos foram produzidos dentro do próprio ambiente utilizado, e que cada histograma foi gerado com suas respectivas curvas de distribuição acumulativa (a curva esverdeada do gráfico) e de densidade (a curva de coloração rosa do gráfico), outro ponto positivo da ferramenta. Além disso, o próprio recurso já distribui o conjunto de dados em intervalos de classes. Este intervalo pode ser parametrizado, e por padrão já vem definido com um valor igual a 10.

Entretanto, uma característica negativa do recurso foi que, em alguns casos, por conta da grande variação dos dados, as medidas do eixo X exibidas pelo histograma tiveram os seus valores transformados em uma notação científica, ou seja, os números foram exibidos com uma notação exponencial, o que dificultou a leitura em um primeiro momento.

Tabela 3.2: Gráficos com os histogramas de cada variável R, F, M e P.

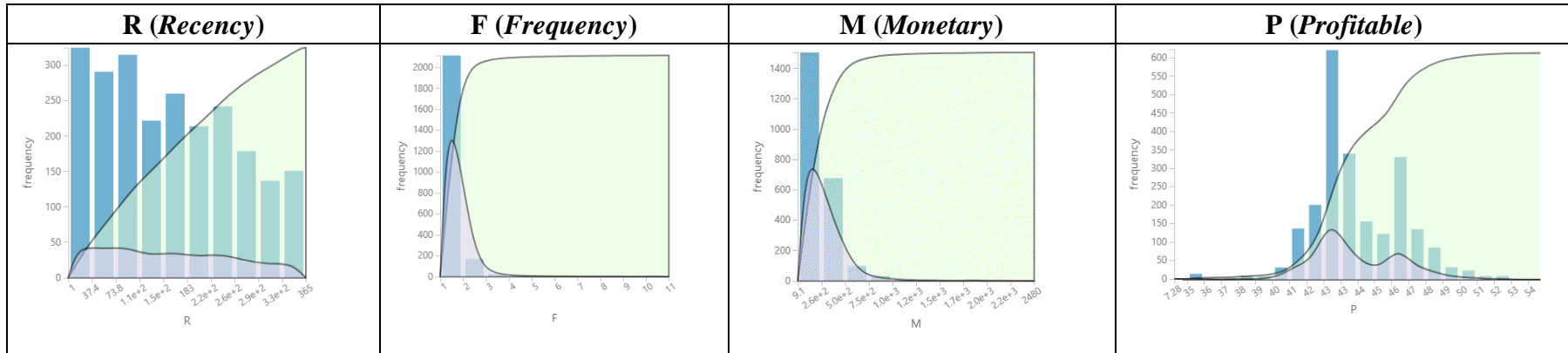
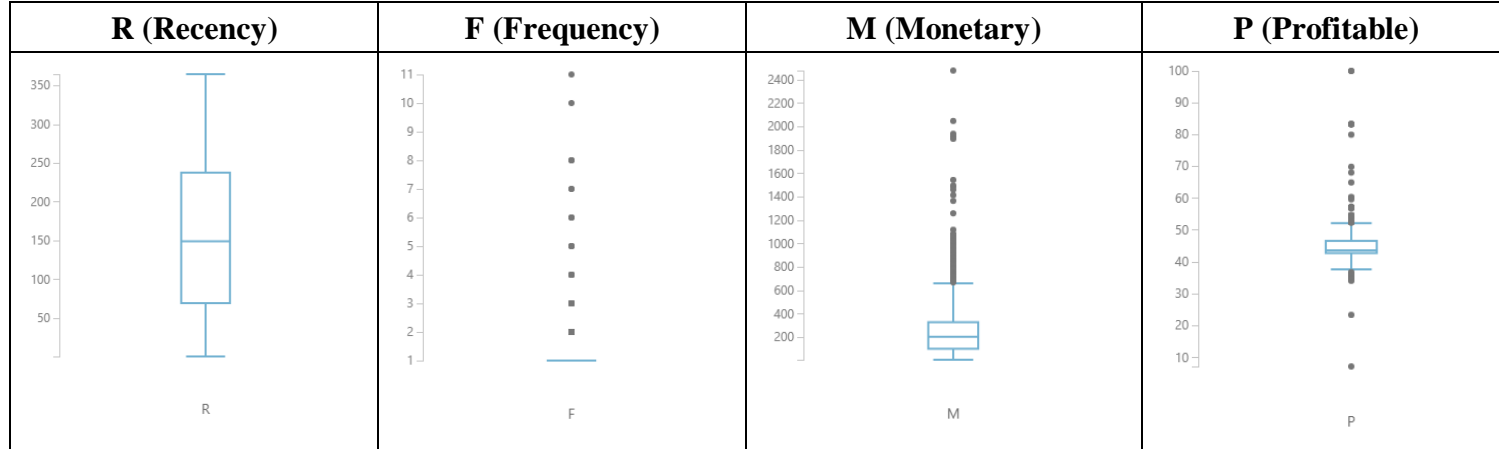


Tabela 3.3: Gráficos *blox-plot* das variáveis R, F, M e P.



Mas para solucionar esta questão bastou passar o cursor do *mouse* em cima do valor exponencial que logo foi mostrado o valor real no gráfico. Além disso, ao passar o cursor do *mouse* sobre cada gráfico foi possível verificar também a quantidade de registros contidos em cada barra do histograma e o percentual que aquela quantidade representava no conjunto de dados.

Após uma breve pausa para descrever mais um recurso da plataforma, o estudo volta sua atenção para de fato analisar os resultados obtidos nos histogramas. E como pode ser visto, o gráfico do atributo R contém uma grande variação na quantidade de valores distintos, sendo este o atributo mais balanceado de todos.

Já o gráfico do atributo F contém uma concentração de registros muito grande em um único valor. Este é o atributo com a maior assimetria quando comparado aos demais, o que poderá afetar na criação dos modelos que serão gerados. O atributo M também possui uma distribuição concentrada, porém um pouco menos distorcida quanto à do atributo F.

E por último, há a distribuição do atributo P. Neste ponto, vale ressaltar que – para uma melhor visualização gráfica dos dados – o estudo optou por aumentar o intervalo das classes do histograma para que ele pudesse refletir melhor os valores contidos neste atributo. Assim, diferente dos demais, o atributo P teve a sua distribuição plotada no histograma com um intervalo de classes maior do que 10. Deste modo, foi possível observar na imagem contida dentro da Tabela 3.2 que o atributo P possui uma distribuição um pouco mais balanceada.

Em seguida, o estudo passou a analisar o gráfico *box-plot* de cada variável. A ideia era verificar os percentis de cada atributo e avaliar a existência ou não de valores discrepantes. Afinal, a existência de *outliers* pode impactar negativamente a interpretação dos dados no processo de criação dos modelos. Neste caso, a Tabela 3.3 ilustra os gráficos *box-plot* dos atributos utilizados por esta análise. Mais uma vez, vale observar que a geração destes gráficos foi realizada diretamente dentro do próprio ambiente *Azure ML*.

Toda via, conforme observado, foi possível perceber que não houve qualquer *outlier* na variável R. Entretanto, este comportamento não foi compartilhado com as demais variáveis do estudo, já que todos os demais atributos possuíam um grande número de valores anômalos.

Mas neste ponto, nenhum outro gráfico chamou tanta atenção quanto o gráfico do atributo F. Nele foi possível verificar uma grande concentração de registros com valores iguais a 1, e que qualquer outro valor além deste foi considerado um *outlier*. Definitivamente este comportamento da variável F é algo crônico e terá um impacto direto na elaboração dos modelos.

Quanto aos gráficos das variáveis M e P foi possível notar a existência de valores discrepantes em ambos. No atributo M, notou-se que os *outliers* surgiram a partir de uma banda superior do gráfico; já em relação ao atributo P, foi possível encontrar *outliers* tanto na banda superior quanto na banda inferior do gráfico.

Ainda assim, tendo em vista todos estes *outliers* apresentados, e mesmo com a noção sobre as influências que eles poderiam ter na geração dos modelos, o estudo não fez neste momento nenhum tratamento para reduzi-los ou removê-los. Isto porque estes valores podem ser importantes no processo de determinação sobre um evento raro no conjunto de dados. Assim, os dados foram imputados no processo de análise e criação de modelos da maneira como foram extraídos do banco de dados, ou seja, com os *outliers* e sem nenhum tratamento específico.

Dando continuidade ao estudo, o passo seguinte foi verificar o resultado obtido a partir da matriz de correlação entre todas as variáveis de entrada. Entretanto, não foi possível encontrar qualquer funcionalidade existente dentro da plataforma que pudesse produzir esta matriz. Por isso, o estudo recorreu a um recurso que permitiu a execução de *scripts* feitos em linguagem *Python* (mas poderia ter sido em linguagem *R* também) para gerar tanto a tabela quanto o gráfico desta matriz. Assim, o resultado produzido pode ser observado na Tabela 3.4 e no gráfico exibido pela Figura 3.1.

Tanto a tabela quanto o gráfico produzido funcionam como uma forma qualitativa de se mostrar a correlação linear entre as variáveis. Neste caso, quanto mais próximo de -1 for o valor de uma relação entre dois atributos, maior será o grau de correlação negativa; e quanto mais próximo de +1, maior será o grau de correlação positiva.

Dessa forma, tanto pelo modelo gráfico (Figura 3.1) quanto pelos valores descritivos (Tabela 3.4) foi possível notar que o conjunto de dados analisado não possui

correlação entre suas variáveis, o que indica uma independência entre os atributos analisados.

Tabela 3.4: Correlação entre as variáveis de entrada utilizadas pelo estudo.

	<b>R</b>	<b>F</b>	<b>M</b>	<b>P</b>
<b>R</b>	1	0,146136	0,044378	0,001022
<b>F</b>	0,146136	1	0,071423	0,023625
<b>M</b>	0,044378	0,071423	1	0,071519
<b>P</b>	0,001022	0,023625	0,071519	1

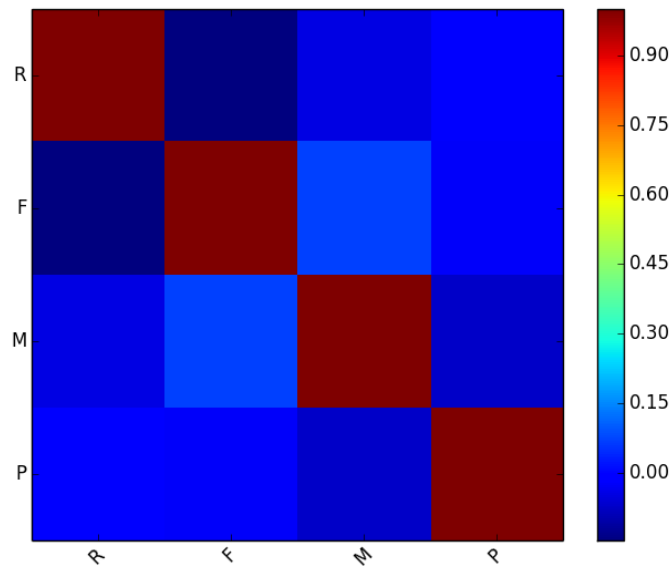


Figura 3.1: Gráfico que contém a matriz de correlação entre as variáveis de entrada utilizadas pelo estudo.

Por último, foi criado o gráfico de projeção para verificar não só a associação entre as variáveis, como também para ajudar a identificar se haveria separação linear entre os atributos. Porém, apesar da plataforma possuir um recurso com tal função, o estudo optou por gerar uma única imagem que pudesse unir os gráficos de todos os atributos para facilitar a interpretação. Neste caso, a imagem foi gerada através de um *script* desenvolvido em linguagem *Python*, e o seu resultado pode ser visto logo a seguir.

Observando os gráficos contidos na Figura 3.2 foi possível perceber uma grande aglutinação dos valores existentes entre as relações das variáveis, o que permitiu também visualizar alguns *outliers*. Fora isso, nenhuma correlação foi encontrada, confirmando o que foi feito na análise anterior.

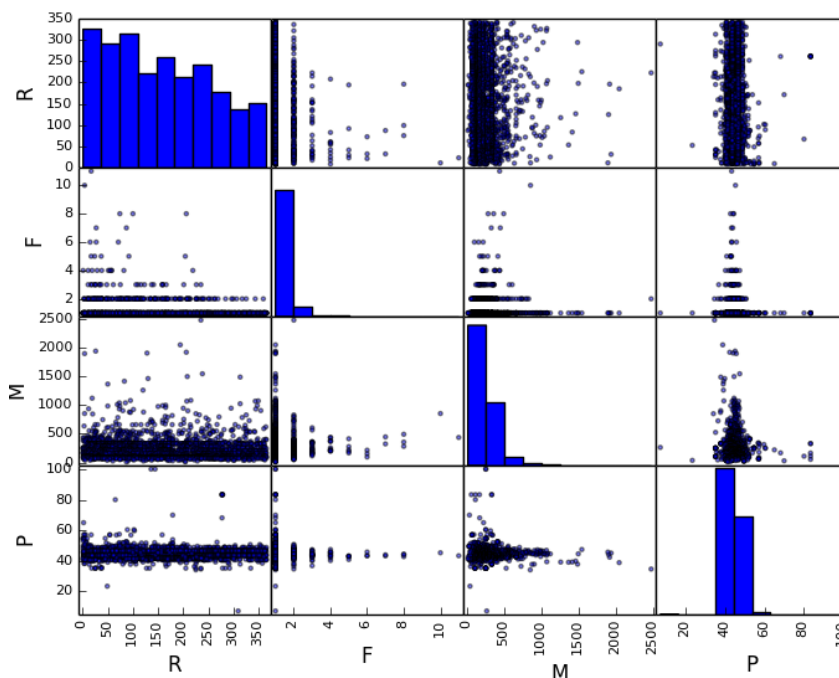


Figura 3.2: Gráfico de projeção entre as variáveis de entrada utilizadas pelo estudo.

Sendo assim, o resultado obtido pela análise das estatísticas básicas foi considerado satisfatório, pois permitiu um melhor entendimento sobre as variáveis e os seus respectivos valores, ou seja, como estavam organizados, como estavam distribuídos e como se relacionavam entre si. Tudo isso propiciou um melhor entendimento sobre a qualidade dos dados que serão trabalhados nas próximas etapas do estudo.

Neste ponto, já é possível ter uma ideia de que os dados são relevantes, mas em alguns casos estão bastante concentrados em algumas faixas de valores, isto é, com baixa distribuição e grande assimetria. Este fato para uma tarefa de classificação não supervisionada pode ser um grande problema, já que pode impactar diretamente nos resultados dos modelos.

De todo modo, a pesquisa encerra esta fase do estudo – que teve por objetivo analisar as estatísticas básicas relativas ao conjunto de dados – e segue adiante com o processo de desenvolvimento.

## 4 Desenvolvimento da Tese

### 4.1 Processo de Desenvolvimento

Como dito anteriormente, neste primeiro experimento foram analisados os dados RFM e RFMP relacionados às vendas efetivadas entre julho de 2016 a junho de 2017, com um total de 2.676 pedidos, feitos por 2.336 clientes distintos.

Os pedidos feitos por clientes recorrentes representam 21% da base de dados analisada, ou seja, 561 pedidos foram realizados por clientes que fizeram mais de 1 pedido dentro deste período. Já os demais 2.115 pedidos foram efetivados por clientes não recorrentes (aqueles que dentro do período de análise fizeram apenas 1 compra). Entretanto, apenas 9% dos clientes são recorrentes, o que causou uma grande distorção (assimetria) no atributo F e que pode afetar os modelos que serão criados. A Tabela 4.1 descreve melhor esta relação entre a frequência de compras dos clientes recorrentes e não recorrentes com os seus respectivos pedidos.

Tabela 4.1: Relação da frequência de compras e clientes recorrentes.

Frequência (quantidade de pedidos)	Total de clientes com o padrão de frequência	Total de pedidos realizados	Percentual
11	1	11	0,41%
10	1	10	0,37%
8	3	24	0,90%
7	2	14	0,52%
6	3	18	0,67%
5	5	25	0,93%
4	11	44	1,64%
3	25	75	2,80%
2	170	340	12,71%
1	2.115	2.115	79,04%
<b>TOTAL</b>	<b>2.336</b>	<b>2.676</b>	<b>100,00%</b>

Além da assimetria do atributo F, vale ressaltar também que a inicialização do algoritmo *k-means* é muito sensível ao conjunto de dados e ao ajuste dos seus parâmetros, pois de acordo com os valores escolhidos – e da ordem de entrada do conjunto de dados

– um resultado diferente poderá ser gerado. Este impacto pode ser observado pela simples mudança na ordem de entrada dos registros, como no caso de um atributo que inicialmente é ordenado de forma crescente, e que depois – em uma nova execução – pode ser ordenado em ordem decrescente.

Por isso foi necessário manter ao máximo o padrão de inicialização do algoritmo para que os grupos criados sofressem a menor influência possível quanto à mudança de inicialização do algoritmo. Neste caso, os registros foram ordenados da seguinte forma:

- **Parâmetro F** – em ordem decrescente;
- **Parâmetro M** – em ordem decrescente;
- **Parâmetro R** – em ordem crescente.

De todo modo, apesar desta ordenação, não há qualquer garantia que futuras análises e modelos sejam inicializados com os mesmos registros de entrada.

Sendo assim, o próximo passo do trabalho foi definir a métrica que iria balizar a performance do estudo. Neste caso, o objetivo era encontrar o melhor número de  $K$  (*clusters*) para o conjunto de dados analisado. Porém, como é notório no algoritmo *k-means*, este método não garante que o número de *clusters* seja de fato o melhor a ser trabalhado, uma vez que este número deve ser definido previamente. E sendo definido previamente, o grande desafio é justamente encontrar o melhor número de  $K$  antes de se iniciar uma determinada análise.

Esta é uma questão muito importante para aqueles que desejam trabalhar com classificação não supervisionada. Pois se de fato o que se está tentando descobrir é justamente a quantidade de *clusters* de um conjunto de dados a partir dele próprio, de que maneira esta quantidade deve ser definida? Com que métrica? Afinal, está é uma dificuldade se comparada com as técnicas de classificação supervisionada ou de regressão, já que estes processos possuem métricas bem definidas para a avaliação dos seus modelos.

Entretanto, várias pesquisas já foram propostas para ajudar na identificação do melhor valor de  $K$  a ser utilizado no algoritmo *k-means*, e uma boa compilação de muitas destas propostas pode ser encontrada nos artigos descritos em (Milligan e Cooper, 1985) e (Mirkin, 2011). Porém, fica muito difícil compará-las já que cada uma das propostas

emprega uma forma específica como critério de performance. Contudo, um outro ponto interessante e pouco discutido é que elas não tratam a formação dos *clusters* a partir de uma orientação que faça algum sentido para as estratégias de negócios que serão utilizadas através da segmentação dos dados.

Por isso, o estudo utilizou como critério inicial de busca pelo melhor valor de  $K$  uma abordagem orientada ao negócio como forma de resolução para este problema. Sendo assim, foi utilizada a seguinte premissa para a formulação da quantidade de *clusters*: buscar dentro do conjunto de dados a maior quantidade possível de segmentos, mas desde que o menor *cluster* não ficasse com menos de 1% dos registros totais.

O objetivo era criar a maior quantidade possível de segmentos de clientes, com base nos seus padrões de compras, para que eles fossem trabalhados em futuras estratégias de marketing, seja por campanhas de emails, redes sociais, anúncios patrocinados e até mesmo por telefone. O limite de 1% se deu pelo volume dos dados, pois um grupo com um volume de registros de clientes menor do que este seria muito pequeno para se trabalhar em uma campanha específica. Porém, o estudo deixa claro que tal valor percentual poderá variar de acordo com o conjunto de dados a ser trabalhado em outras análises.

E foi assim, a partir desta premissa, que o estudo deu início ao seu desenvolvimento funcional através da plataforma de aprendizado de máquina em nuvem escolhida para o desenvolvimento da tese. A seguir são apresentados os detalhes da execução e os resultados obtidos neste primeiro experimento:

O primeiro passo foi carregar o conjunto de dados para dentro da plataforma. Neste ponto nenhuma dificuldade foi encontrada, até porque o conjunto de dados não era muito grande. Isto foi feito em uma só etapa, a partir de um arquivo no formato *ARFF*. A única limitação encontrada foi que, uma vez feito o envio do arquivo, não foi possível encontrar uma forma de renomeá-lo. Consegue-se até atualizar os dados, mas renomear o conjunto não foi possível.

Em seguida, foi feita uma transformação através de comandos SQL para garantir o ordenamento e a sequência da entrada de dados conforme descrito anteriormente, garantindo assim que todos os modelos fossem inicializados da mesma forma. Apesar de ser possível fazer este ordenamento fora da ferramenta, ou seja, subir o arquivo com o

padrão desejado, foi muito bom saber que havia uma maneira para se criar comandos SQL dentro da própria plataforma, já que este recurso pode ajudar na manipulação dos dados e facilitar a persistência para uso em outros ambientes ou modelos.

Todavia, os recursos de SQL disponíveis são limitados, pois o que a ferramenta possui é apenas uma versão mais enxuta destes comandos. Vale também ressaltar um ponto de melhoria: é que ele só informa quando há um erro após a tentativa de execução, e com uma descrição bem limitada. Neste caso, como sugestão, ele poderia mostrar o erro de sintaxe antes da execução.

Dando continuidade, a próxima etapa teve por objetivo criar a maior quantidade de modelos para comparar os resultados obtidos com a premissa estabelecida anteriormente. Esta tarefa poderia levar a uma execução bem trabalhosa, mas logo de início foi encontrado um método bastante interessante – denominado *Sweep Clustering* – que ajudou bastante no decorrer de toda a pesquisa. É que este recurso permite inferir o melhor número de  $K$  através de alguns índices existentes na plataforma.

Assim, este recurso tem por objetivo a construção e testes de vários modelos a partir de diferentes configurações de parâmetros, como por exemplo, um intervalo de número de *clusters* estimado e a escolha de uma métrica de precisão (método matemático usado para estimar o ajuste do modelo; no caso, o melhor valor para  $K$ ). Então, através das configurações informadas, o recurso executa o seu processo de construção e testes de modelos iterativos até que seja encontrado aquele com o melhor conjunto de *clusters* com base na métrica selecionada.

Estas métricas são usadas para medir a similaridade entre os *clusters* e a precisão dos centroides. Porém, dentro do ambiente *Azure ML* há como alternativa as seguintes opções: *Simplified Silhouette* (Hruschka *et al.*, 2004), *Davies-Bouldin* (Davies e Bouldin, 1979), *Dunn* (Dunn†, 1974) e *Average Deviation* (Michael *et al.*, 1999).

Ainda assim, apesar de ser um excelente recurso prático, vale destacar que cada uma destas métricas possui o seu próprio critério de performance, ou seja, não traz uma resposta definitiva para a solução do problema de identificação do melhor valor de  $K$  para o algoritmo *k-means*. Elas apenas ajudam na tentativa de se buscar uma resposta a partir de um critério justificável. Por conta disso, o estudo fez uso destas métricas apenas como ponto de partida para encontrar o número de  $K$  mais apropriado de acordo com a premissa

estabelecida, ou seja, encontrar o maior número de *clusters*, desde que o *cluster* com a menor quantidade de registros não ficasse com menos de 1% do total dos dados.

Por isso, o estudo criou um modelo específico para cada métrica existente (quatro no total) e configurou os parâmetros de quantidade de *clusters* entre um intervalo de 2 a 20 *clusters*. Desta maneira foram criados 19 modelos por métrica (executados com o valor de *K* variando entre 2 e 20 *clusters* – inclusive). Como foram testadas 4 métricas por execução, um total de 76 modelos foram criados por análise. Porém, vale ressaltar que o estudo analisou somente o melhor modelo gerado por cada métrica.

Outro ponto importante e que deve ser informado é que a inicialização do algoritmo não foi efetuada com nenhuma escolha de parâmetros randômicos. Esta é uma forma de garantia para o estudo, pois assegura assim a replicação dos modelos caso necessário. De todo modo, os parâmetros utilizados nesta análise podem ser vistos na Figura 4.1.

▲ K-Means Clustering

Create trainer mode  
Parameter Range ▼

Range for Number of Cent... ☰  
 Use Range Builder  
2 - 20

Initialization for sweep  
K-Means++ ▼

Random number seed ☰  
[Empty text box]

Number of seeds to sweep ☰  
[Empty text box]

Metric ☰  
Euclidean ▼

Iterations ☰  
100

Assign Label Mode ☰  
Ignore label column ▼

Figura 4.1: Parâmetros de customização do algoritmo *k-means* no ambiente *Azure ML*.

A descrição de cada parâmetro e suas respectivas opções de valores são apresentadas logo em seguida:

- **Create trainer mode:** especifica como o modelo deverá ser treinado, ou seja, se com um valor específico de parâmetros e sem variação (*Single Parameter* – para uso do *k-means* clássico) ou com o uso do *Sweep Clustering* (*Parameter Range*) para utilizar vários parâmetros e permitir que o recurso encontre a configuração ideal. Como descrito anteriormente, o estudo fez o uso deste último recurso, e por isso utilizou o valor *Parameter Range*;
- **Range for Number of Centroids:** intervalo com o número de *clusters* que o algoritmo iria trabalhar. O valor utilizado nesta análise ficou entre 2 e 20;
- **Initialization for Sweep:** escolha do algoritmo de inicialização dos centroides. Opções de parametrização:
  - *First N* – também chamado de método *Forgy*, onde um número inicial de pontos é escolhido no conjunto de dados e usado como o meio inicial;
  - *Random* – também chamado de método de partição aleatória, onde o algoritmo coloca aleatoriamente um ponto de dados em um *cluster*, para em seguida calcular a média inicial para ser o centroide dos pontos atribuídos aleatoriamente ao *cluster*;
  - *K-Means++* – método padrão para inicialização dos *clusters*, aprimora os meios utilizados pelo algoritmo *k-means* tradicional ao usar um método próprio para escolher os centroides dos *clusters* iniciais. Este foi o valor utilizado nas análises;
  - *K-Means++ rápido* – uma variante do algoritmo *K-means++*, porém mais otimizado;
  - *Evenly* – onde os centroides estão localizados equidistantes uns dos outros no espaço de *D*-Dimensional de *N* pontos de dados.
- **Random number seed:** parâmetro opcional, define um valor a ser usado para a inicialização de cada *cluster*. Este parâmetro não foi utilizado nesta análise;

- **Number of seeds to Sweep:** número total de valores aleatórios a serem usados como pontos de partida para inicialização dos *clusters*. Este parâmetro também não foi utilizado por esta análise;
- **Metric:** métrica usada para calcular a distância entre os *clusters*. Opções de parametrização:
  - Euclidiana – a distância Euclidiana normalmente é usada como uma medida de dispersão dos *clusters*. Essa métrica é preferencial porque minimiza a distância média entre pontos e os centroides. Esta foi a opção utilizada nas análises;
  - Cosseno – como alternativa é possível utilizar a função cosseno para medir a similaridade dos *clusters*. A similaridade por cosseno é útil em casos onde não se importa com o comprimento de um vetor, somente com seu ângulo.
- **Iterations:** número de iterações a ser usado pelo algoritmo. O valor utilizado nesta análise foi igual a 100;
- **Assign label mode:** caso haja alguma coluna no arquivo de dados que contenha algum rótulo, é possível utilizá-la para orientar a seleção dos *clusters* ou especificar que os valores sejam ignorados. Este parâmetro não foi utilizado por esta análise.

Quanto aos parâmetros do recurso *Sweep Clustering* e suas opções de valores, eles podem ser vistos na Figura 4.2 e são descritos logo a seguir:

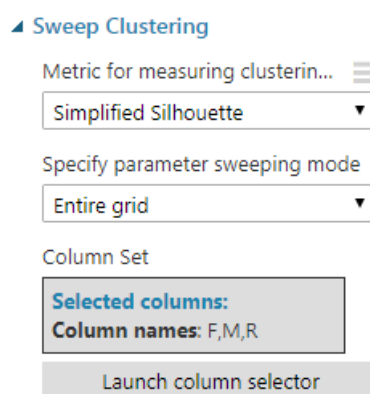


Figura 4.2: Parâmetros de customização do recurso *Sweep Clustering*.

- **Metric for measuring Clustering:** especifica o método matemático a ser usado para estimar o ajuste do modelo (*Simplified Silhouette*, *Davies-Bouldin*, *Dunn* e *Average Deviation*);
- **Specify parameter sweeping mode:** se a execução será realizada com todos os valores especificados (*Entire grid*) ou com uma escolha randômica (*Random Sweep*) a partir dos valores parametrizados. O valor utilizado nesta análise foi o *Entire grid*;
- **Column Set:** escolha dos campos do conjunto de dados que serão usados na análise.

Todo este desenvolvimento explicado até aqui pode ser visto resumidamente – de forma ilustrativa – na Figura 4.3, pois ela representa exatamente o processo de execução destas tarefas dentro do ambiente *Azure ML*.

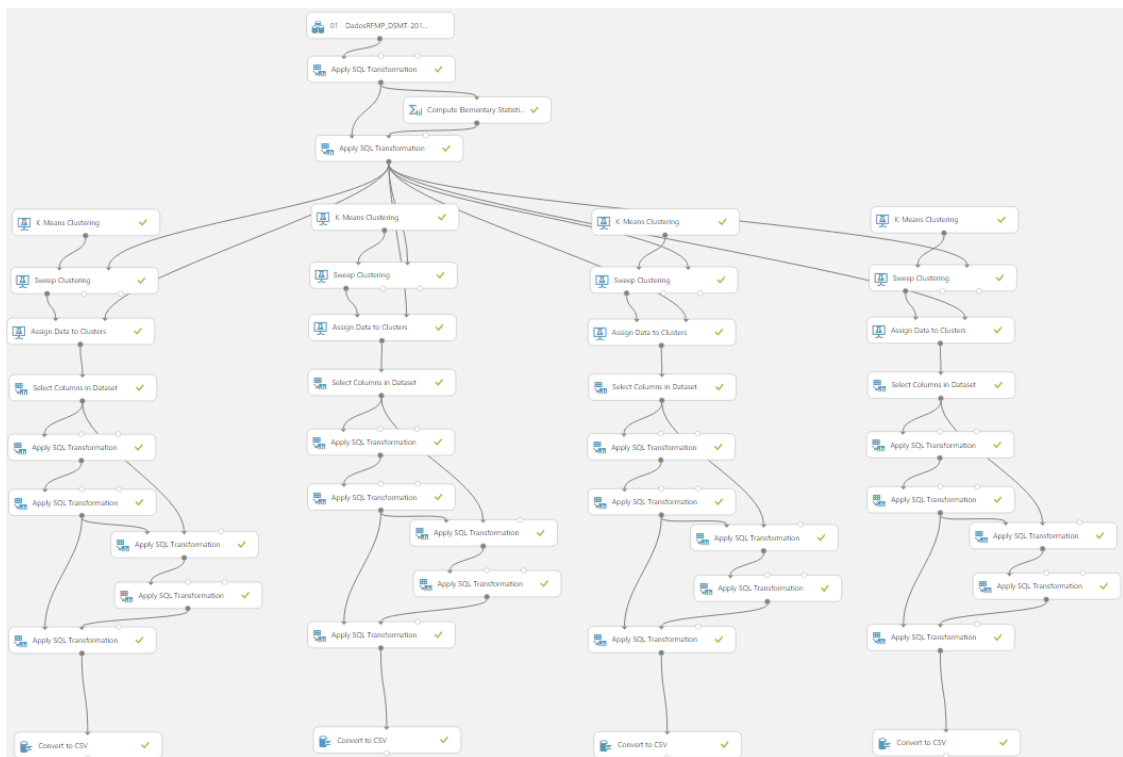


Figura 4.3: Processo de desenvolvimento e execução dos modelos.

Nesta figura é possível observar no topo uma caixa que representa o conjunto de dados analisado; em seguida, a execução das consultas SQL e do cálculo das estatísticas básicas. Na sequência, são observadas quatro raias, cada uma com a execução do algoritmo *k-means* a partir de uma métrica definida pelo recurso *Sweep Clustering*. Ao

final de cada uma delas, um encadeamento de comandos SQL até a produção final dos resultados que gerou a tabela auxiliar de avaliação do estudo, que foi exportada através de um arquivo no formato *CSV* (outro recurso bastante útil da ferramenta).

Explicado o processo de desenvolvimento, a análise segue sua execução para dar continuidade à pesquisa. Mas antes, uma observação: toda vez que for encontrada a notação RFM/P, o estudo estará tratando sempre dos modelos RFM e RFMP, ou seja, será sobre algo que se aplica aos dois modelos.

## 4.2 Processo de Avaliações dos Modelos

Definidos os valores de parametrização e o formato de inicialização do algoritmo, o estudo passou a analisar os resultados gerados pelo melhor modelo criado a partir de cada métrica utilizada. Deste modo, a pesquisa avaliou um único resultado por métrica e fez uma comparação entre eles para averiguar o melhor desempenho e se atendia ao critério de performance definido anteriormente.

Mas antes disso, vale descrever as formas de avaliação disponibilizadas pelo ambiente de execução e que foram utilizados pela tese:

- **Gráfico de PCA** – exibe a formação de cada *cluster* gerado a partir de um determinado modelo. Nele é possível visualizar as diferenças entre os *clusters* e observar se são separados linearmente ou se há sobreposição entre eles. Desta forma, este gráfico calcula dois eixos utilizando componentes principais para resumir as diferenças multidimensionais entre os *clusters*; assim, as atribuições de cada *cluster* são exibidas ao longo desses eixos;
- **Gráfico com a Quantidade de Registros** – exibe a quantidade de registros contida em cada um dos *clusters*;
- **Histograma do Modelo** – exibe um histograma com a frequência acumulada em cada *cluster*.

Entretanto, como os resultados dos modelos de segmentação não são perfeitos, o estudo gerou uma tabela auxiliar – com os detalhes dos resultados de cada *cluster* – para complementar o processo de avaliação e ajudar a entender o conteúdo de cada grupo, ou seja, como eles se comportaram e de que maneira eles se separaram em relação aos

demais. A apresentação dos grupos nesta tabela será exibida em ordem decrescente pela coluna que contém a quantidade de registros por *cluster*, e irá conter os seguintes campos:

- **Grupo** – identificador de cada *cluster*;
- **Tamanho** – quantidade de registros por *cluster*;
- **%** – percentual de registros por *cluster*;
- **RFM+** – identificação positiva dos atributos RFM de cada *cluster*;
- **RFM-** – identificação negativa dos atributos RFM de cada *cluster*;
- **RFMP+** – identificação positiva dos atributos RFMP de cada *cluster*;
- **RFMP-** – identificação negativa dos atributos RFMP de cada *cluster*;
- **Média R** – valor médio do atributo R contido em cada *cluster*;
- **Min R** – valor mínimo do atributo R contido em cada *cluster*;
- **Max R** – valor máximo do atributo R contido em cada *cluster*;
- **Média F** – valor médio do atributo F contido em cada *cluster*;
- **Min F** – valor mínimo do atributo F contido em cada *cluster*;
- **Max F** – valor máximo do atributo F contido em cada *cluster*;
- **Média M** – valor médio do atributo M contido em cada *cluster*;
- **Min M** – valor mínimo do atributo M contido em cada *cluster*;
- **Max M** – valor máximo do atributo M contido em cada *cluster*;
- **Média P** – valor médio do atributo P contido em cada *cluster*;
- **Min P** – valor mínimo do atributo P contido em cada *cluster*;
- **Max P** – valor máximo do atributo P contido em cada *cluster*;
- **% Acertos R** – percentual de acertos do atributo R. Neste caso, se um *cluster* for considerado como R+, será calculado um percentual com todos os

registros que contenham o atributo R menor ou igual ao valor da média global. Caso contrário, ou seja, se o *cluster* for considerado como R-, será calculado um percentual com todos os registros que contenham o atributo R maior que o valor da média global. Isto acontece porque quanto menor for o valor de R, mais recente é o cliente. Assim, um valor menor é considerado positivo, enquanto que um valor maior é considerado negativo;

- **% Acertos F** – percentual de acertos do atributo F. Neste caso, se um *cluster* for considerado como F+, será calculado um percentual com todos os registros que contenham o atributo F maior ou igual ao valor da média global. Caso contrário, ou seja, se o *cluster* for considerado como F-, será calculado um percentual dos registros que contenham o atributo F menor que o valor da média global. Isto acontece porque quanto maior for o valor de F, mais frequente ou leal é o cliente. Assim, um valor maior é considerado positivo, enquanto que um valor menor é considerado negativo;
- **% Acertos M** – percentual de acertos do atributo M. Neste caso, se um *cluster* for considerado como M+, será calculado um percentual com todos os registros que contenham o atributo M maior ou igual ao valor da média global. Caso contrário, ou seja, se o *cluster* for considerado como M-, será calculado um percentual com todos os registros que contenham o atributo M menor que o valor da média global. Isto acontece porque quanto maior for o valor de M, maior é o gasto do cliente. Assim, um valor maior é considerado positivo, enquanto que um valor menor é considerado negativo;
- **% Acertos P** – percentual de acertos do atributo P. Neste caso, se um *cluster* for considerado como P+, será calculado um percentual com todos os registros que contenham o atributo P maior ou igual ao valor da média global. Caso contrário, ou seja, se o *cluster* for considerado como P-, será calculado um percentual com todos os registros que contenham o atributo P menor que o valor da média global. Isto acontece porque quanto maior for o valor de P, maior é a lucratividade do cliente. Assim, um valor maior é considerado positivo, enquanto que um valor menor é considerado negativo.

### 4.3 Métricas de Avaliações dos Modelos

Além do processo de avaliação que foi discutido no tópico anterior, o estudo propõe a criação de três novos índices de mensuração para avaliar os resultados obtidos por cada modelo. Estes índices foram desenvolvidos para suprir a ausência de indicadores existentes no processo de criação de modelos RFM/P. Afinal, sem isto, o estudo não teria como comparar os modelos entre si, e muito menos mensurar a consistência e a qualidade dos *clusters* e modelos produzidos.

Vale destacar que estes índices foram elaborados a partir da criação da tabela auxiliar gerada com base nos resultados de cada modelo. A seguir são apresentadas as descrições referentes a cada um destes novos índices:

O primeiro índice está associado à qualidade individual de cada *cluster* produzido em correspondência com a classificação RFM/P atribuída. Este índice de avaliação irá depender do tipo de análise, ou seja, se é uma análise RFM ou RFMP, conforme descrito a seguir:

- **% Acertos RFM** – percentual de acertos em todos os atributos RFM, ou seja, se um registro contido em um determinado *cluster* foi classificado de forma correta nos três atributos analisados seguindo as regras apresentadas anteriormente nos campos: % Acertos R, % Acertos F e % Acertos M, isto é, se a classificação do registro é igual à do seu respectivo *cluster*; ou
- **% Acertos RFMP** – percentual de acertos em todos os atributos RFMP, ou seja, se um registro contido em um determinado *cluster* foi classificado de forma correta nos quatro atributos analisados seguindo as regras apresentadas anteriormente nos campos: % Acertos R, % Acertos F, % Acertos M e % Acertos P, isto é, se a classificação do registro é igual à do seu respectivo *cluster*.

O segundo índice irá mensurar a qualidade média dos *clusters* gerados pelos modelos, conforme apresentado na sequência:

- **Taxa Média de Acertos dos Clusters** – somatório do percentual da taxa de acertos RFM ou RFMP de todos os *clusters* (% Acertos RFM ou % Acertos RFMP), dividido pelo total de *clusters* obtidos por cada modelo.

E por último, e mais importante, o índice que irá determinar a qualidade e a assertividade geral do modelo:

- **Taxa de Acertos RFM ou RFMP do Modelo** – percentual com os acertos de todos os registros do conjunto de dados que foram classificados corretamente pela classificação RFM ou RFMP quando comparados às classificações dos seus respectivos *clusters*. Como esta taxa representa o somatório de acertos do modelo, ela não será apresentada na tabela auxiliar, mas sim de forma descritiva ao final de cada análise.

Com estas premissas, o estudo segue adiante com sua pesquisa empírica para avaliar os diversos resultados produzidos através da criação de vários modelos.

# 5 Desenvolvimento dos Modelos

## 5.1 Análise RFM

Explicados todos os critérios utilizados para avaliação dos modelos, o estudo passa agora a analisar em detalhes os resultados encontrados, que são apresentados logo a seguir.

### 5.1.1 *Simplified Silhouette*

A execução realizada com o recurso *Sweep Clustering* computa os valores de qualidade que indicam a pontuação alcançada por cada modelo através da métrica selecionada. Como resultado, a ferramenta disponibiliza uma tabela ordenada pela melhor pontuação obtida por cada modelo e o número de centroides utilizado na execução, conforme pode ser visto na Tabela 5.1. Entretanto, vale ressaltar que para a métrica *Simplified Silhouette*, quanto maior o valor da pontuação, melhor é o modelo.

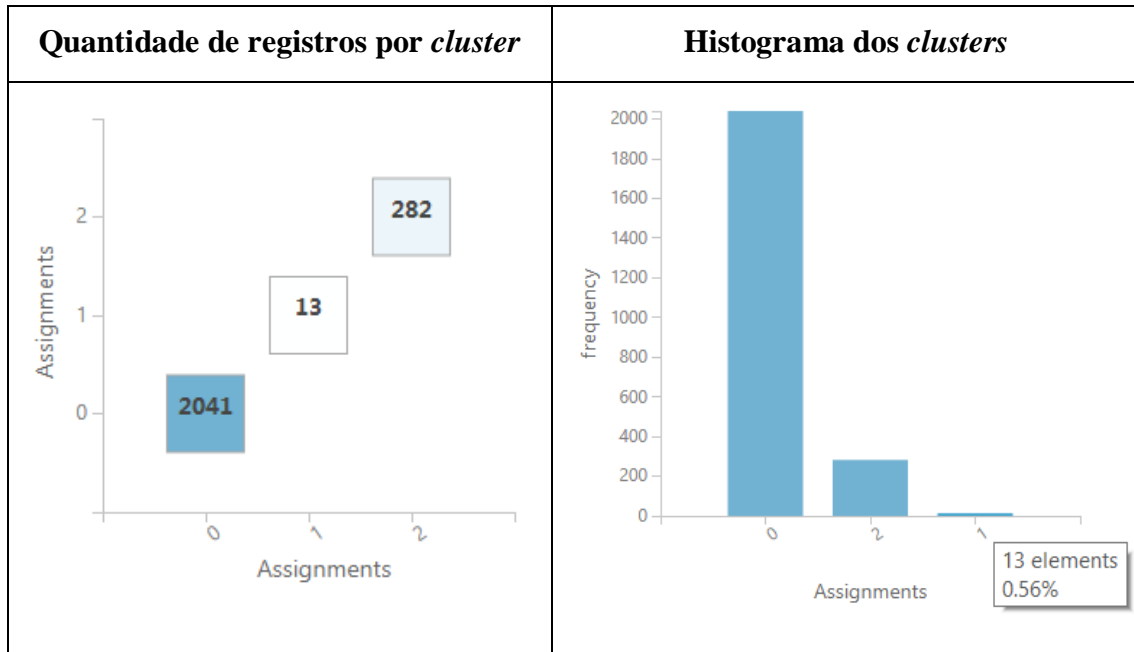
Tabela 5.1: Pontuação da métrica *Simplified Silhouette* da análise RFM.

Pontuação	Número de <i>Clusters</i>	Pontuação	Número de <i>Clusters</i>
0,642888	3	0,545174	5
0,63924	2	0,544318	18
0,606748	8	0,544017	20
0,602901	7	0,533451	12
0,587567	9	0,533147	17
0,573255	6	0,525783	14
0,566735	4	0,524665	13
0,548443	19	0,523648	15
0,547631	10	0,515615	16
0,546379	11		

Como pode ser observado, com esta métrica o melhor modelo encontrado produziu somente 3 *clusters* distintos, no qual o maior agrupamento possui 2.014 registros e o menor apenas 13 registros, o que representa apenas 0,56% dos registros e não satisfaz ao critério de performance do estudo. Deste modo, percebe-se que este modelo

concentrou grande parte de todos os registros em um único *cluster*, desbalanceando-o em detrimento dos demais, conforme pode ser visto nos gráficos da Tabela 5.2.

Tabela 5.2: Gráficos dos resultados da métrica *Simplified Silhouette* da análise RFM.



Além disso, como pode ser visto na Figura 5.1, os grupos ficaram sobrepostos, como se um estivesse contido dentro do outro, ou seja, sem nenhuma separação linear entre eles. Isto se deve em parte pelo fato de que o resultado é sempre mostrado por um gráfico de componentes principais, a partir de uma visão bidimensional da representação dos *clusters*.

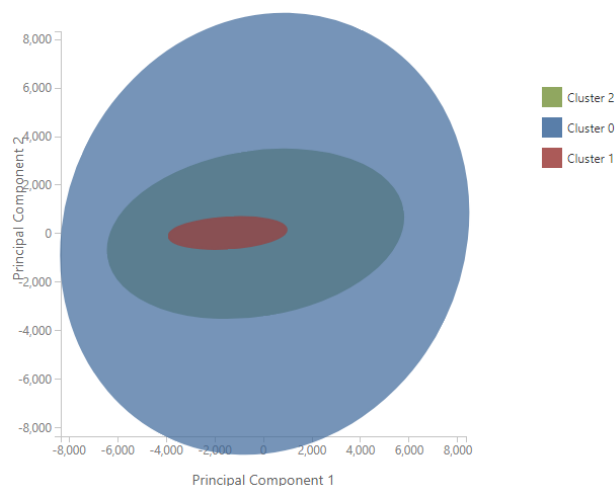


Figura 5.1: Gráfico da visualização dos *clusters* da 1ª análise RFM com a métrica *Simplified Silhouette*.

Tudo isso dificulta a visualização de cada registro por *cluster*, o que significa dizer que não há como compreender de que maneira ficou a distribuição dos dados dentro de cada elipse que representa um determinado agrupamento. Isto é, não há como saber o quão distante está um registro do outro dentro de um mesmo *cluster*; e nem para saber se há uma grande concentração de registro em uma determinada área ou se estão dispersos por todo domínio do *cluster*.

Deste modo, este gráfico permitiu apenas ter uma ideia superficial de como ficou o modelo, o que limitou uma avaliação mais precisa a respeito do resultado produzido por este recurso. Neste ponto, o estudo avaliou como negativa a performance do recurso apresentado pela plataforma utilizada, já que era possível exibir todos os dados analisados dentro do gráfico gerado sem qualquer redução de performance.

Dando continuidade, a próxima etapa desta análise fez uso da tabela complementar produzida pelo estudo para auxiliar no entendimento sobre a formação de cada *cluster*, e o seu resultado pode ser visto na Tabela 5.7. Além disso, esta tabela também permitiu analisar o resultado RFM de cada um dos agrupamentos gerados – mesmo sabendo que este modelo não atendia ao critério de performance estabelecido previamente, já que o menor *cluster* ficou com menos registros do que o estabelecido.

Como pode ser visto, o *cluster* 0 possui 2.041 registros, o que corresponde a 87,37% dos dados. Este grupo foi classificado como RFM-, ou seja, a média do atributo R (158,62) ficou maior do que a média global; já a média do atributo F (1,12) foi menor do que a média global, assim como a do atributo M (183,41). Desta forma, este grupo pode ser considerado com baixo valor para a empresa a partir da classificação do modelo RFM, já que ambos os atributos foram considerados negativos, pois não superaram a média em relação à performance global. De uma forma resumida, os clientes deste grupo não são recentes, não são frequentes e nem possuem um tíquete médio com um alto valor.

Entretanto, ao analisar o campo % de acertos RFM, pode-se verificar que o resultado não foi muito satisfatório, uma vez que a quantidade de registros classificados corretamente em todos os três atributos analisados (R, F e M) não passou de 34,59%. Isto porque só no atributo R o percentual de acertos ficou em 48,11%, enquanto que no atributo F foi de 91,08%, e no atributo M foi de 70,26%.

Seguindo com a análise, foi possível observar que o *cluster 2* possui 282 registros, o que corresponde a 12,07% dos dados. Este grupo foi classificado como RFM+, ou seja, a média do atributo R (148,42) ficou menor do que a média global, a média do atributo F (1,12) ficou maior do que a média global, assim como a do atributo M (183,41). Desta forma, este grupo pode ser considerado um grupo VIP segundo o modelo RFM, já que ambos os atributos foram considerados positivos, pois superaram a média em relação à performance global. Em suma, pela classificação RFM este grupo é composto por clientes recentes, frequentes e com um tíquete médio maior do que a média global.

Mas novamente, ao observar o campo % de acertos RFM, é possível verificar que o percentual de acertos foi de apenas 8,87% dos registros. Um resultado nem um pouco confiável para a classificação deste grupo. Ainda assim, vale notar que o percentual de acertos dos atributos R (54,96%) e do atributo M (100%) foram maiores do que os encontrados na análise do grupo anterior. Porém, o que penalizou o resultado deste grupo foi o atributo F, com um percentual de acertos de apenas 13,48%.

Por último, o grupo 1 que possui apenas 13 registros, com 0,56% do volume total, e que não atende ao critério de performance estabelecido pela pesquisa. De todo modo, este grupo foi classificado como RM+ e F-, ou seja, média de R (148,84) menor do que a média global, média de M (1.260,05) maior do que a média global, e média de F (1,07) menor do que a média global. Com estes valores, os clientes deste grupo foram considerados recentes e monetários; porém, não frequentes.

Do mesmo modo que foi feito anteriormente, é preciso analisar o campo % de acertos RFM, que neste caso foi de 53,85%, a maior taxa entre os três *clusters* analisados. Porém, o que contribuiu para este resultado foram os percentuais de acertos dos atributos R (53,85%), F (92,31%) e M (100%), e claro a baixa quantidade de registros contidos neste *cluster*, já que com um número menor de registros fica mais fácil ter uma taxa de acertos maior.

Feitas as análises individuais de cada *cluster*, e suas respectivas representações do modelo RFM, o estudo passa agora a analisar o comportamento do modelo, isto é, como os *clusters* foram gerados e de que forma foram separados.

Neste contexto, percebe-se que a métrica *Simplified Silhouette* utilizou somente o atributo M para separar os *clusters* entre si. Pois como pode ser visto, a média dos atributos R e F de cada *cluster* ficaram muito parecidas; enquanto que a média do atributo M variou bastante.

Além disso, corrobora para esta afirmação os valores encontrados nos campos Min M e Max M, já que foi possível perceber uma variação de escala entre os *clusters*; de modo que o grupo 0 possui os registros que contêm os valores entre 9,10 e 373,09; enquanto que o grupo 2 possui os registros com valores entre 372,79 (muito próximo ao valor máximo do grupo 0) e 1.119,99; e o grupo 1 possui valor mínimo de 1.260,05 (próximo também ao valor máximo do grupo 2) e máximo de 2.480,00.

Sendo assim, com esta métrica não foi possível encontrar o modelo ideal que pudesse atender aos requisitos de performance do estudo. Além disso, a qualidade, a quantidade e a distribuição dos registros entre os *clusters* encontrados também não foi satisfatória (grande concentração de registros em um único *cluster*). Por tudo isso, a pesquisa descarta de imediato o uso deste modelo. Até porque, a taxa média de acertos dos *clusters* foi de apenas 32,47% (impactada negativamente pelo resultado do *cluster* 2, com apenas 8,87% de taxa de acertos RFM), enquanto que a taxa de acertos RFM do modelo produzido por esta métrica foi de apenas 31,59%.

Desta maneira, é dada como concluída a análise dos resultados produzidos por esta métrica. E assim, o estudo segue adiante com a análise das demais.

### **5.1.2 *Davies-Bouldin***

O melhor modelo gerado a partir desta métrica encontrou um total de 10 *clusters* distintos, conforme pode ser visto na Tabela 5.3. Entretanto, diferente da anterior, para esta métrica quanto menor o valor da pontuação, melhor é o modelo.

Neste caso, como pode ser visto na Tabela 5.4, o maior agrupamento encontrado possui 697 registros, e o menor somente 1 registro (o que representa apenas 0,04% dos registros e não satisfaz ao critério de performance estabelecido). De todo modo, com este modelo o estudo conseguiu perceber uma ligeira melhora na distribuição dos dados. Porém, chama a atenção os três *clusters* com menos de 1% dos registros, o que indica um potencial padrão para a descoberta de um comportamento de clientes atípicos, isto é, os eventos raros de um conjunto de dados. A conferir.

Tabela 5.3: Pontuação da métrica *Davies-Bouldin* da análise RFM.

Pontuação	Número de <i>Clusters</i>	Pontuação	Número de <i>Clusters</i>
0,666951	10	0,75593	17
0,68459	8	0,768643	12
0,689041	3	0,772548	6
0,704428	19	0,796288	15
0,714249	11	0,809666	16
0,719915	18	0,82136	14
0,726493	20	0,822089	13
0,732642	9	0,838731	4
0,735934	2	0,8732	5
0,737832	7		

Já pela análise do gráfico da Figura 5.2 foi observado que os *clusters* continuaram sobrepostos. Porém, se avaliados somente os *clusters* 1, 9 e 2, é possível verificar que há entre eles uma perspectiva de separação linear, com uma ligeira sobreposição em uma parte do *cluster* 9 e outra do *cluster* 2. Novamente, a análise deste gráfico ficou prejudicada pela forma como ele foi exibido, já que inviabilizou identificar onde estavam concentrados os dados em cada parte das elipses que representam os *clusters*.

Seguindo com esta análise, foi feito novamente o uso da tabela auxiliar para uma melhor compreensão do comportamento de cada grupo. Os dados desta tabela podem ser encontrados na Tabela 5.8, e são descritos com mais detalhes logo a seguir.

O *cluster* 6 é o maior de todos, com 697 registros, o que corresponde a 29,84% dos dados. Este grupo foi classificado com RF+ e M-, sendo que a média de R foi de 79,37, a média de F foi de 1,15 e a média de M foi de 129,28. Isto significa dizer que este grupo ficou formado por clientes recentes, frequentes, mas não monetários (tíquete médio com baixo valor). Quanto aos acertos RFM, ele obteve uma taxa de 11,05%, o que representa uma classificação RFM inconsistente, com taxa de acerto de R em 94,26%, de F em 11,48% e de M em 100,00%. Neste caso, nota-se que os resultados do atributo F foram os que mais impactaram negativamente para o péssimo resultado da taxa de acertos RFM do *cluster*.

Tabela 5.4: Gráficos dos resultados da métrica *Davies-Bouldin* da análise RFM.

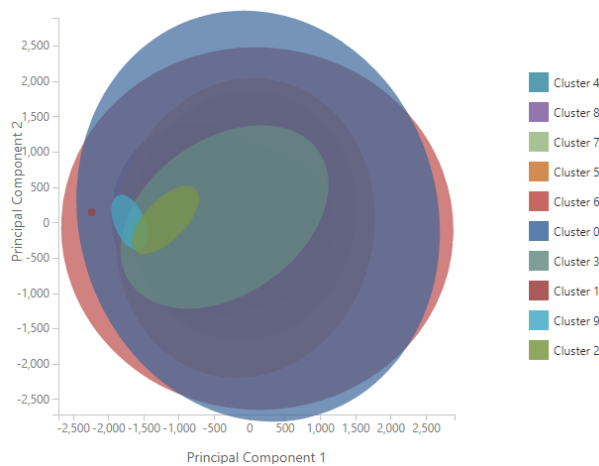
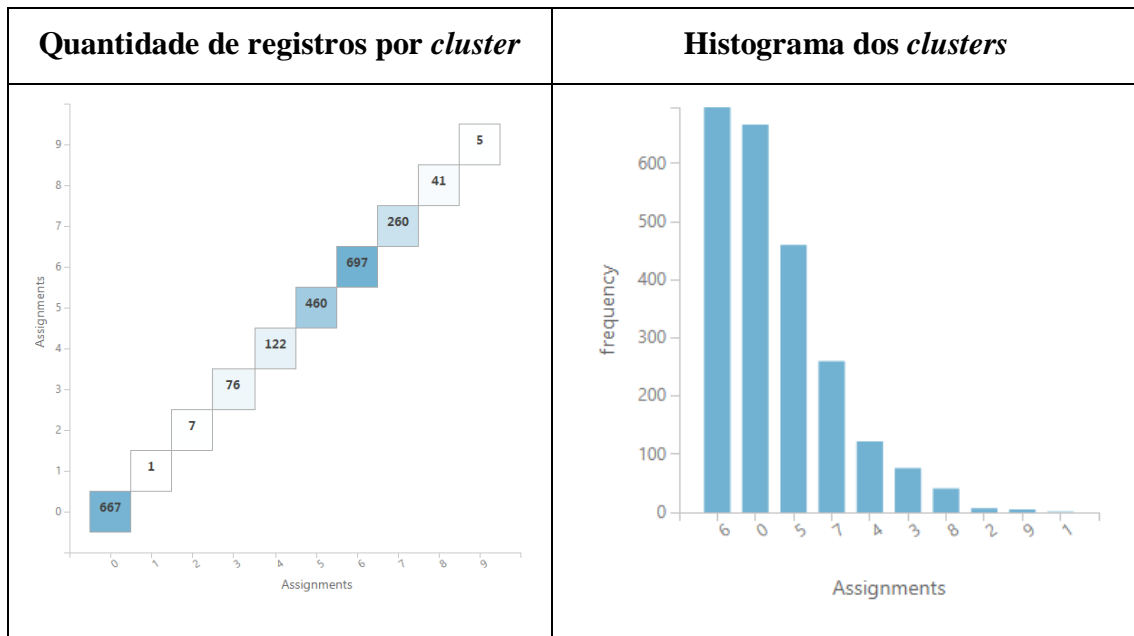


Figura 5.2: Gráfico da visualização dos *clusters* da 1ª análise RFM com a métrica *Davies-Bouldin*.

Já o *cluster* 0, o segundo maior deste modelo, possui 667 registros (28,55% do total) e foi classificado como RFM-, ou seja, os clientes deste grupo não são recentes, nem frequentes, e muito menos monetários. Neste grupo, foram encontradas as seguintes médias de R (255,65), de F (1,04) e de M (122,20). E no que diz respeito aos acertos RFM, este grupo teve taxa de sucesso de 96,70%, com taxa de acerto de R em 100%, de F em 96,70% e de M em 100,00%. Estes resultados representam uma melhora significativa se comparados aos encontrados no grupo anterior, em especial ao do atributo

F. Desta forma, a segmentação RFM atribuída a este segmento pode ser considerada consistente.

O *cluster 5* contém 460 registros (19,69%) e foi classificado como RFM+, o que significa dizer que este é um *cluster* formado por clientes recentes, frequentes e monetários, com média de R em 83,10, de F em 1,23 e de M em 301,02. Em compensação, a taxa de acertos RFM voltou a ficar muito baixa, com um valor de 11,96%, influenciada novamente pelas taxas do atributo F, que foi de apenas 14,78%; enquanto que os atributos R e M tiveram acertos de 90,87% e 88,26% respectivamente. Mais uma vez, os resultados encontrados não foram consistentes em relação à atribuição da classificação RFM.

Quanto ao *cluster 7*, ele possui 260 registros (11,13%) e foi classificado como M+ e RF-, ou seja, são clientes monetários, mas não são recentes e nem frequentes. As médias encontradas neste grupo foram de: R (260,59), F (1,10) e M (312,42). Em relação aos acertos RFM, foi possível encontrar uma taxa de 87,31%, com taxas de R, F e M em 100%, 93,08% e 93,85% respectivamente. Estas taxas mostram que o acerto do atributo F tem sido fundamental para o sucesso da taxa de acertos RFM dos *clusters*.

Em relação ao *cluster 4*, com 122 registros (5,22%), ele foi considerado como RFM+, com média de R em 77,81, de F em 1,39 e de M em 473,96. Sua taxa de acertos RFM foi de 15,57%, tendo sido impactada novamente pelos resultados do atributo F, que obteve uma taxa de 16,39%, enquanto que R obteve 87,70% e M um valor de 100%.

Logo na sequência vem o *cluster 3*, com 76 registros (3,25%) e classificação M+ e RF-, semelhante ao *cluster 7* descrito anteriormente, mas com média de R em 229,96, de F em 1,13 e M em 598,77. A taxa de acertos RFM deste grupo foi de 72,37%, com acertos de R, F e M de: 82,89%, 88,16% e 100,00% respectivamente.

O *cluster 8* com 41 registros (1,76%) e com classificação FM+ e R- foi o próximo a ser analisado. Com esta classificação, este grupo ficou composto por clientes frequentes, monetários, mas não recentes; com média de R, F e M de: 165,80, 1,26 e 930,67 respectivamente. E apesar de ser um *cluster* pequeno, ele obteve apenas 4,88% de taxa de acertos RFM, a menor taxa entre todos os modelos avaliados até agora. Porém neste caso, além do atributo F ter contribuído com uma taxa de 7,32%, o atributo R também teve influência com uma taxa de apenas 53,66%, um valor muito abaixo dos

encontrados anteriormente por esta análise. Já o atributo M, manteve a taxa de acerto em 100%.

Chegando aos 3 últimos *clusters* desta análise, observa-se que cada um deles representou um conjunto muito pequeno dos dados. A começar pelo *cluster* 2, com 7 registros (0,30%), classificação M+ e RF-, com média de R em 163,14, de F em 1,00 e de M em 1.433,59. Já a taxa de acertos RFM ficou com um valor de 42,86%. Neste caso, chama atenção a taxa de acerto de F que foi de 100%; afinal, este grupo contém os clientes com apenas 1 frequência de compra. Já o acerto de R foi de 42,86% – o mais baixo até agora – e o acerto de M foi de 100%.

O *cluster* 9 foi um pouco menor do que o anterior, com 5 registros (0,21%) e classificação RM+ e F-, com média de R em 111,60, de F em 1 e de M em 1.942,41. Já a taxa de acertos RFM ficou com uma taxa de 60%, com uma taxa de 100% de acerto nos atributos F e M, mas com taxa de 60% no atributo R.

Por último o menor *cluster*, e o que chamou mais atenção, o *cluster* 1. Este *cluster* contém somente um único registro que representa apenas 0,04% dos dados. Isto indica um padrão muito fora da curva, um grande *outlier*, pois do contrário ele estaria associado a outro agrupamento. Neste caso, foi encontrado um valor médio de M de 2.480,00, um valor médio de F igual a 2,00, e um valor médio de R de 235,00. Com isto, este registro ficou classificado como FM+ e R-, o que significa um cliente monetário, frequente, mas não recente. De todo modo, a taxa de acertos RFM ficou com 100%.

Encerradas as análises individuais, o estudo analisou o comportamento entre os *clusters* para tentar identificar de que maneira eles foram criados e por quais atributos ou valores eles foram separados.

E logo de início foi possível perceber que os grupos 4, 5 e 6 possuem as médias de R muito próximas, entre 77,81 e 83,10; porém, há uma diferença entre eles no valor da média do atributo M muito grande. Isto porque o modelo separou estes três grupos como se fossem uma relação complementar do valor de M entre eles, isto é, uma escala, variando de 9,10 a 215,34 no *cluster* 6, de 217,01 a 384,99 no *cluster* 5 e de 389,49 a 659,98 no *cluster* 4. Além disso, os resultados encontrados pelas taxas de acertos RFM entre eles também ficaram bem semelhantes, com um percentual de sucesso muito baixo e pouco confiável. E como registrado anteriormente, nota-se que os resultados do atributo

F foram os que mais impactaram negativamente para isto. Assim, os resultados encontrados em cada um destes *clusters* não ficaram consistentes em relação à atribuição da classificação RFM de cada grupo.

Por outro lado, os *clusters* 0, 3 e 7 possuem valores médios do atributo R muito próximos, variando entre 229,61 a 260,59. E assim como na análise anterior, o que os diferenciaram foram os valores do atributo M, já que o grupo 0 possui valores entre 11,90 e 211,00, o grupo 7 entre 217,90 e 450,60, e o grupo 3 entre 461,19 e 784,60. A diferença em relação à análise anterior é que a taxa acertos RFM foi muito maior, justificando em todos casos a classificação RFM associada a cada grupo. E isto se deve exclusivamente à taxa de acerto do atributo F.

Portanto, pela observação destas últimas duas análises fica claro que os *clusters* foram separados pela atribuição do valor do parâmetro M, pois se dependesse somente do atributo R talvez estes 6 *clusters* fossem classificados em apenas 2. Uma outra maneira de se analisar estes *clusters* seria olhar a comparação das médias de M em relação às médias de R. Assim, é possível comparar os *clusters* 0 e 6, já que possuem valores médios de M bem semelhantes (122,10 x 129,28), mas com uma diferença no valor de R, pois o valor do *cluster* 0 é de 255,65 e o do *cluster* 6 é de 79,37. Neste caso, o primeiro grupo é composto por clientes perdidos, enquanto que o segundo é composto por clientes recentes.

Este mesmo comportamento pode ser observado entre os *clusters* 5 e 7, com médias de M de 301,02 e 312,42, e com médias de R de 83,10 e 260,59 respectivamente. Neste caso, o primeiro é composto por clientes recentes, enquanto que o segundo por clientes perdidos. Da mesma forma, os *clusters* 4 e 3 também possuem comportamentos semelhantes a este, já que as médias de M foram de 473,96 e 598,77 (uma variação da média um pouco maior que os outros dois exemplos, mas que dentro do contexto da análise ainda justifica a semelhança entre os grupos), porém com médias de R em 77,81 e 229,96. Sendo assim, conforme observado, em ambos os casos ficou comprovado que a separação entre os *clusters* foi feita pela diferença dos valores do atributo R e não do atributo M.

Já os *clusters* 8, 2, 9 e 1 foram exclusivamente separados pela faixa de valores do atributo M. Juntos, os clientes destes *clusters* possuem 2,31% do total de registros e representam os clientes mais monetários, ou seja, aqueles que possuem um tíquete médio mais alto. Em ambos os casos, o % de Acertos M foi de 100%, o que justifica a

classificação do atributo M pela classificação RFM. Desta forma, é possível concluir que estes são os clientes especiais, com um comportamento diferente dos demais, e que merecem uma atenção dos estrategistas com o intuito de mantê-los fidelizados.

Por último, observa-se que a taxa média de acertos dos *clusters* foi de apenas 50,27%, enquanto que a taxa de acertos RFM do modelo foi de 46,53% – valores bem acima dos encontrados pela métrica anterior, mas ainda assim muito aquém para as expectativas da pesquisa. Desta forma, o estudo dá por encerradas as análises feitas a partir da métrica *Davies-Bouldin* e segue adiante dando continuidade às demais.

### 5.1.3 *Dunn*

O melhor modelo encontrado por esta métrica obteve um total de 10 *clusters*, conforme pode ser visto na Tabela 5.5. Este é o mesmo valor de *K* encontrado na análise da métrica anterior. Assim, o estudo não fará qualquer avaliação, já que os *clusters* e os seus respectivos conteúdos ficaram idênticos.

Mas de qualquer maneira, vale ressaltar que a pontuação do índice encontrada foi diferente, assim como a ordenação do número de *K*. Para esta métrica, quanto maior a pontuação, melhor é o modelo.

Tabela 5.5: Pontuação da métrica *Dunn* da análise RFM.

<b>Pontuação</b>	<b>Número de <i>Clusters</i></b>		<b>Pontuação</b>	<b>Número de <i>Clusters</i></b>
0,717043	10		0,456264	16
0,574956	18		0,456264	15
0,571718	17		0,401157	7
0,564139	19		0,383209	8
0,505996	13		0,373733	9
0,488532	3		0,29973	2
0,486586	14		0,249979	6
0,480615	11		0,248166	4
0,470244	12		0,242115	5
0,460534	20			

Dessa maneira, o estudo segue adiante para avaliar os resultados com a última métrica apreciada.

### 5.1.4 Average Deviation

Com esta métrica o melhor modelo encontrado produziu um total de 20 *clusters* distintos, conforme pode ser visto na Tabela 5.6. Nela, quanto menor a pontuação, melhor é o modelo.

Tabela 5.6: Pontuação da métrica *Average Deviation* da análise RFM.

Pontuação	Número de <i>Clusters</i>	Pontuação	Número de <i>Clusters</i>
41,613956	20	69,426041	10
42,92757	19	69,69846	9
46,384604	18	71,865234	8
51,330497	17	81,8166	7
52,494542	16	82,753018	6
52,571814	15	103,102992	5
53,836602	14	108,92223	4
56,61034	13	137,28337	3
57,251479	12	145,670851	2
62,122759	11		

Pela Figura 5.3 e Figura 5.4 pode-se observar que o maior *cluster* possui 307 registros (13% do total), e o menor somente 1 registro (0,04%). Com esta condição este modelo também não cumpre com o critério de desempenho estipulado pelo estudo. Ainda assim, este foi o modelo com o maior número de *clusters* encontrado e foi o que obteve a melhor distribuição dos registros. Contudo, foram identificados cinco *clusters* com menos de 1% dos dados, totalizando 21 registros, o que ampliou o número de clientes com um possível comportamento atípico.

Quanto ao gráfico da Figura 5.5, foi possível observar que os *clusters* continuaram sobrepostos, mas com um pequeno distanciamento entre alguns, como foi o caso dos *clusters* 1 (em vermelho, à esquerda do gráfico), 9 (também à esquerda, em azul) e o *cluster* 2 (em verde, próximo ao *cluster* 9). Ainda assim, se analisados isoladamente com outros *clusters*, percebe-se que o *cluster* 17 (em verde, à direita do gráfico) e o 10 (em laranja, quase sobreposto pelo *cluster* 17) foram separados linearmente quando comparados aos *clusters* 1, 9 e 2.

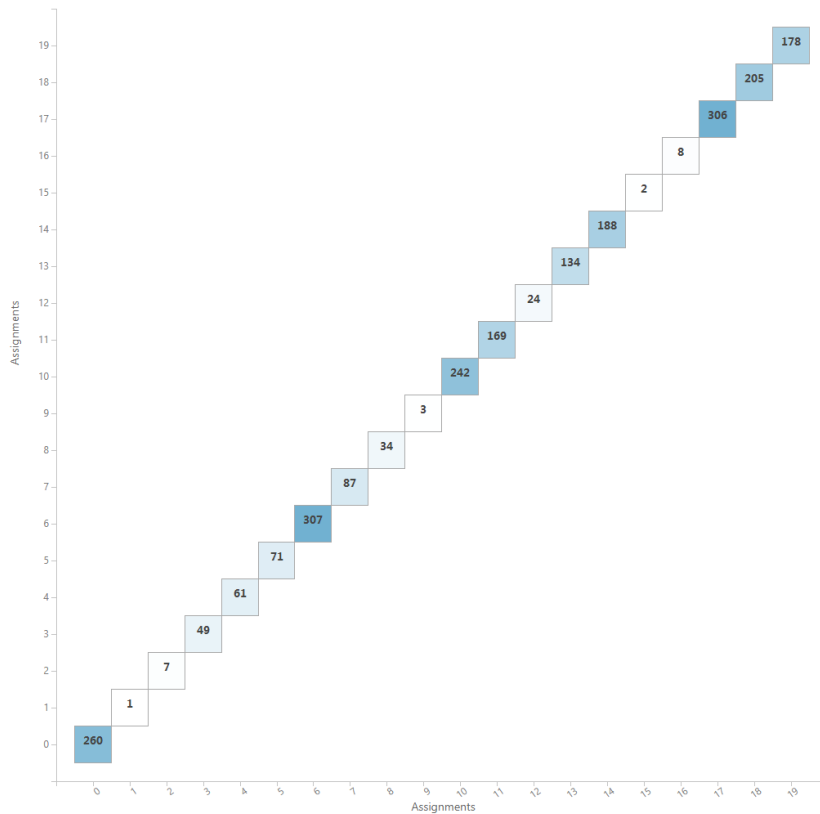


Figura 5.3: Quantidade de registros por *cluster* da 1º análise RFM com a métrica *Average Deviation*.

Aqui vale uma forte crítica ao gráfico gerado pela plataforma, já que a ferramenta poderia exibir cada *cluster* a partir da quantidade de registros em ordem decrescente. Deste modo, seria garantido que um *cluster* menor ficaria sempre à frente de um *cluster* maior, evitando assim que um bloqueasse a visualização do outro por conta da sobreposição.

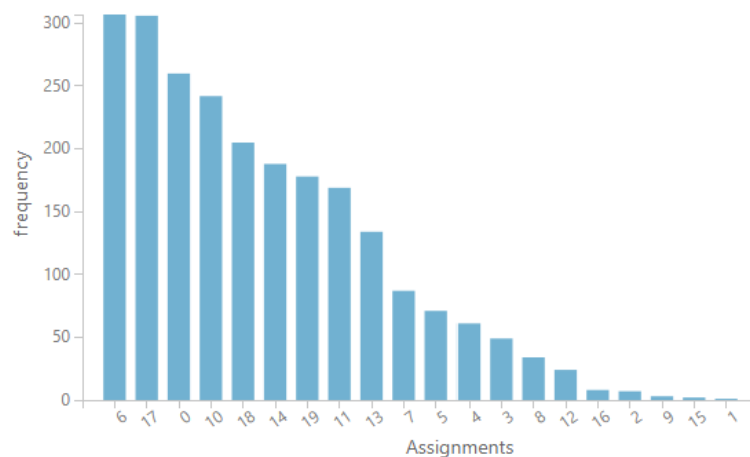


Figura 5.4: Histograma dos *clusters* da 1º análise RFM com a métrica *Average Deviation*.

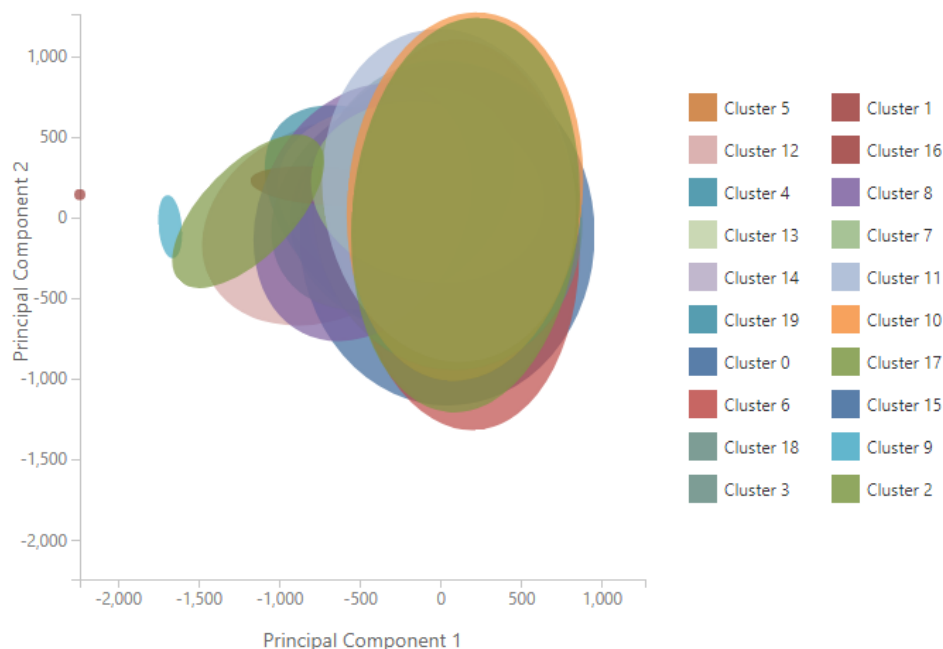


Figura 5.5: Gráfico da visualização dos *clusters* da 1ª análise RFM com a métrica *Average Deviation*.

No gráfico da Figura 5.5 foi possível observar que o *cluster* 16 (em vermelho, posicionado no meio do gráfico, por trás do *cluster* 2 e quase imperceptível) teve a sua forma e posicionamento encobertos por outros *clusters*. Este foi só um exemplo das dificuldades impostas pela ferramenta para interpretar os agrupamentos a partir do gráfico de saída produzido.

De todo modo, o estudo segue com sua análise a partir do uso da tabela auxiliar para obter uma melhor compreensão da formação de cada *cluster*. Os dados desta tabela podem ser encontrados na Tabela 5.9, e são detalhados logo em seguida.

O *cluster* de número 6 – o maior de todos – com 307 registros (13,14%) foi classificado como R+ e FM-, isto é, clientes recentes, mas não frequentes e nem monetários. Neste *cluster*, a média de R foi de 56,25, de F foi de 1,10 e de M foi de 91,48. A taxa de acertos RFM encontrada foi de 91,86%, o que demonstra uma consistência na classificação RFM.

O próximo *cluster* analisado foi o 17, com um volume de dados muito semelhante ao anterior, já que possui 306 registros (13,10%). Entretanto, ele foi classificado como RFM-, ou seja, clientes não recentes e nem frequentes e monetários. A média de R deste *cluster* foi de 181,07, de F foi de 1,02 e de M foi de 86,41. Quanto à

taxa de acertos RFM, o valor obtido foi de 68,30%, não muito significativo, pois foi impactado pelo acerto do atributo R que obteve somente 70,59% de taxa de sucesso.

Já o *cluster* 0 possui um total de 260 registros (11,13%) e foi classificado como RF+ e M-, o que quer dizer: clientes recentes, frequentes, mas não monetários. A média de R ficou em 57,10, de F em 1,30 e de M em 196,06. A taxa de acertos RFM foi de apenas 18,85%, impactada diretamente pela taxa de acerto de F, que foi de 20%.

O *cluster* 10 obteve um total de 242 registros (10,36%) e foi classificado como RFM-, mesma interpretação do *cluster* 17. Sua média de R ficou em 294,47, de F em 1,01 e de M em 93,00. A taxa de acertos RFM foi bastante satisfatória, com um valor de 98,35%.

Em seguida o *cluster* 18, com 205 registros (8,78%) e classificação F+ e RM-, ou seja, clientes frequentes, mas não recentes e nem monetários. A média de R ficou em 176,22, de F em 1,19 e de M em 197,98. Já a taxa de acertos RFM foi de meros 6,83%, impactada pelas taxas de acertos de R (68,29%) e F (14,63%), indicando uma classificação inexata.

Os próximos dois *clusters* analisados – o 14 e o 19 – ficaram muito parecidos, com 188 (8,05%) e 178 (7,62%) registros respectivamente. Ambos foram classificados como RFM+, ou seja, clientes importantes, pois são: recentes, frequentes e monetários. As médias encontradas no *cluster* 14 ficaram em: 42,70 em R; 1,20 em F e 313,25 em M. Já as médias do *cluster* 19 ficaram em: 125,05 em R; 1,21 em F e 313,00 em M. Em ambos os casos, a taxa de acertos RFM foi um pouco maior do que 13%. Resultados insignificantes produzidos pela taxa de acerto do atributo F (13,30% no *cluster* 14 e 14,61% no *cluster* 19).

Já o *cluster* 11, com 169 registros (7,23%) e classificação RFM-, representa justamente o inverso dos dois *clusters* analisados anteriormente, pois estes clientes não são recentes, nem frequentes e muito menos monetários. A média de R foi de 297,58, de F foi de 1,03 e de M foi de 192,02, com taxa de acertos RFM de 87,57%. O que demonstra um *cluster* consistente na classificação RFM.

Da mesma forma, o *cluster* 13 – com 134 registros (5,74%) – e o *cluster* 7 – com 87 registros (3,72%) – também obtiveram um bom resultado na taxa de acertos RFM, com valores de 91,79% e 98,85% respectivamente, e classificação RFM idênticas de M+

e RF-, isto é, clientes monetários, mas não recentes e nem frequentes. As médias encontradas no *cluster* 13 ficaram em: 215,79 em R; 1,14 em F e 332,41 em M. Já as médias do *cluster* 7 ficaram em: 315,35 em R; 1,01 em F e 324,30 em M.

Por outro lado, os *clusters* 5 e 4 ficaram com classificação RFM semelhante as dos *clusters* 14 e 9, pois foram classificados como RFM+; e da mesma forma, obtiveram uma taxa de acertos RFM muito baixa, com 19,72% e 11,48% respectivamente, impactados pela taxa de acerto de F (19,72% e 11,48%, nesta ordem). Estes *clusters* possuem 71 (3,04%) e 61 (2,61%) registros cada. De modo que as médias do *cluster* 5 foram de: 52,07 em R, 1,54 em F e 416,09 em M; e as do *cluster* 4 foram de: 90,6 em R, 1,23 em F e 539,15 em M.

O *cluster* 3, com 49 registros (2,10%) foi classificado como M+ e RF-, comportamento semelhante aos dos *clusters* 13 e 7, mas com média de R em 270,30, de F em 1,08 e de M em 538,80. A taxa de acertos RFM foi bem positiva, com um total de 93,88%.

Quanto ao *cluster* 8, com 34 registros (1,46%) e classificação FM+ e R-, representa um conjunto de clientes frequentes e monetários, mas não recentes. Já a média de R ficou em 178,76, de F em 1,20 e de M em 723,50. A taxa de acertos RFM foi de apenas 17,65%, impactada mais uma vez pela taxa de acerto de F, que foi de 20,59%.

Já o *cluster* 12, com 24 registros (1,03%), foi o que teve a menor taxa de acertos RFM, com um insignificante resultado de 4,17%, impactado pelas taxas de acerto de R (70,83%) e F (4,17%), e classificação RFM+, com média de R em 105,58, de F em 1,37 e de M em 966,49.

Os dois *clusters* seguintes possuem classificações semelhantes com M+ e RF-. Estes são os *clusters* 16 e 2, com 8 (0,34%) e 7 (0,30%) registros cada. As médias encontradas no *cluster* 16 foram de: 333,75 em R; 1,12 em F e 970,75 em M. Já as médias do *cluster* 2 foram de: 163,14 em R; 1,00 em F e 1.433,59 em M. A grande diferença é que enquanto o *cluster* 16 teve uma taxa de acertos RFM de 87,50%, o *cluster* 2 teve uma taxa de acertos RFM de apenas 42,86%, impactada negativamente pelo atributo R (42,86%).

Em seguida o *cluster* 9, com 3 registros (0,13%) e classificação RM+ e F-, ou seja, clientes recentes, monetários, mas não frequentes, com média de R em 52,66, de F em 1,00 e de M em 1.914,15, com taxa de acertos RFM de 100%.

O próximo foi o *cluster* 15, com apenas 2 registros (0,09%) e classificação M+ e RF-. Sua média de R foi de 200,00, de F foi de 1,00 e de M foi de 1.984,80; com taxa de acertos RFM de 100%.

E por último, novamente foi encontrado um *cluster* com apenas 1 registro. O mesmo da análise da métrica anterior, com apenas 0,04% dos dados. O que demonstra que mesmo com um número maior de *clusters*, ele continua isolado. E como foi observado no gráfico da Figura 5.5, cada vez mais distante dos demais, com um valor médio de M de 2.480,00, um valor médio de F igual a 2, e um valor médio de R de 235,00. Classificado como FM+ e R-, e com resultado da taxa de acertos RFM igual a 100%.

Finalizadas as avaliações de cada *cluster*, o estudo segue adiante para entender o modelo gerado e distinguir o comportamento de separação entre os *clusters*.

Em seguida, o estudo dá início à análise dos *clusters* 0, 5, 6, 9 e 14, que juntos representam 35,49% do conjunto de dados analisado. Ambos possuem um valor semelhante no atributo R, com médias que variam entre 42,70 a 57,70. Em compensação, as médias do atributo M ficaram muito diferentes, o que de fato separa os *clusters* entre eles. Porém, mais uma vez foi encontrada uma escala no atributo M quando observados os valores mínimos e máximos contidos em cada *cluster*. No caso, o *cluster* 6 contém os registros com valores entre 9,10 e 143,26; o *cluster* 0 possui os registros com valores entre 144,09 e 252,29; o *cluster* 14 tem os registros com valores entre 252,40 e 365,89; o *cluster* 5 entre 364,99 e 478,99; e o *cluster* 9 com valores entre 1.896,75 e 1.939,87.

Em seguida, foi analisado o comportamento entre os *clusters* 3, 7, 10, 11 e 16, que representam 23,76% dos registros totais. Todos apresentaram valores parecidos para a média de R, que variaram entre 270,30 e 333,75. Novamente, o que diferenciou os *clusters* foi o valor médio de M. Inclusive, com alguns intervalos de valores semelhantes aos encontrados na análise dos *clusters* 0, 5, 6, 9 e 14. Neste caso, o *cluster* 10 contém os registros com valores entre 36,40 e 142,00; o *cluster* 11 possui os registros com valores entre 144,09 e 254,89; o *cluster* 7 tem os registros com valores entre 262,99 e 419,00; o *cluster* 3 entre 444,89 e 661,28; e o *cluster* 16 com valores entre 828,80 e 1.072,80.

Comportamento parecido – e com médias de R e F muito mais próximas – foi encontrado na comparação entre os *clusters* 8 e 18, onde as médias de R foram de 178,76 e de 176,22; e as médias de F de 1,19 e 1,20 respectivamente. Juntos eles representam um pouco mais de 10% dos registros. Estes *clusters* foram separados somente pelo valor do atributo M, que no *cluster* 18 foi de 197,98 e no *cluster* 8 foi de 723,50.

Entretanto, foi possível também encontrar semelhanças nas médias de M e avaliar as separações entre os *clusters* pelo atributo R. Como no caso dos *clusters* 0, 11 e 18, que possuem média de M entre 192,02 e 197,98, e que juntos representam 27,14% dos clientes. Nesta análise, observa-se que a média de R ficou muito distante entre estes três *clusters*, já que o *cluster* 0 ficou com média de 57,10 (portanto é recente), o *cluster* 18 com média de 176,22 (não recente), assim como o *cluster* 11, que ficou com média de 297,58. Curioso também foi que a média do atributo F teve uma ligeira variação, pois o *cluster* 11 ficou com média de 1,03 (não frequente), o *cluster* 18 com média de 1,19 (não frequente), enquanto que o *cluster* 0 com média de 1,30 (frequente). A questão é que como o atributo F ficou muito assimétrico, as taxas de acertos deste atributo ficaram baixas em alguns *clusters*, o que dificultou a interpretação a partir dos valores encontrados pelas médias do atributo de F em cada grupo.

Outros dois segmentos com comportamentos semelhantes ao que foi reportado anteriormente foram encontrados nos *clusters* 14 e 19, com média de M igual a 313,00 e 313,25 respectivamente. Neste caso, o comportamento de compra foi praticamente o mesmo, inclusive no que diz respeito ao atributo F, com médias de 1,20 e 1,21. A única diferença encontrada foi no valor de R que foi de 42,70 (recente) para o *cluster* 14 e 125,05 (não recente) para o *cluster* 19. Como observação, vale ressaltar que os estrategistas de marketing da empresa deveriam aprender, com o conjunto de clientes do *cluster* 19, a identificar o motivo de uma quantidade de clientes ter este comportamento de compra e não voltar a fazer mais negócios com a empresa. Assim, eles poderiam evitar, por exemplo, que os clientes do *cluster* 14 pudessem se tornar clientes não recentes; pois como se sabe, clientes não recentes são clientes perdidos ou próximos disso. Neste caso específico, fazer com que os clientes do *cluster* 14 não se tornem futuramente clientes do *cluster* 19 pode fazer com que a taxa de evasão de clientes (*Churn*) seja reduzida em 8%.

Deste modo, foi possível encontrar também esta semelhança de comportamento no atributo M entre os *clusters* 3 e 4; entre os *clusters* 7 e 13; e entre os *clusters* 6, 10 e

17. Ambos os casos possuem uma forte diferenciação entre os *clusters* pelo valor do atributo R. Sendo assim, se compreendidos os motivos pelos quais os clientes não voltaram a fazer negócio com a empresa, poderia se evitar a tempo que um conjunto de clientes ativos se tornassem não ativos.

Além disso, quando comparados os *clusters* 1, 2, 9, 15 e 16 (cada um com menos de 1% do conjunto de dados, e que juntos não chegam a este valor) nota-se que eles foram compostos por clientes com um valor médio de M muito maior do que os demais; e que juntos com os clientes do *cluster* 12, representam os clientes mais monetários do conjunto de dados. Neste caso, o valor do atributo M, em conjunto com o valor do atributo R, os diferenciaram entre si. Mas para os estrategistas, bastaria apenas agrupá-los em um único *cluster*, sem a necessidade de distinção pelo valor de M para um acompanhamento, já que fazem parte de um seleto grupo de clientes que precisam de atenção máxima, dado o valor médio gasto por eles.

Por fim, em uma rápida avaliação da qualidade do modelo produzido pela métrica *Average Deviation*, pode-se observar que a taxa média de acertos dos *clusters* foi de apenas 58,32% e a taxa de acertos RFM do modelo foi de 55,14%. Resultados ligeiramente melhores do que os encontrados pela métrica *Davies-Bouldin*, mas que ainda assim estão muito abaixo do que o esperado.

Sendo assim, o estudo conclui a última avaliação realizada pelo recurso *Sweep Clustering* e passa agora para uma apreciação geral dos modelos encontrados na análise RFM inicial.

### **5.1.5 Síntese dos Resultados RFM Encontrados**

De um modo geral, esta análise permitiu avaliar de forma positiva a utilização do ambiente *Azure ML* para a geração de modelos de classificação não supervisionada.

Nela foi possível compreender as funcionalidades dos recursos disponíveis, tanto para a geração quanto para a validação dos modelos. Na maioria das vezes, os recursos apresentaram uma boa performance e praticidade, como foi o caso dos gráficos e análises gerados no processo de entendimento das estatísticas básicas. Em outros, como no gráfico de PCA – que representa de forma visual a formação dos *clusters* – o resultado deixou a desejar, visto que a saída é muito confusa e incompleta.

Além disso, outro recurso que chamou bastante a atenção foi o *Sweep Clustering*, já que sua aplicação foi bastante explorada durante todo o processo. Isto porque ele auxiliou na criação de diversos modelos sem que houvesse a necessidade de várias execuções, o que conseqüentemente trouxe uma economia de tempo e esforço.

É que com este recurso a busca pelo melhor valor de  $K$  consegue encontrar respaldo através dos resultados produzidos por cada uma das métricas disponibilizadas pelo ambiente. Mas como dito anteriormente, uma vez que cada métrica utiliza o seu próprio critério de desempenho, fica muito difícil compará-las, já que cada uma emprega um índice específico de performance.

Por isso, a simples utilização deste recurso não foi suficiente para encontrar o melhor valor de  $K$  para o conjunto de dados analisado. E como alternativa, o estudo avaliou o comportamento interno de cada *cluster* para entender os seus respectivos conteúdos através dos valores encontrados em cada atributo. Este processo ajudou a compreender a formação interna de cada *cluster* e a inseri-los dentro de um contexto de negócio.

Neste caso, como foi observada durante toda a análise, a divisão dos *clusters* foi feita pelas médias contidas nos atributos M e R, com o atributo de M se sobrepondo ao de R. Assim, foi possível ter uma ideia de que o atributo que representava o valor monetário possuía uma força maior do que os demais no processo de separação e formação dos *clusters*.

Também foi observado que à medida que a quantidade do número de *clusters* aumentava, a separação e a distribuição dos registros entre as classes também melhoravam. Esta declaração pode ser justificada pelas taxas médias de acertos globais dos *clusters* e pelas taxas de acertos dos modelos, visto que o melhor modelo produzido pela métrica *Simplified Silhouette*, com 3 *clusters*, obteve um resultado médio de acertos dos *clusters* de apenas 32,43% e resultado de acertos RFM do modelo de 31,59%; enquanto que os produzidos pelas métricas *Davies-Bouldin* e *Dunn*, com 10 *clusters*, conseguiram uma taxa média de acertos dos *clusters* de 50,27% e taxa de acertos RFM do modelo de 46,53%; e por último, com 20 *clusters*, a partir da métrica *Average Deviation*, foi encontrada uma taxa média de acertos dos *clusters* de 58,32% e taxa de acertos RFM do modelo de 55,14%.

De todo modo, estas taxas de acertos atribuídas à classificação RFM foram insuficientes e precisam ser melhoradas. Boa parte deste desempenho se deve à distorção (assimetria) do atributo F. Com isso, já foi possível identificar uma causa para o baixo desempenho e inferir possíveis melhorias para aumentar a qualidade do modelo. Estas melhorias serão tratadas em uma análise futura neste próprio estudo, a fim de se evitar ou contornar a assimetria causada por esta disfunção.

Além disso, não foi possível encontrar uma segmentação que fosse capaz de separar linearmente todos os *clusters* (bem-vindos ao mundo real). Em poucos casos, foram encontrados alguns *clusters* com separação entre eles; mas no geral, foi observada uma grande sobreposição entre a maioria dos agrupamentos. Outro ponto importante é que nenhum modelo conseguiu produzir um total de *clusters* que ficasse dentro do limite de critério estabelecido inicialmente pelo estudo, ou seja, com o menor *cluster* possuindo não menos do que 1% dos registros.

Por tudo isso, é imperativo tentar aumentar o resultado da taxa de acertos RFM através da busca e criação de novos modelos. Talvez isto possa ser feito a partir de uma melhor distribuição dos dados (redução da concentração de um único valor em um campo específico – como é o caso do atributo F).

Mas antes disso, o estudo irá avaliar a criação de novos modelos a partir da inclusão do atributo P seguindo as mesmas etapas desta análise. O objetivo é descobrir se a inclusão deste atributo poderá impactar no processo de segmentação dos clientes.

## 5.1.6 Tabelas Auxiliares – RFM

Tabela 5.7: Tabela auxiliar com os resultados da métrica *Simplified Silhouette* da análise RFM.

G.	Qtd.	%	RFM+	RFM-	Media R	Min R	Max R	Media F	Min F	Max F	Media M	Min M	Max M	% Acertos R	% Acertos F	% Acertos M	Qtd. Acertos RFM	% Acertos RFM
0	2.041	87,37%		RFM	158,62	1	365	1,12	1	8	183,41	9,10	373,09	48,11	91,08%	70,26%	706	34,59%
2	282	12,07%	RFM		148,42	1	365	1,28	1	11	561,90	372,79	1.119,99	54,96	13,48%	100%	25	8,87%
1	13	0,56%	RM	F	148,84	4	313	1,07	1	2	1.709,78	1.260,05	2.480,00	53,85	92,31%	100%	7	53,85%

Tabela 5.8: Tabela auxiliar com os resultados das métricas *Davies-Bouldin* e *Dunn* da análise RFM.

G.	Qtd.	%	RFM+	RFM-	Media R	Min R	Max R	Media F	Min F	Max F	Media M	Min M	Max M	% Acertos R	% Acertos F	% Acertos M	Qtd. Acertos RFM	% Acertos RFM
6	697	29,84%	RF	M	79,37	1	170	1,15	1	6	129,28	9,10	215,34	94,26%	11,48%	100%	77	11,05%
0	667	28,55%		RFM	255,65	166	365	1,04	1	5	122,20	11,90	211,00	100%	96,70%	100%	645	96,70%
5	460	19,69%	RFM		83,10	1	176	1,23	1	8	301,02	217,07	384,99	90,87%	14,78%	88,26%	55	11,96%
7	260	11,13%	M	RF	260,59	171	365	1,10	1	8	312,42	217,90	450,60	100%	93,08%	93,85%	227	87,31%
4	122	5,22%	RFM		77,81	1	209	1,39	1	11	473,96	389,49	659,98	87,70%	16,39%	100%	19	15,57%
3	76	3,25%	M	RF	229,96	66	364	1,13	1	3	598,77	461,19	784,60	82,89%	88,16%	100%	55	72,37%
8	41	1,76%	FM	R	165,80	4	354	1,26	1	10	930,67	746,69	1.119,99	53,66%	7,32%	100%	2	4,88%
2	7	0,30%	M	RF	163,14	50	313	1,00	1	1	1.433,59	1.260,05	1.545,40	42,86%	100%	100%	3	42,86%
9	5	0,21%	RM	F	111,60	4	206	1,00	1	1	1.942,41	1.896,75	2.049,24	60,00%	100%	100%	3	60,00%
1	1	0,04%	FM	R	235,00	235	235	2,00	2	2	2.480,00	2.480,00	2.480,00	100%	100%	100%	1	100%

Tabela 5.9: Tabela auxiliar com os resultados da métrica *Average Deviation* da análise RFM.

G.	Qtd.	%	RFM+	RFM-	Media R	Min R	Max R	Media F	Min F	Max F	Media M	Min M	Max M	% Acertos R	% Acertos F	% Acertos M	Qtd. Acertos RFM	% Acertos RFM
6	307	13,14%	R	FM	56,25	1	118	1,10	1	6	91,48	9,100	143,26	100%	91,86%	100%	282	91,86%
17	306	13,10%		RFM	181,07	119	237	1,02	1	2	86,41	11,900	138,00	70,59%	97,71%	100%	209	68,30%
0	260	11,13%	RF	M	57,10	1	117	1,30	1	6	196,06	144,090	252,29	100%	20,00%	93,08%	49	18,85%
10	242	10,36%		RFM	294,47	237	365	1,01	1	2	93,00	36,400	142,00	100%	98,35%	100%	238	98,35%
18	205	8,78%	F	RM	176,22	117	238	1,19	1	5	197,98	141,330	261,96	68,29%	14,63%	90,73%	14	6,83%
14	188	8,05%	RFM		42,70	1	83	1,20	1	8	313,25	252,400	365,89	100,00%	13,30%	100%	25	13,30%
19	178	7,62%	RFM		125,05	85	171	1,21	1	7	313,00	247,700	408,19	85,39%	14,61%	100%	24	13,48%
11	169	7,23%		RFM	297,58	238	365	1,03	1	2	192,02	144,090	254,89	100%	96,45%	89,35%	148	87,57%
13	134	5,74%	M	RF	215,79	165	265	1,14	1	8	332,41	268,750	435,99	100%	91,79%	100%	123	91,79%
7	87	3,72%	M	RF	315,35	263	365	1,01	1	2	324,30	262,990	419,00	100%	98,85%	100%	86	98,85%
5	71	3,04%	RFM		52,07	1	150	1,54	1	11	416,09	364,990	478,99	100%	19,72%	100%	14	19,72%
4	61	2,61%	RFM		90,63	3	179	1,23	1	8	539,15	455,000	659,98	85,25%	11,48%	100%	7	11,48%
3	49	2,10%	M	RF	270,30	191	364	1,08	1	3	538,80	444,890	661,28	100%	93,88%	100%	46	93,88%
8	34	1,46%	FM	R	178,76	40	331	1,20	1	2	723,50	654,990	849,99	64,71%	20,59%	100%	6	17,65%
12	24	1,03%	RFM		105,58	4	215	1,37	1	10	966,49	844,990	1.119,99	70,83%	4,17%	100%	1	4,17%
16	8	0,34%	M	RF	333,75	295	354	1,12	1	2	970,75	828,800	1.072,80	100%	87,50%	100%	7	87,50%
2	7	0,30%	M	RF	163,14	50	313	1,00	1	1	1.433,59	1.260,050	15.45,40	42,86%	100%	100%	3	42,86%
9	3	0,13%	RM	F	52,66	4	128	1,00	1	1	1.914,15	1.896,750	1.939,87	100%	100%	100%	3	100%
15	2	0,09%	M	RF	200,00	194	206	1,00	1	1	1.984,80	1.920,370	2.049,24	100%	100%	100%	2	100%
1	1	0,04%	FM	R	235,00	235	235	2,00	2	2	2.480,00	2.480,000	2.480,00	100%	100%	100%	1	100%

## 5.2 Análise RFMP

Nesta nova etapa do estudo o que se pretende é avaliar se a inclusão do parâmetro  $P$  trará alguma influência no processo de segmentação de clientes e na formação dos seus grupos. O intuito é confrontar se haverá algum impacto entre os modelos criados a partir da classificação RFM com os novos que serão gerados a partir do modelo RFMP.

Este processo será realizado seguindo a mesma regra de performance da análise anterior. E a busca pelo melhor valor de  $K$  será realizada também através do uso do recurso *Sweep Clustering*. A seguir os resultados obtidos.

### 5.2.1 Simplified Silhouette

Como é observado na tabela a seguir, o melhor valor de  $K$  a partir do uso da métrica *Simplified Silhouette* foi com 3 *clusters*, resultado idêntico ao encontrado na análise RFM anterior.

Tabela 5.10: Pontuação da métrica *Simplified Silhouette* da análise RFMP.

Pontuação	Número de <i>Clusters</i>		Pontuação	Número de <i>Clusters</i>
0,642806	3		0,544955	5
0,639138	2		0,543336	10
0,604019	7		0,542931	18
0,572911	6		0,542195	19
0,56654	4		0,541977	14
0,561887	16		0,537876	12
0,560579	9		0,534887	13
0,552075	17		0,532246	11
0,551969	8		0,527719	20
0,545708	15			

Apesar disso, nota-se pela Tabela 5.10 que a pontuação máxima foi um pouco diferente que a encontrada na análise RFM com esta métrica, e que a ordem subsequente com os demais valores de  $K$  também apresentou pequenas mudanças. Mas nada disso trouxe algum resultado efetivo que diferenciasse o modelo gerado, uma vez que o melhor número de *clusters* sugerido foi o mesmo da análise anterior. Inclusive, o conteúdo dos agrupamentos foi exatamente igual, ou seja, a quantidade de registros por *cluster* e seus

respectivos valores – tanto das médias quando dos valores mínimos e máximos – não sofreram qualquer alteração. Por conta disso, será feito aqui apenas uma rápida observação sobre o conteúdo do atributo P na composição de cada *cluster*. A partir deste ponto o estudo passa a fazer uso da Tabela 5.16 (tabela auxiliar).

E como pode ser observado no *cluster* 0, a média de P foi de 44,47%, com valor mínimo de 7,28 e máximo de 100. Este *cluster* foi classificado como RFMP-, ou seja, um conjunto de clientes não recentes, não frequentes, não monetários e nem lucrativos. A taxa de acerto no campo % de acertos de P foi de 61,59%, o que também impactou o resultado da taxa de % de acertos RFMP do *cluster*, que passou a ser de 21,07% (valor inferior ao observado na primeira análise, que foi de 34,59%).

Já o *cluster* 2 ficou com média de P de 44,68 – com valor mínimo de 34,15 e máximo de 52,62 – classificação RFMP+ e % de acertos de P de 53,19%. Isto quer dizer que este *cluster* representa um comportamento contrário ao anterior, ou seja, clientes recentes, frequentes, monetários e lucrativos. Quanto ao % de acertos RFMP, o valor final foi de apenas 4,26%.

Por último, o *cluster* 3 com média de P de 41,41, valor mínimo de 34,68 e máximo de 47,46. Este grupo foi classificado como RM+ e FP-, isto é, clientes recentes e monetários, porém não frequentes e não lucrativos. A taxa de % de acertos de P foi de 69,23%, com taxa de acertos RFMP de 38,46%.

Com estes resultados, a taxa média de acertos dos *clusters* produzida por esta métrica foi de apenas 21,26%, um valor muito abaixo dos 32,43% encontrados na análise anterior. Além disso, a taxa de acertos RFMP do modelo foi de 19,14%, muito inferior ao resultado de 31,59% obtido pela análise RFM através desta métrica, o que tornou inconsistentes as classificações RFMP atribuídas.

Os motivos apresentados para este baixo desempenho foram os mesmos da análise RFM com esta métrica: uma grande concentração dos registros em um único *cluster*, o que prejudicou a distribuição dos dados. Outro ponto que chamou a atenção foi que as médias dos agrupamentos para o atributo P ficaram próximas à média global.

Deste modo, foi observado que apesar de não ter tido qualquer impacto nas formações dos *clusters* e nem nos seus respectivos conteúdos, a inclusão do atributo P trouxe uma piora nos valores apresentados pela taxa de acertos RFMP. Por tudo isso, o

estudo descarta este modelo e segue adiante com as demais métricas para avaliar os impactos do atributo P.

### 5.2.2 *Davies-Bouldin*

O próximo passo foi analisar a formação dos *clusters* criados por esta métrica. E como pode ser observado pelos resultados da Tabela 5.11, o melhor modelo produziu um total de 9 agrupamentos. Este resultado já foi um pouco diferente do que foi encontrado na análise RFM anterior – uma vez que produziu um *cluster* a menos no melhor modelo sugerido – e alterou não só a pontuação da métrica, como também a ordem de performance do valor de *K* indicado.

Esta mudança pode ser observada ao confrontar os dados desta análise com os encontrados na Tabela 5.3, onde é possível observar que a pontuação de um modelo com 9 *clusters* ficou apenas na décima posição. Já o modelo com 10 *clusters* (o melhor daquela análise), foi apenas o terceiro na análise com o parâmetro P (análise atual). Um ponto interessante a se notar é que em ambos os casos o modelo sugerido com 8 *clusters* ficou em segundo lugar.

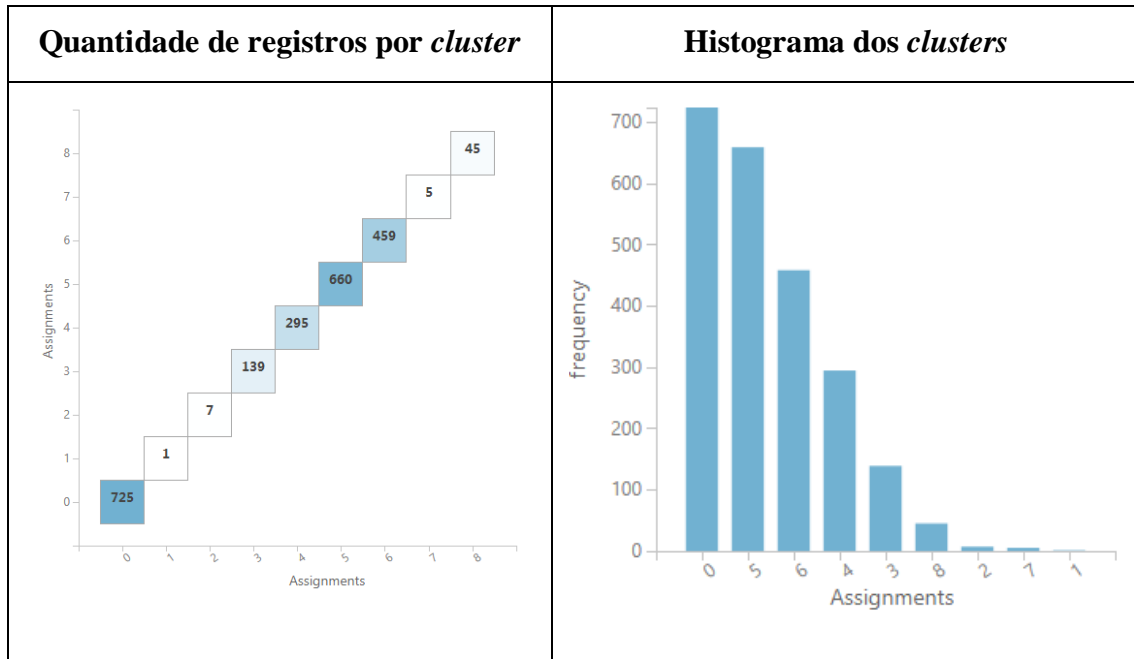
Tabela 5.11: Pontuação da métrica *Davies-Bouldin* da análise RFMP.

<b>Pontuação</b>	<b>Número de <i>Clusters</i></b>		<b>Pontuação</b>	<b>Número de <i>Clusters</i></b>
0,614738	9		0,736157	2
0,649746	8		0,740643	13
0,679089	10		0,743741	15
0,683093	16		0,750446	14
0,689233	3		0,750641	11
0,702324	17		0,763732	20
0,713097	18		0,773307	6
0,718735	19		0,839337	4
0,732913	7		0,873869	5
0,736023	12			

Porém, analisando os resultados da Tabela 5.17, pode-se observar que o maior agrupamento possui um total de 725 registros (31,04% do total), enquanto que o menor possui somente 1 registro. De todo modo, com um *cluster* a menos já foi possível perceber

que o maior *cluster* ficou com uma concentração de dados maior, uma vez que na análise anterior desta métrica o maior *cluster* ficou com um total de 697 registros.

Tabela 5.12: Gráficos dos resultados da métrica *Davies-Bouldin* da RFMP.



Já pela análise do gráfico da Figura 5.6, observa-se que os *clusters* continuaram sobrepostos, com um pequeno destaque para os *clusters* 2, 7 e 1. Por sinal, os menores de todos.

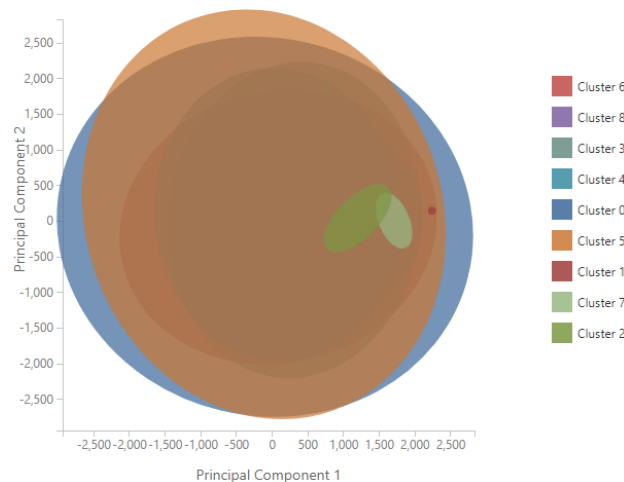


Figura 5.6: Gráfico da visualização dos *clusters* da 1ª análise RFMP com a métrica *Davies-Bouldin*.

Dando continuidade, o estudo fez novamente uso da tabela auxiliar para melhor compreender a composição dos *clusters*. Os dados desta tabela podem ser encontrados na Tabela 5.17, e são descritos com mais detalhes logo a seguir.

Como pode ser visto, o *cluster 0* – que possui 725 registros (31,04% do total) – ficou classificado como RFP+ e M-, ou seja, clientes recentes, frequentes e lucrativos, mas com um tíquete médio baixo. As médias de R, F, M e P ficaram, respectivamente, em: 80,90; 1,16; 132,42 e 44,71. Já a taxa de acertos RFMP deste *cluster* foi de apenas 5,24%, um resultado muito abaixo do que foi encontrado no maior *cluster* da análise RFM anterior feita por esta métrica. Este resultado foi impactado pelas taxas de acertos dos atributos F (11,45%) e P (47,72%), o que fez com que este *cluster* ficasse bastante incoerente com a classificação RFMP.

O próximo grupo analisado foi o *cluster 5*, com 660 registros (28,25%) e classificação P+ e RFM-, isto é, clientes lucrativos, mas não recentes, nem frequentes e muito menos monetários. Este grupo parece ser o *cluster 0* da análise RFM anterior, com respectivas médias de R, F, M e P de: 256,55; 1,04; 121,62; e 44,51. A diferença é que alguns registros foram atribuídos a outros grupos. Entretanto, ao incluir o atributo P, o valor da taxa de acertos RFM ficou em 38,18%, com grande influência do acerto do campo P, que foi de apenas 40%. Isto começa a sinalizar que a inclusão do atributo P está trazendo uma complexidade maior para a taxa de acertos RFMP.

Na sequência, o *cluster 6*, com 459 registros (19,65%) e classificação RFM+ e P-, com médias R, F, M e P de: 72,98; 1,28; 319,19 e 44,37. Este grupo é formado por clientes recentes, frequentes, monetários e não lucrativos. Sua taxa de acertos RFMP foi de somente 9,80%, sofrendo influências das taxas de acertos do atributo F (16,34%) e em parte do atributo P (76,47%). Apesar disso, a taxa do atributo P foi a segunda maior na análise deste modelo, perdendo apenas para a do *cluster 8* que será discutido mais à frente.

Agora o *cluster 4*, com 295 registros (12,63%) e classificação M+ e RFP- (clientes monetários, mas não recentes, não frequentes e nem lucrativos). As médias obtidas por este grupo foram de: 249,76 (R); 1,10 (F); 310,37 (M) e 43,97 (P). Dos *clusters* com mais de 1 registro, este foi o que obteve a maior taxa de acertos RFMP, com um valor de 61,02%, e que ainda assim representa um resultado pouco expressivo. Neste caso, o acerto de P foi o que mais pesou para este valor, com taxa de 66,10%.

Na próxima análise o *cluster 3*, com 139 registros (5,95%) e classificação FMP+ e R-, isto é, clientes frequentes, monetários, lucrativos, mas não recentes. As médias encontradas foram de: 163,56 (R); 1,15 (F); 555,96 (M); e 44,85 (P). A taxa de acertos

RFMP deste grupo foi a pior de todas, com apenas 2,16%, influenciada pelas seguintes taxas: 50,36% (R), 10,07% (F), 100% (M) e 58,27% (P).

O próximo *cluster* foi o de número 8, com 45 registros (1,93%) e classificação FMP+ e R-. Comportamento idêntico ao do *cluster* 3, mas com médias de R, F, M e P de: 171,15 (R); 1,28 (F); 915,26 (M); e 44,50 (P). Já a taxa de acertos RFMP foi de 8,89%, influenciada pelas seguintes taxas: 55,56% (R), 11,11% (F), 100,00% (M) e 77,78% (P).

Já os *clusters* 2, 7 e 1 continuaram os mesmos, ou seja, idênticos aos apresentados pela métrica *Davies-Bouldin* na análise RFM anterior. A única diferença foi que ambos ganharam um atributo P- a mais, isto é, representam segmentos de clientes não rentáveis, apesar de terem os maiores valores monetários. O valor da média de P no *cluster* 2 foi de 41,06; já no *cluster* 5 foi de 43,26 e no *cluster* 1 foi de apenas 34,68 (o menor valor encontrado por um *cluster* neste modelo). Já as taxas de sucesso apresentadas pelos *clusters* 2 e 7 tiveram um recuo na performance, caindo para 14,29% no *cluster* 2 e 20% no *cluster* 7.

No *cluster* 1, a taxa se manteve em 100%, até porque este grupo conta com a existência de um único registro. Porém, o que mais chamou a atenção foi que, apesar de ter o maior valor para a média do atributo M (o que é positivo), este *cluster* possui o menor valor para a média do atributo P, com apenas 34,68% (o que não é bom).

De todo modo, dando continuidade a esta análise, o estudo passa agora a comparar os *clusters* do modelo para tentar entender a separação entre eles.

E como pode ser visto, os grupos 6 e 0 possuem médias do atributo R de 72,98 e 80,90 respectivamente, porém possuem médias de M muito distintas, que são de: 319,19 e 132,42. Mais uma vez, foi possível encontrar uma escala de valores no atributo M entre estes dois *clusters*, já que o *cluster* 0 possui valores de M entre 9,10 e 226,90; enquanto que o *cluster* 6 possui valores entre 222,75 e 470,09. Vale ressaltar que na análise RFM anterior com esta mesma métrica, um *cluster* a mais foi encontrado com média de R próxima a destes *clusters*, mas com média de M de 473,93 e valores entre 389,49 e 659,98.

Além disso, os valores das médias de P ficaram muito próximos, já que a média deste atributo no *cluster* 6 foi de 44,37 e no *cluster* 0 foi de 44,71. Entretanto, como a média de P de todo o conjunto de dados ficou em 44,49, isto fez com que o *cluster* 6 fosse classificado como P- (clientes não rentáveis) e o *cluster* 0 com P+ (clientes rentáveis).

Este é um ponto que chamou muita atenção, pois será mesmo que uma diferença percentual de 0,34 é suficiente para classificar um conjunto de dados entre positivo e negativo? Esta é uma questão que merece reflexão e que deve ser analisada com mais detalhes pelo estudo mais à frente.

Seguindo no mesmo raciocínio, os *clusters* 4 e 5 possuem médias de R iguais a 249,76 e 256,55, com uma escala de valores que variam entre 11,90 e 211 no *cluster* 5; e entre 217,90 e 465,20 no *cluster* 4. Comportamento semelhante ao encontrado nos *clusters* 6 e 0, mas com a diferença dos valores no atributo R.

Outro destaque na separação dos *clusters* pelo atributo M pode ser visto entre os *clusters* 3, 8, 2 e 7, onde foi possível observar que a separação dos *clusters* seguiu uma escala do valor contido em cada agrupamento por este atributo. No *cluster* 3 estão contidos os registros com valores de M entre 426,49 e 734,89; no *cluster* 8 os que possuem valores entre 743,99 e 1.119,99; no *cluster* 2 os que estão entre 1.260,05 e 1.545,40; e no *cluster* 7 os que estão entre 1.896,75 e 2.049,24. Dessa forma, percebe-se que os *clusters* foram separados pelos valores contidos no atributo M, já que o valor de R entre os *clusters* ficou muito semelhante.

Ainda assim, se analisados apenas os *clusters* 4 e 6, percebe-se que ambos possuem as médias de M bem semelhantes, com valores de 310,73 e 319,19 respectivamente. Porém, os valores de R ficaram bem distintos, já que a média de R do grupo 4 foi de 249,76 e do grupo 6 foi de 72,98. Deste modo, com esta análise foi possível perceber que os *clusters* também foram separados pelos valores do atributo R e não apenas pelo atributo M.

Por último o *cluster* 1 – que ficou exatamente igual ao *cluster* 7 da análise RFM feita por esta métrica – e que possui um único registro com o maior valor da média do atributo M e o menor valor da média do atributo P.

Com o término desta análise, nota-se que a inclusão do atributo P alterou o modelo proposto anteriormente na análise RFM através da métrica *Davies-Bouldin*. A primeira observação foi na quantidade de *clusters* produzida – um a menos que a anterior – e consequentemente no conteúdo dos registros contidos neles.

Além disso, foi possível observar que nem sempre o pedido com o maior valor monetário é o mais lucrativo. Isto se deve muitas vezes pelas estratégias das empresas em

ganhar mais a partir de uma venda concentrada e de maior volume, mas com uma lucratividade menor. Porém, uma venda deste tipo pode centralizar as receitas das empresas nas mãos de poucos consumidores, o que torna o modelo de negócio mais sensível à evasão dos clientes. Por isso, analisar esta relação pode ser saudável para as empresas.

Outro ponto que chamou a atenção foi que as médias dos agrupamentos para o atributo P ficaram próximas à média global, o que trouxe uma dificuldade tanto para o acerto deste atributo quanto para a própria interpretação do modelo. Pois como visto na análise entre os *clusters* 0 e 6, a diferença entre a média de um e outro foi de apenas 0,34, de modo que um ficou classificado como P+ (lucrativo) e outro como P- (não lucrativo). Por isso, é necessário analisar melhor esta relação que torna a classificação positiva ou negativa a partir de um único valor absoluto (no caso a média).

Por último, a taxa média de acertos dos *clusters* ficou em apenas 28,84%, um valor muito abaixo do que o encontrado na análise RFM com esta métrica (que foi de 50,27%). Além disso, a taxa de acertos RFM do modelo foi de apenas 22,47%, resultado inferior aos 46,53% obtidos na análise RFM.

Deste modo, é possível afirmar que a inclusão do atributo P trouxe uma dificuldade a mais para a assertividade dos *clusters*. Ainda assim, não foi encontrado o modelo que pudesse atender a premissa do estudo.

### **5.2.3 *Dunn***

A análise seguinte foi feita com a métrica *Dunn*, que encontrou como melhor índice um modelo com um total de 8 *clusters* (Tabela 5.13). Este resultado foi diferente da análise RFM anterior realizada a partir desta métrica. Afinal, esta análise gerou dois *clusters* a menos e alterou tanto a pontuação dos índices quanto a ordem de performance do valor de *K*.

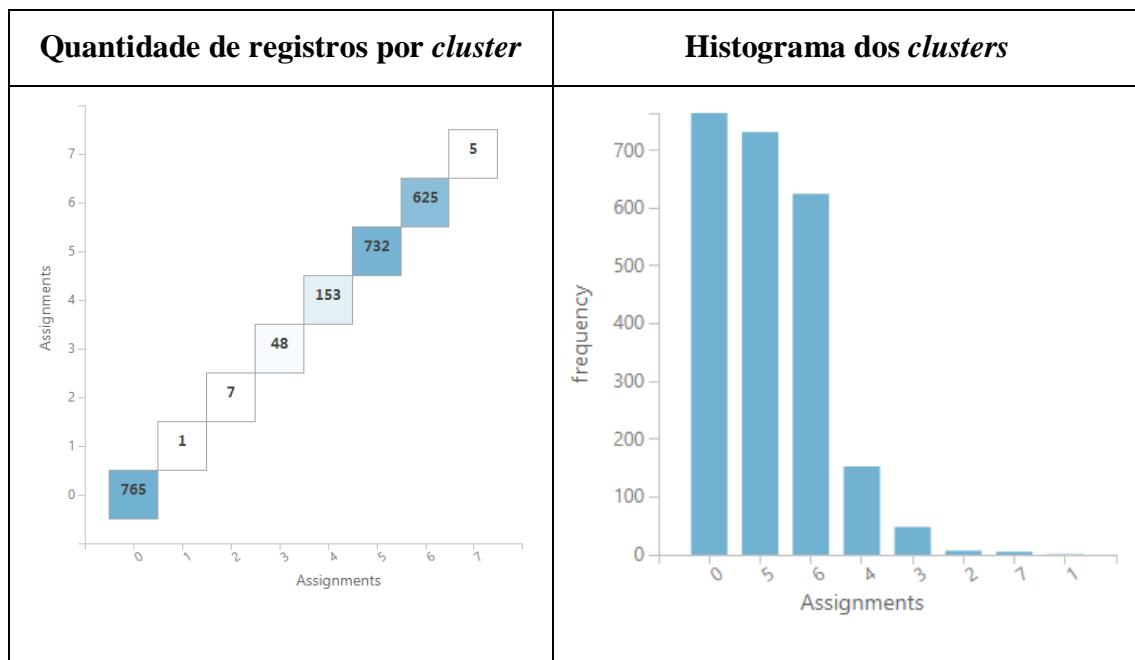
Conforme pode ser visto nos resultados da Tabela 5.14, com este modelo o maior *cluster* ficou com 765 registros (32,75% dos dados) e o menor com apenas 1 registro (com 0,04% dos dados), sem atender ao critério de performance estabelecido. Vale ressaltar que foram encontrados também outros 2 *clusters* com menos de 1% dos dados, e ambos com valores de *M* acima de 1.000,00, o que indica cada vez mais que os clientes com este

comportamento de compra são excepcionais, pois se destacam dos demais justamente pelo valor.

Tabela 5.13: Pontuação da métrica *Dunn* da análise RFMP.

Pontuação	Número de <i>Clusters</i>	Pontuação	Número de <i>Clusters</i>
0,692401	8	0,488513	3
0,689898	9	0,458993	20
0,689189	10	0,455967	14
0,603297	17	0,42287	15
0,603297	16	0,40093	7
0,600793	18	0,299725	2
0,553202	19	0,249968	6
0,53752	12	0,248158	4
0,53322	13	0,242107	5
0,515649	11		

Tabela 5.14: Gráficos dos resultados da métrica *Dunn* da análise RFMP.



Com isso, percebe-se novamente que a inclusão do atributo P está influenciando o resultado do modelo, tornando o conjunto de dados menos distribuído. Esta observação pode ser feita tanto na análise desta métrica quanto na que foi discutida no tópico anterior.

Já pela análise do gráfico da Figura 5.7, nota-se que os *clusters* continuaram sobrepostos, com um pequeno destaque para os *clusters* 1, 2 e 7, que representam justamente os *clusters* com menos de 1%.

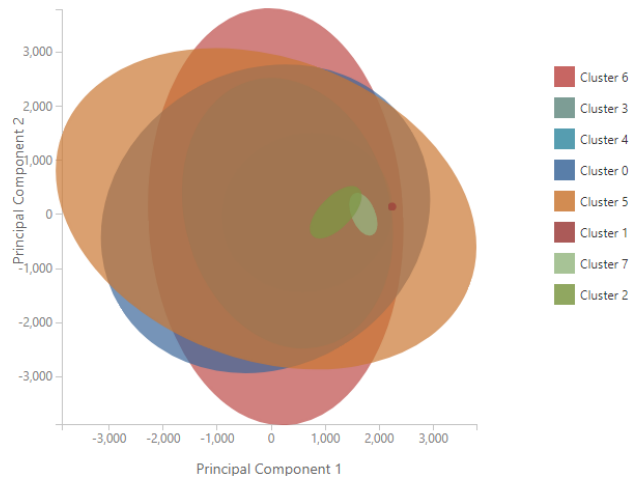


Figura 5.7: Gráfico da visualização dos *clusters* da 1ª análise RFMP com a métrica *Dunn*.

Em seguida, foi feito uso da tabela auxiliar (Tabela 5.18) para melhor compreender a composição dos *clusters*, que possui os seus resultados descritos com mais detalhes a seguir.

O *cluster* 0 foi o maior de todos, com 765 registros (32,75%) e classificação RFP+ e M-, o que significa um cliente recente, frequente, lucrativo, mas não monetário. As médias de R, F, M e P ficaram em: 82,79; 1,16; 130,42 e 44,76. A taxa de acertos RFMP foi de apenas 5,23% influenciada pelas taxas de F (11,37%) e de P (48,10%).

Já o *cluster* 5 possui 732 registros (31,34%) e classificação RFMP-, ou seja, clientes não recentes, não frequentes, não monetários e nem lucrativos. As médias de R, F, M e P ficaram em: 266,40; 1,04; 146,46; e 44,44. A taxa de acertos RFMP foi de 51,37%, com taxa de acerto de P em 57,51%.

O *cluster* seguinte foi o 6, com 625 registros (26,76%) e classificação RFM+ e P-, mesmo padrão encontrado no *cluster* 0. As médias deste *cluster* para os atributos R, F, M e P ficaram em: 116,21; 1,23; 320,11 e 44,17. A taxa de acertos RFMP foi de somente 7,20%, influenciada pela taxa de F (13,60%) e, com exceção do *cluster* 1 (com apenas um registro), a taxa de P deste *cluster* foi a maior de todas dentro deste modelo, com um resultado de 79,04%.

Na sequência o *cluster* 4, com 153 registros (6,55%) e classificação MP+ e RF-, ou seja, clientes monetários, lucrativos, mas não recentes e nem frequentes. As médias encontradas para R, F, M e P foram de: 171,67; 1,13; 538,39; e 44,82. Já a taxa de acertos RFMP foi de 31,37%, impactada pelas taxas de R (52,29%) e de P (57,52%).

Em seguida o *cluster* 3, com 48 registros (2,05%) e classificação FM+ e RP-, isto é, clientes frequentes, monetários, mas não recentes e nem lucrativos. As médias R, F, M e P encontradas foram de: 174,12; 1,31; 903,68; e 44,34. Quanto à taxa de acertos RFMP o valor foi de apenas 4,17%, impactada pelas taxas de R (58,33%); F (14,58%) e P (25%).

Quanto aos *clusters* 2, 7 e 1 eles possuem exatamente o mesmo conteúdo da análise RFMP encontrada pela métrica *Davies-Bouldin*, por isso não será feita qualquer nova observação sobre eles.

Assim, o estudo parte agora para a avaliação sobre os atributos e valores de separação entre os *clusters*.

Já de início destaca-se a relação entre os *clusters* 7 e 6, com médias de R de 111,60 e de 116,21, mas com valores de M de 1.942,41 e 320,11 respectivamente. Apesar dessa grande diferença, ambos foram classificados com M+. Além disso, no *cluster* 7 os clientes ficaram com média de F de apenas 1 (clientes não frequentes), enquanto que no *cluster* 6 há clientes frequentes e não frequentes. Porém, como a média ficou em 1,23, todos os clientes deste grupo foram marcados como frequentes. Consequentemente, a taxa de acerto do atributo F ficou muito baixa, apenas 13,60%, o que impactou no resultado e na consistência da classificação RFMP sugerida pela análise. Já em relação ao atributo P, as médias ficaram próximas, com valores de 43,62 e 44,17 respectivamente, com ambos sendo classificados com P+ (lucrativos).

Outra relação que vale observar é a dos *clusters* 0 e 5, com médias de M de 130,42 e 146,46, mas com médias de R de 82,79 e 266,40 respectivamente. Neste caso, os grupos foram separados pelo valor de R e não pelo de M. Quanto ao atributo F, os valores mínimo e máximo de ambos os *clusters* ficaram bem semelhantes, o que demonstra que o modelo não utilizou o conteúdo deste atributo para separar os *clusters*; apesar de o primeiro ter ficado com uma média de 1,16 (frequente) e o segundo com 1,04 (não frequente). Isto se deve à assimetria do atributo F. Quanto ao atributo P, o primeiro

ficou com média de 44,76 (P+, clientes lucrativos) e o segundo com média de 44,44 (P-, não lucrativo). Esta diferença tão curta entre os valores de P foi suficiente para distinguir os *clusters* na classificação RFMP, e isto se deve pela comparação entre os valores absolutos da média de P em relação à média global de P.

E por último a comparação dos *clusters* 2, 4 e 3, que ficaram com as médias de R com os respectivos valores: 163,14; 171,67 e 174,12. Isto porque o modelo separou estes três grupos como se fosse uma relação complementar do valor de M entre eles, isto é, uma escala, variando de 387,50 a 719,88 no *cluster* 4, de 725,42 a 1.119,99 no *cluster* 3 e de 1.260,05 a 1.545,40 no *cluster* 2.

Deste modo, o estudo chega ao final da análise desta métrica. Nela, foi possível avaliar novamente que o atributo P trouxe um impacto na geração do modelo, sendo percebido tanto na pontuação quanto no ordenamento do valor de *K*. Também foi possível notar que este impacto não foi positivo, visto que a qualidade do modelo e as respectivas classificações RFMP dos *clusters* não foram consistentes, já que a taxa média de acertos dos *clusters* foi de apenas 29,20%, enquanto que a taxa de acertos RFMP do modelo foi de 22%.

Ainda assim, está ficando cada vez mais claro a compreensão dos modelos e o que está impactando na formação e na separação dos *clusters*. Tudo isso torna positivo o que foi feito até aqui.

Desta forma, o estudo segue para avaliação da última métrica da análise RFMP.

#### **5.2.4 *Average Deviation***

O melhor modelo indicado pela métrica *Average Deviation* encontrou um total de 19 *clusters*, conforme pode ser visto na Tabela 5.15, um *cluster* a menos do que o encontrado por esta mesma métrica na análise RFM. Esta foi a métrica que produziu o modelo com a maior quantidade de *clusters* entre as métricas avaliadas na análise RFMP.

Além disso, como pode ser observado na Figura 5.8 e na Figura 5.9 o maior *cluster* obteve um total de 312 registros (13,36% dos dados), e o menor continuou com apenas 1 registro (com 0,04% dos dados). Além do *cluster* 1, outros cinco *clusters* ficaram com menos de 1% dos registros, todos com um valor médio do atributo M acima de 956.

Tabela 5.15: Pontuação da métrica *Average Deviation* da análise RFMP.

Pontuação	Número de <i>Clusters</i>		Pontuação	Número de <i>Clusters</i>
43,559803	19		68,811258	10
43,87662	20		71,750435	9
46,389253	18		81,706712	8
47,014679	17		81,979171	7
47,091983	16		82,874815	6
48,194539	15		103,200011	5
53,128563	14		109,011208	4
54,535035	13		137,356068	3
57,943787	12		145,744753	2
62,271623	11			

Novamente foi possível perceber que o atributo P influenciou o resultado do modelo, imputando uma concentração maior de registros em cada um dos *clusters* encontrados. Já pela análise do gráfico da Figura 5.7, observa-se que os *clusters* continuaram sobrepostos, com um pequeno destaque para os *clusters* 2, 7 e 16.

Na continuidade, o estudo indica como referência a Tabela 5.19 que guiará os próximos passos.

De início foi avaliado o *cluster* 11 com 312 registros (13,36%) e com classificação RP+ e FM-, ou seja, clientes recentes, lucrativos, mas não frequentes e nem monetários. As médias de R, F, M e P ficaram em: 56,44; 1,10; 92,34 e 45,83. Já a taxa de acertos RFMP foi de 60,58%, sofrendo impacto direto da taxa de acerto de P, que foi de apenas 67,63%. Com este resultado, nota-se que houve um impacto negativo para a consistência do *cluster* trazido pelo atributo P.

Na sequência, foi analisado o *cluster* 17 com 306 registros (13,10%) e classificação RFMP-, isto é, não são recentes, não são frequentes, não são monetários e nem lucrativos (o pior cenário para um grupo de clientes). As médias de R, F, M e P ficaram em: 181,07; 1,02; 86,41 e 43,88. Já a taxa de acertos RFMP foi de 47,71%, sendo impactado pelas taxas de P (67,65%) e R (70,59%).

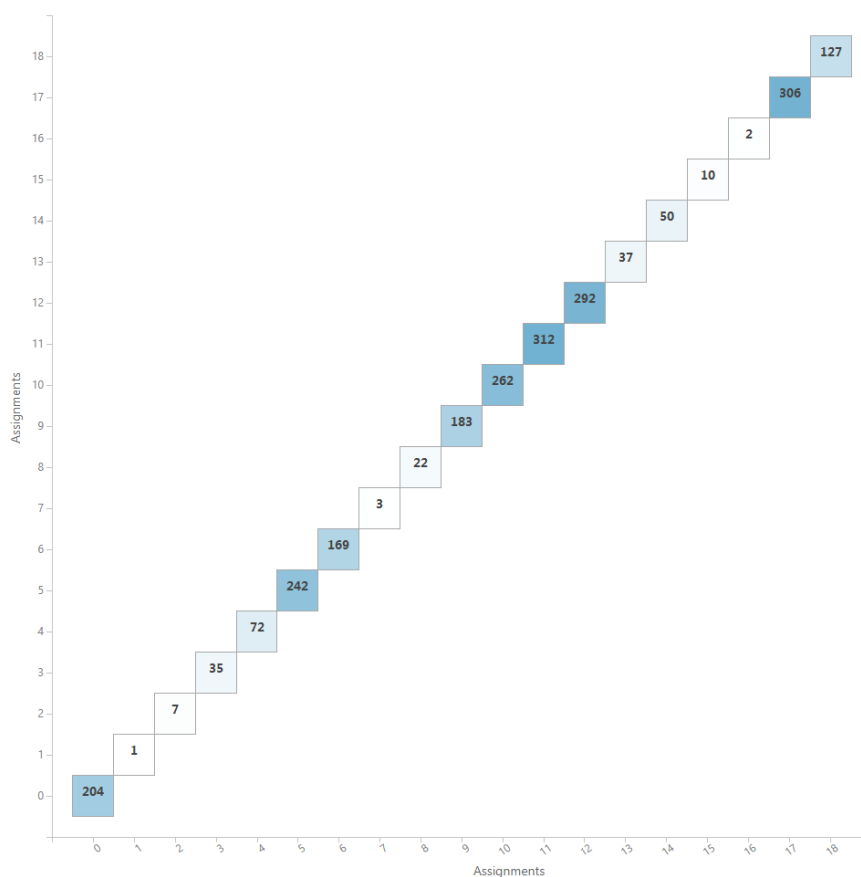


Figura 5.8: Quantidade de registros por *cluster* da 1ª análise RFMP com a métrica *Average Deviation*.

O *cluster* 12 foi o próximo a ser analisado. Com 292 registros (12,50%) e classificação RFM+ e P-, este *cluster* representa os clientes recentes, frequentes, monetários, mas não lucrativos. As médias de R, F, M e P ficaram em: 59,22; 1,22; 328,30 e 43,93. Já a taxa de acertos RFMP foi de apenas 10,27%, sofrendo impacto direto dos resultados obtidos pelas taxas de F (13,36%). Neste *cluster*, a taxa de acerto de P não teve influência, pois o resultado foi de 82,53%.

Já o *cluster* 10 ficou com 262 registros (11,22%) e classificação RF+ e MP-, o que significa um grupo de clientes recentes, frequentes, mas não monetários e nem lucrativos. As médias de R, F, M e P ficaram em: 53,92; 1,32; 200,94 e 44,38. Já a taxa de acertos RFMP foi de somente 11,45%, influenciada pelos resultados das taxas de F (20,61%) e P (61,83%).

Em seguida o *cluster* 5, que ficou com 242 registros (10,36%) e classificação P+ e RFM-, ou seja, clientes lucrativos, mas não recentes, não frequentes e nem monetários. As médias de R, F, M e P ficaram em: 294,47; 1,01; 93,00 e 45,46. Já a taxa de acertos

RFMP foi de somente 50,41%, com impacto direto dos resultados obtidos pela taxa do atributo P (51,65%).

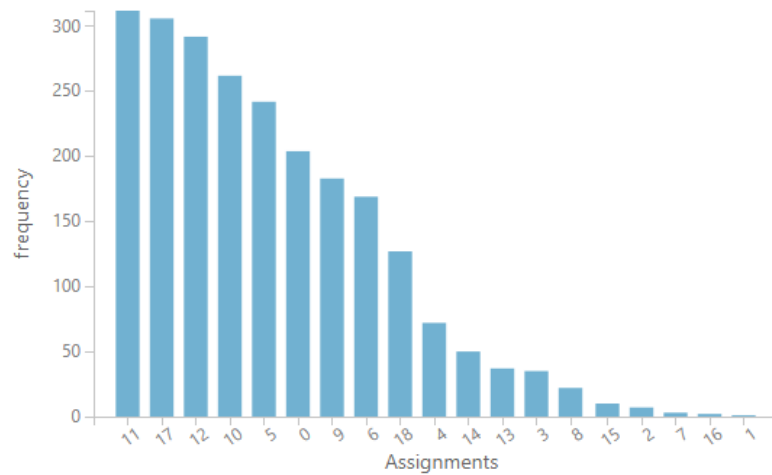


Figura 5.9: Histograma dos *clusters* da 1ª análise RFMP com a métrica *Average Deviation*.

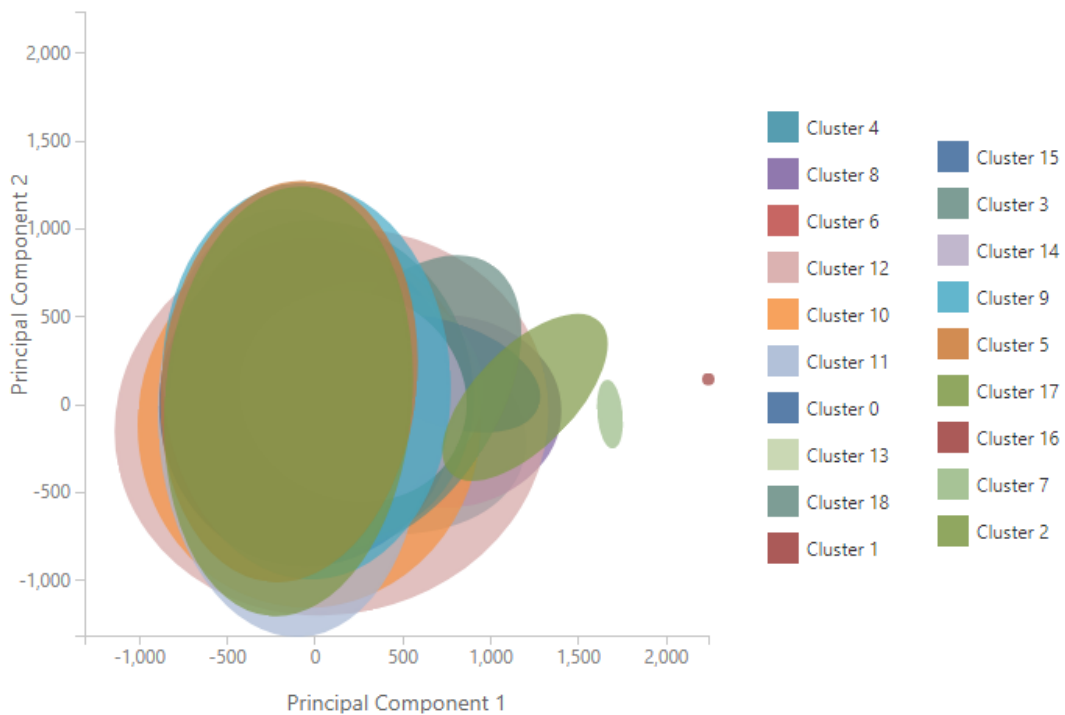


Figura 5.10: Gráfico da visualização dos *clusters* da 1ª análise RFMP com a métrica *Average Deviation*.

O próximo foi o *cluster* 0, com 204 registros (8,73%) e classificação F+ e RMP-, isto é, clientes frequentes, mas não recentes, não monetários e nem lucrativos. As médias de R, F, M e P ficaram em: 168,25; 1,20; 198,05 e 44,30. Já a taxa de acertos RFMP foi

muito baixa, com somente 2,45% de acertos, influenciada pelas taxas de R (61,27%), F (15,20%) e P (66,18%).

Agora o *cluster* 9, com 183 registros (7,83%) e classificação RFMP-, comportamento semelhante ao encontrado pelo *cluster* 17. As médias de R, F, M e P ficaram em: 292,73; 1,03; 192,91 e 44,18. Já a taxa de acertos RFMP foi de 52,46%, sendo influenciada pela taxa de P (58,47%).

O *cluster* 6 ficou com 169 registros (7,23%) e classificação FM+ e RP-, ou seja, clientes frequentes, monetários, mas não recentes e nem lucrativos. As médias de R, F, M e P ficaram em: 170,73; 1,17; 316,82 e 43,27. Já a taxa de acertos RFMP foi uma das mais baixas do modelo, com apenas 2,37% de acertos, influenciada pelas taxas de R (60,36%) e F (11,24%), enquanto que a taxa de P foi muito boa, com um resultado de 88,76%.

Na continuidade, o *cluster* 18, com 127 registros (5,44%) e classificação MP+ e RF-, o que sugere um conjunto de clientes monetários, lucrativos, mas não recentes e nem frequentes. As médias de R, F, M e P ficaram em: 294,33; 1,03; 322,53 e 44,65. Enquanto que a taxa de acertos RFMP foi de 42,52%, influenciada pela taxa de P (44,88%).

Já o *cluster* 4 ficou com 72 registros (3,08%) e classificação RFMP+, o que representa um grupo com clientes importantes (recentes, frequentes, monetários e lucrativos). As médias de R, F, M e P ficaram em: 64,65; 1,51; 504,19 e 44,65. Em compensação a taxa de acertos RFMP foi de 6,94%, influenciada pelas taxas de F (16,67%) e P (50%).

Na sequência os *clusters* 14 e 13. O primeiro ficou com 50 registros (2,14%), enquanto que o segundo com 37 registros (1,58%). A classificação de ambos foi de MP+ e RF-, comportamento idêntico ao encontrado no *cluster* 18. Já as médias de R ficaram em 199,94 e 283,27 respectivamente; as de F ficaram em 1,10 para ambos os *clusters*; as de M ficaram em 452,00 e 574,97; e as de P ficaram em 44,69 e 45,34. Quanto à taxa de acertos RFMP, o *cluster* 14 obteve uma taxa de 32%, influenciada pela taxa de P (44%); enquanto que o *cluster* 13 obteve uma taxa de acertos RFMP de 70,27% (um dos maiores apresentados por este modelo), com taxa de acerto de P de 75,68%.

Os *clusters* 3 e 8 também merecem ser avaliados em conjunto, pois ficaram com classificação RFM+ e P-, igual à encontrada no *cluster* 12. Nos detalhes, o *cluster* 3 possui

35 registros (1,50%) e o *cluster* 8 possui 22 registros (0,94%). As médias de R ficaram em 155,05 e 96 respectivamente; as de F ficaram em 1,20 e 1,40; as de M 719,66 e 956,58; e as de P ficaram em 43,64 e 44,30. Já a taxa de acertos RFMP foi péssima, com um valor de 2,86% no *cluster* 3 e 0% no *cluster* 8! Ambos impactados pelas taxas de F (20% no *cluster* 3 e 4,55% no *cluster* 8) e de P (54,29% no *cluster* 3 e 22,73% no *cluster* 8).

Em relação ao *cluster* 15, com 10 registros (0,43%) e classificação MP+ e RF-, ele ficou com um comportamento semelhante ao encontrado nos *clusters* 14 e 13. Suas médias de R, F, M e P ficaram em: 309,02; 1,10; 991,70 e 45,42. Já a taxa de acertos RFMP foi de 70%, com uma taxa de acerto de P de 80% (um dos maiores obtidos neste modelo).

Por último os *clusters* 2, 7, 16 e 1. Eles representam exatamente os mesmos *clusters* encontrados por esta métrica na análise RFM. A diferença é que o *cluster* 7 ganhou um atributo P+ a mais (clientes lucrativos), enquanto que os demais ganharam um atributo P- (clientes não lucrativos). Já a média do atributo P para cada um destes *clusters* ficou em: 41,06; 45,01; 40,63 e 34,68 respectivamente. Em relação à taxa de acertos RFMP, os resultados obtidos foram de: 14,29%; 66,67%; 100% e 100%. Influenciaram nos resultados dos dois primeiros *clusters* as taxas de P (71,43% no *cluster* 2 e 66,67% no *cluster* 7). Além disso, a taxa do atributo R encontrada no *cluster* 2 foi de apenas 42,86%.

Finalizadas as análises individuais dos *clusters* gerados pela métrica *Average Deviation*, o estudo avança para entender o modelo produzido e distinguir o comportamento de separação entre os *clusters*.

Inicialmente foi tratada a relação entre os *clusters* 11, 12, 10, 4 e 7 que possuem as seguintes médias de R: 56,44; 59,22; 53,92; 64,65 e 52,66 respectivamente. Apesar de todos serem considerados recentes, a classificação RFMP de cada um varia em relação aos outros. Fica claro na comparação destes *clusters* que o atributo M foi o critério de separação entre eles, já que no *cluster* 11 encontram-se os registros entre 9,10 e 145,90; no *cluster* 10 os registros de 150,00 a 265,80; no 12 os registros entre 266,38 e 413,99; e no *cluster* 4 os que estão entre 423,03 e 635,29. O único que escapa desse efeito de escala é o *cluster* 7, com valores definidos entre 1.896,75 e 1.939,87. Destes cinco *clusters* apenas o 7 e o 11 conseguiram uma taxa de acertos RFMP maior do que 60%, mas apenas o último possui um bom volume de registros. Em compensação, o *cluster* 12 foi o que

teve a maior taxa de acerto de P entre os *clusters* com mais de 1% dos registros, com um resultado de 82,53%.

Os *clusters* 17, 6, 0, 14 e 16 possuem comportamentos semelhantes no atributo R, com médias: 181,07; 168,25; 170,73; 199,94 e 200,00. Ambos formados por clientes não recentes, porém a classificação RFMP é diferente em cada agrupamento. Novamente o critério de separação foi o atributo M, seguindo os mesmos critérios adotados nos *clusters* anteriores. No caso, o *cluster* 17 possui registros entre 11,90 e 138,00; o *cluster* 0 possui os registros que estão entre 141,33 e 255,09; o *cluster* 6 entre 261,96 e 384,59; e o *cluster* 14 entre 385,50 e 539,98. O *cluster* 16 foi o único a fugir deste padrão com um valor bem acima dos demais, entre 1.920,37 e 2.049,24. Destes *clusters*, o único que foi considerado como P+ (lucrativo) foi o *cluster* 14. Neste caso, se comparado então com os *clusters* da análise do parágrafo anterior, pode-se até inferir que os clientes mais recentes são mais lucrativos que os mais antigos. Mas com uma classificação RMFP tão inconsistente, é preciso um pouco de prudência antes de se fazer este tipo de afirmação.

Também foi possível encontrar este padrão de separação pelo atributo M entre os *clusters* 3, 8 e 2, com médias de R de: 155,05; 96,00 e 163,14. Neste caso, o *cluster* 3 possui os valores contidos entre 654,99 e 849,99; o *cluster* 8 entre 844,99 e 1.119,99; e o *cluster* 2 entre 1.260,05 e 1.545,40. As taxas de acertos de R nestes 3 *clusters* ficaram baixas, isto porque a média deste atributo está justamente próxima à faixa de valores dos dados contidos nestes segmentos. De fato, esta é uma questão que deve ser enfrentada para melhorar a análise dos modelos RFM e RFMP.

Padrão semelhante foi encontrado também nos *clusters* 5, 9, 18, 13 e 15, com médias de R de: 294,47; 292,73; 294,33; 283,27 e 309,20. Nestes *clusters* o intervalo de M ficou em: 36,40 e 142,00 no *cluster* 5; 144,09 e 254,89 no *cluster* 9; 262,99 e 419,00 no *cluster* 18; 492,5 e 729,44 no *cluster* 13; e 828,8 e 1.082,86 no *cluster* 15. Além disso, as taxas de acerto do atributo P ficaram baixas, seguindo o mesmo padrão do que foi relatado no parágrafo anterior com o atributo R, ou seja, a média do atributo P ficou próxima à faixa dos valores contidos em cada *cluster*.

Entretanto, assim como nas análises anteriores, os *clusters* não foram separados somente pelo valor do atributo M, como foi o caso dos *clusters* 11, 17 e 5. Ambos tinham valores de M muito próximos, com médias de 86,41; 92,34 e 93,00 respectivamente. Entretanto, as médias de R ficaram distintas, com valores de: 181,07; 56,44 e 294,47. O

mesmo ocorreu com os *clusters* 10, 0 e 9, que ficaram com médias de M de: 200,94; 198,05 e 192,91. Porém, com médias de R de: 53,92; 168,25 e 292,73. Além destes *clusters*, este comportamento foi compartilhado também pelos agrupamentos 12, 6 e 18, onde foi possível observar que as médias de M ficaram semelhantes, mas em compensação a variação nas médias de R ficaram diferentes.

Desta forma o estudo chega à parte final da avaliação da métrica *Average Deviation* a partir do modelo de classificação RFMP. Pelos resultados deste índice, 19 *clusters* foram criados com uma taxa média de acertos dos *clusters* de 39,12% (muito inferior aos 58,32% encontrados na análise RFM) e taxa de acertos RFMP do modelo de apenas 31,55% (abaixo dos 55,14% da análise RFM). Sendo assim, o estudo dá por encerrada a avaliação do recurso *Sweep Clustering* para a criação de modelos RFMP.

### **5.2.5 Síntese dos Resultados – RFMP**

O objetivo desta análise foi avaliar se a inclusão do atributo P traria algum impacto na formação dos *clusters* em contrapartida aos que foram criados pelo modelo RFM. A ideia era tentar diferenciar os grupos de clientes não só pelo valor monetário – conforme o padrão de comportamento dos modelos RFM – mas também pela rentabilidade.

Como primeira observação foi possível notar que a qualidade dos modelos criados pela classificação RFMP ficou inferior àquelas apresentadas pela análise RFM. Afinal, nenhum modelo gerado teve a taxa de acertos RFMP superior às apresentadas na análise RFM. Isto se deve não só pela inclusão do novo atributo, o que fez com que o processo de acerto e classificação ficasse um pouco mais impreciso e complexo, mas também pelo fato da maioria das médias dos *clusters* terem ficado muito próximas à da média global, o que tornou inconsistente a classificação pelo atributo P.

Além disso, outros motivos – conforme já descrito na análise RFM – foram observados como causa do baixo desempenho, como por exemplo, a grande concentração dos registros em um único valor para o atributo F e a baixa distribuição dos dados entre os *clusters*. Estes itens justificam em parte a grande sobreposição encontrada nos gráficos de visualização apresentados pelo estudo.

Deste modo, é possível afirmar que a inclusão do atributo P trouxe uma complexidade a mais para a assertividade dos *clusters*, tornando inconsistentes as

classificações RFMP atribuídas a eles. Isto se torna evidente ao analisar as taxas médias de acertos dos *clusters* e as taxas globais de acertos dos modelos, uma vez que o melhor modelo produzido pela métrica *Simplified Silhouette*, com 3 *clusters*, obteve um resultado médio de acertos dos *clusters* de apenas 21,26%, e resultado de acertos RFMP do modelo de 19,14%; enquanto que as produzidas pela métrica *Davies-Bouldin*, com 9 *clusters*, conseguiram uma taxa média de acertos dos *clusters* de 28,84%, e taxa de acertos RFMP do modelo de 22,47%; já as produzidas pela métrica *Dunn* (com 8 *clusters*) ficaram com uma taxa média de acertos dos *clusters* de 29,20%, e taxa de acertos RFMP do modelo de 22%; e por último, com 19 *clusters*, a partir da métrica *Average Deviation*, foi encontrada uma taxa média de acertos dos *clusters* de 39,12% e uma taxa de acertos RFMP do modelo de 31,55%.

Além disso, mais uma vez não foi possível encontrar um modelo que atendesse a premissa da pesquisa. Ainda assim, esta análise ajudou a compreender a formação dos modelos e os critérios que afetaram a separação entre os *clusters*, como o fato da distorção (assimetria) do atributo F estar interferindo na qualidade dos modelos, assim como a comparação por um valor absoluto (no caso a média global) não trazer um resultado tão adequado. Ambos os fatores já foram identificados e serão tratados mais à frente.

## 5.2.6 Tabelas Auxiliares – RFMP

Tabela 5.16: Tabela auxiliar com os resultados da métrica *Simplified Silhouette* da análise RFMP.

G	Qtd.	%	RFMP +	RFMP -	Med R	Min R	Max R	Med F	Min F	Max F	Med M	Min M	Max M	Med P	Min P	Max P	% de Acertos				Qtd. Acertos RFMP	% Acertos RFMP
																	% R	% F	% M	% P		
0	2.041	87,37		RFMP	158,61	1	365	1,12	1	8	183,41	9,10	373,09	44,47	7,28	100	48,11	91,08	70,26	61,59	430	21,07
2	282	12,07	RFMP		148,42	1	365	1,28	1	11	561,90	372,79	1.119,99	44,68	34,15	52,62	54,96	13,48	100,00	53,19	12	4,26
1	13	0,56	RM	FP	148,84	4	313	1,07	1	2	1.709,78	1.260,05	2.480,00	41,41	34,68	47,46	53,85	92,31	100,00	69,23	5	38,46

Tabela 5.17: Tabela auxiliar com os resultados da métrica *Davies-Bouldin* da análise RFMP.

G	Qtd.	%	RFMP +	RFMP -	Med R	Min R	Max R	Med F	Min F	Max F	Med M	Min M	Max M	Med P	Min P	Max P	% de Acertos				Qtd. Acertos RFMP	% Acertos RFMP
																	% R	% F	% M	% P		
0	725	31,04	RFP	M	80,90	1	173	1,16	1	6	132,42	9,10	226,90	44,71	23,47	80	92,97	11,45	100,00	47,72	38	5,24
5	660	28,25	P	RFM	256,55	166	365	1,04	1	5	121,62	11,90	211	44,51	36,33	83,31	100	96,67	100,00	40,00	252	38,18
6	459	19,65	RFM	P	72,98	1	165	1,28	1	11	319,19	222,75	470,09	44,37	35,23	100	98,26	16,34	94,34	76,47	45	9,80
4	295	12,63	M	RFP	249,76	159	365	1,10	1	8	310,37	217,90	465,20	43,97	7,28	83,3	100	92,88	92,20	66,10	180	61,02
3	139	5,95	FMP	R	163,56	3	364	1,15	1	8	555,96	426,49	734,89	44,85	35,23	51,35	50,36	10,07	100,00	58,27	3	2,16
8	45	1,93	FMP	R	171,15	4	354	1,28	1	10	915,26	743,99	1.119,99	44,50	34,15	47,27	55,56	11,11	100,00	77,78	4	8,89
2	7	0,30	M	RFP	163,14	50	313	1	1	1	1.433,59	1.260,05	1.545,40	41,06	37,87	47,46	42,86	100	100,00	71,43	1	14,29
7	5	0,21	RM	FP	111,60	4	206	1	1	1	1.942,41	1.896,75	2.049,24	43,26	38,87	45,74	60,00	100	100,00	60,00	1	20,00
1	1	0,04	FM	RP	235	235	235	2	2	2	2.480	2.480,00	2.480,00	34,68	34,68	34,68	100	100	100,00	100,00	1	100,00

Tabela 5.18: Tabela auxiliar com os resultados da métrica *Dunn* da análise RFMP.

G	Qtd.	%	RFMP +	RFMP -	Med R	Min R	Max R	Med F	Min F	Max F	Med M	Min M	Max M	Med P	Min P	Max P	% de Acertos				Qtd. Acertos RFMP	% Acertos RFMP
																	% R	% F	% M	% P		
0	765	32,75	RFP	M	82,79	1	179	1,16	1	6	130,42	9,1	238,9	44,76	23,47	80	89,54	11,37	99,87	48,10	40	5,23
5	732	31,34		RFMP	266,40	170	365	1,04	1	5	146,46	11,9	362,99	44,44	7,28	83,31	100	96,17	87,84	57,51	376	51,37
6	625	26,76	RFM	P	116,21	1	334	1,23	1	11	320,11	217,07	455,29	44,17	35,23	100	69,76	13,60	95,68	79,04	45	7,20
4	153	6,55	MP	RF	171,67	3	365	1,13	1	8	538,39	387,5	719,88	44,82	35,23	51,35	52,29	91,50	100	57,52	48	31,37
3	48	2,05	FM	RP	174,12	4	354	1,31	1	10	903,68	725,42	1.119,99	44,34	34,15	47,27	58,33	14,58	100	25	2	4,17
2	7	0,30	M	RFP	163,14	50	313	1	1	1	1433,59	1.260,05	1.545,40	41,06	37,87	47,46	42,86	100	100	71,43	1	14,29
7	5	0,21	RM	FP	111,60	4	206	1	1	1	1942,41	1.896,75	2.049,24	43,26	38,87	45,74	60,00	100	100	60	1	20,00
1	1	0,04	FM	RP	235	235	235	2	2	2	2.480	2.480,00	2.480,00	34,68	34,68	34,68	100	100	100	100	1	100,00

Tabela 5.19: Tabela auxiliar com os resultados da métrica *Dunn* da análise RFMP *Average Deviation*.

G	Qtd.	%	RFMP +	RFMP -	Med R	Min R	Max R	Med F	Min F	Max F	Med M	Min M	Max M	Med P	Min P	Max P	% de Acertos				Qtd. Acertos RFMP	% Acertos RFMP
																	% R	% F	% M	% P		
11	312	13,36	RP	FM	56,44	1	118	1,10	1	6	92,34	9,10	145,90	45,83	23,47	80	100	91,35	100	67,63	189	60,58
17	306	13,10		RFMP	181,07	119	237	1,02	1	2	86,41	11,90	138	43,88	36,33	69,93	70,59	97,71	100	67,65	146	47,71
12	292	12,50	RFM	P	59,22	1	115	1,22	1	8	328,30	266,38	413,99	43,93	37,88	59,71	100	13,36	100	82,53	30	10,27
10	262	11,22	RF	MP	53,92	1	110	1,32	1	6	200,94	150	265,80	44,38	35,67	65	100	20,61	87,02	61,83	30	11,45
5	242	10,36	P	RFM	294,47	237	365	1,01	1	2	93,00	36,40	142	45,46	38,86	83,31	100	98,35	100	51,65	122	50,41
0	204	8,73	F	RMP	168,25	112	230	1,20	1	5	198,05	141,33	255,09	44,30	35,23	100	61,27	15,20	89,71	66,18	5	2,45
9	183	7,83		RFMP	292,73	230	365	1,03	1	2	192,91	144,09	254,89	44,18	35,23	83,30	100	96,17	89,07	58,47	96	52,46
6	169	7,23	FM	RP	170,73	110	233	1,17	1	8	316,82	261,96	384,59	43,27	37,88	50,35	60,36	11,24	100	88,76	4	2,37
18	127	5,44	MP	RF	294,33	233	365	1,03	1	3	322,53	262,99	419	44,65	7,28	83,3	100	97,64	100	44,88	54	42,52

4	72	3,08	RFMP		64,65	1	138	1,51	1	11	504,19	423,03	635,29	44,65	35,23	49,64	100	16,67	100	50,00	5	6,94
14	50	2,14	MP	RF	199,94	142	321	1,1	1	2	452,00	385,50	539,98	44,69	39,27	51,35	86,00	90,00	100	44,00	16	32,00
13	37	1,58	MP	RF	283,27	192	364	1,10	1	3	574,97	492,50	729,44	45,34	40,06	48,85	100	91,89	100	75,68	26	70,27
3	35	1,50	RFM	P	155,05	25	317	1,2	1	2	719,66	654,99	849,99	43,64	34,15	48,85	45,71	20,00	100	54,29	1	2,86
8	22	0,94	RFM	P	96	4	209	1,40	1	10	956,58	844,99	1.119,99	44,30	37,72	46,71	77,27	4,55	100	22,73	0	0,00
15	10	0,43	MP	RF	309,20	207	354	1,1	1	2	991,70	828,80	1.082,86	45,42	39,45	47,27	100	90,00	100	80,00	7	70,00
2	7	0,30	M	RFP	163,14	50	313	1	1	1	1433,59	1.260,05	1.545,40	41,06	37,87	47,46	42,86	100	100	71,43	1	14,29
7	3	0,13	RMP	F	52,66	4	128	1	1	1	1914,15	1.896,75	1.939,87	45,01	44,18	45,74	100	100	100	66,67	2	66,67
16	2	0,09	M	RFP	200	194	206	1	1	1	1984,80	1.920,37	2.049,24	40,63	38,87	42,39	100	100	100	100	2	100
1	1	0,04	FM	RP	235	235	235	2	2	2	2.480,00	2.480,00	2.480,00	34,68	34,68	34,68	100	100	100	100	1	100

## 5.3 Análise RFM (Z-SCORE)

Para tentar melhorar um pouco mais os resultados encontrados até o momento, o estudo passa agora a avaliar a criação de novos modelos a partir da normalização dos dados através do método *Z-Score*. A princípio, a proposta é a mesma da que foi apresentada nos modelos anteriores, isto é, tentar encontrar o melhor número de *K* fazendo uso do recurso *Sweep Clustering*. Em seguida, validar se o resultado atende ao critério de performance estabelecido pela pesquisa, para depois avaliar a qualidade dos *clusters* e dos modelos gerados.

### 5.3.1 Simplified Silhouette

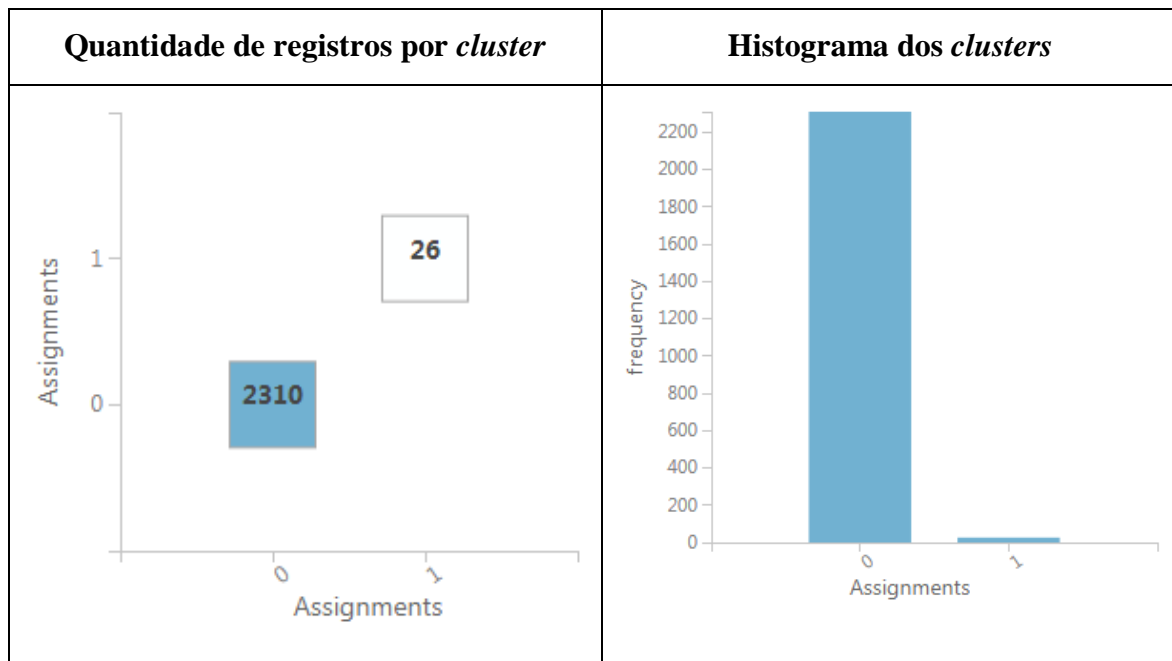
Com esta métrica o melhor modelo sugerido obteve um resultado com apenas dois *clusters*, conforme pode ser visto na Tabela 5.20. Este é um resultado diferente do que foi encontrado pela primeira análise RFM, já que naquela o modelo produzido por esta métrica encontrou um valor com 3 *clusters*. Além disso, tanto a pontuação máxima quanto a ordem subsequente com os demais valores de *K* também apresentaram variações.

Tabela 5.20: Pontuação da métrica *Simplified Silhouette* da análise RFM (*Z-Score*).

Pontuação	Número de <i>Clusters</i>	Pontuação	Número de <i>Clusters</i>
0,728182	2	0,549924	20
0,599744	8	0,547327	16
0,597767	4	0,547152	19
0,592846	7	0,546326	18
0,588271	5	0,543869	15
0,572104	6	0,543782	11
0,569978	9	0,542736	17
0,558577	3	0,541688	13
0,552223	12	0,539866	14
0,549968	10		

Ainda assim, de acordo com a Tabela 5.21, o maior agrupamento ficou com 2.310 registros (98,89%) e o menor com 26 (1,11%). Este é um resultado que atende ao critério estabelecido. Porém, como os dados ficaram concentrados em apenas dois *clusters*, a qualidade obtida pelo modelo ficou comprometida.

Tabela 5.21: Gráficos dos resultados da métrica *Simplified Silhouette* da análise RFM (Z-Score).



O gráfico da Figura 5.11 ilustra a representação destes dois *clusters*. E novamente pode-se observar que eles continuaram sobrepostos.

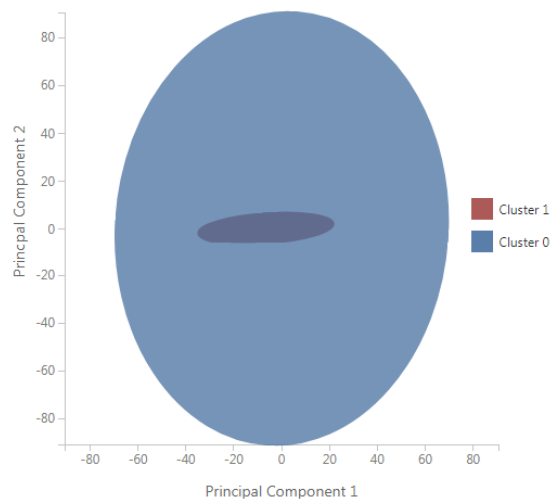


Figura 5.11: Gráfico da visualização dos *clusters* da análise RFM (Z-Score) com a métrica *Simplified Silhouette*.

Na sequência, é detalhado o comportamento de cada *cluster* conforme demonstrado na Tabela 5.25 (tabela auxiliar).

O *cluster* 0 possui classificação RFM- (clientes não recentes, não frequentes e nem monetários), com médias de R, F e M de: 158,40; 1,09 e 236,71. A taxa de acertos

RFM foi de 30,56%, impactada pelas taxas de R (45,14%) e M (61,60%). Entretanto, é interessante notar que a taxa de F foi bastante elevada, com um valor de 91,56%.

Já o *cluster* 1 possui classificação RFM+ (clientes recentes, frequentes e monetários), um comportamento justamente contrário ao do anterior. As médias de R, F e M ficaram em: 62,34; 5,61 e 316,52. A taxa de acertos RFM foi de 53,85%, influenciada pela taxa de M (54,69%).

Quanto à comparação entre os dois *clusters*, é possível observar que desta vez o modelo utilizou o atributo F como critério de separação, deixando os clientes com frequência entre 1 e 3 para o *cluster* 0, e os clientes com frequência maior ou igual a 4 para o *cluster* 1. Uma separação de valores razoável para este atributo.

Já em relação à taxa média de acertos dos *clusters*, o resultado obtido foi de apenas 42,20% (valor superior aos 32,47% encontrados na primeira análise RFM), enquanto que a taxa de acertos RFM do modelo foi de 30,82% (valor inferior aos 31,59% encontradas na primeira análise RFM). Apesar da melhora nos resultados obtidos, as classificações dos *clusters* não ficaram consistentes.

De todo modo, o estudo já consegue perceber uma ligeira alteração na criação do modelo, inclusive pela possibilidade de separação dos *clusters* através do atributo F. Porém, os resultados apresentados não podem ser considerados coerentes com o que a pesquisa deseja.

Sendo assim, é finalizada a análise desta métrica com os dados padronizados a partir da normalização *Z-Score*.

### **5.3.2 *Davies-Bouldin***

Esta métrica obteve como melhor valor de *K* sugerido o mesmo valor encontrado na métrica *Simplified Silhouette* (2 *clusters*), por isso o estudo não fará qualquer avaliação em relação aos *clusters* produzidos, já que os conteúdos ficaram idênticos. De qualquer forma, vale ressaltar que a pontuação encontrada pelo índice foi diferente, assim como a ordenação do número de *K*, conforme pode ser visto na Tabela 5.22.

Tabela 5.22: Pontuação da métrica *Davies-Bouldin* da análise RFM (*Z-Score*).

<b>Pontuação</b>	<b>Número de <i>Clusters</i></b>		<b>Pontuação</b>	<b>Número de <i>Clusters</i></b>
0,549727	2		0,754029	10
0,71385	7		0,759901	18
0,717102	8		0,766354	17
0,720899	12		0,772628	11
0,722881	14		0,790349	4
0,728557	13		0,802092	9
0,743439	20		0,824489	5
0,744457	16		0,830081	6
0,745139	19		0,882793	3
0,751596	15			

### 5.3.3 *Dunn*

Assim como na análise anterior, esta métrica também sugeriu como melhor valor de *K* um modelo com 2 *clusters*. Mas novamente, vale destacar que tanto a pontuação encontrada quanto a ordenação do número de *K* ficaram diferentes, como demonstra a Tabela 5.23.

Tabela 5.23: Pontuação da métrica *Dunn* da análise RFM (*Z-Score*).

<b>Pontuação</b>	<b>Número de <i>Clusters</i></b>		<b>Pontuação</b>	<b>Número de <i>Clusters</i></b>
0,655842	2		0,405955	12
0,466432	19		0,40385	18
0,461927	20		0,396417	11
0,460713	6		0,390535	8
0,457894	5		0,334583	14
0,428861	16		0,331235	13
0,428861	15		0,301198	4
0,422076	17		0,299019	3
0,416066	7		0,270584	9
0,406326	10			

### 5.3.4 Average Deviation

Das quatro métricas avaliadas nesta análise, esta foi a única que obteve um resultado diferente das demais. Neste caso, o melhor modelo sugerido ficou com um valor de  $K$  igual a 20 (valor semelhante ao encontrado por esta métrica na análise RFM). De todo modo, o valor da pontuação obtido nesta análise – conforme pode ser visto na Tabela 5.24 – ficou diferente da que foi apresentada na primeira análise RFM.

Tabela 5.24: Pontuação da métrica *Average Deviation* da análise RFM (*Z-Score*).

Pontuação	Número de <i>Clusters</i>	Pontuação	Número de <i>Clusters</i>
0,421769	20	0,69031	10
0,438438	19	0,692245	9
0,459932	18	0,75049	8
0,490611	17	0,758808	7
0,495687	16	1,054807	6
0,497835	15	1,126279	5
0,505771	14	1,160806	4
0,510323	13	1,193762	3
0,544725	12	1,303976	2
0,61535	11		

Além disso, o conteúdo dos *clusters* também ficou diferente, já que o maior ficou com um total de 348 registros (14,90%) e o menor com apenas 1 registro (0,04%), conforme pode ser visto na Figura 5.12 e na Figura 5.13. Este é um resultado que não atende ao critério estabelecido. Porém, esta análise possui uma distribuição muito melhor do que as outras. Resta apenas saber se a qualidade do modelo está consistente com a classificação RFM atribuída a cada um dos *clusters*.

Já o gráfico da Figura 5.14 ilustra a representação dos *clusters* produzidos. Sendo que mais uma vez foi possível notar que a sobreposição das elipses não permitiu uma análise mais aprofundada. Principalmente por não ser possível observar a representação de todos os *clusters*.

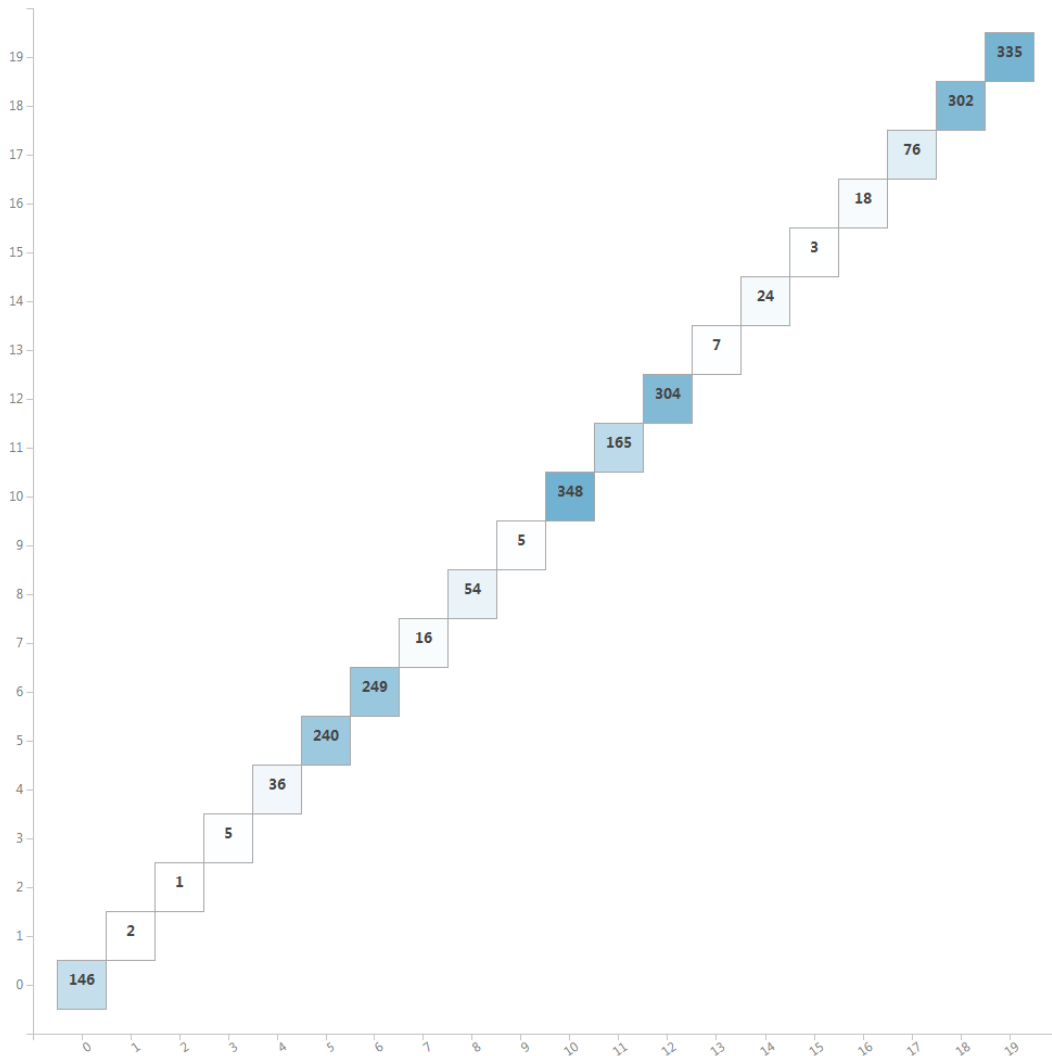


Figura 5.12: Quantidade de registros por *cluster* da análise RFM (*Z-Score*) com a métrica *Average Deviation*.

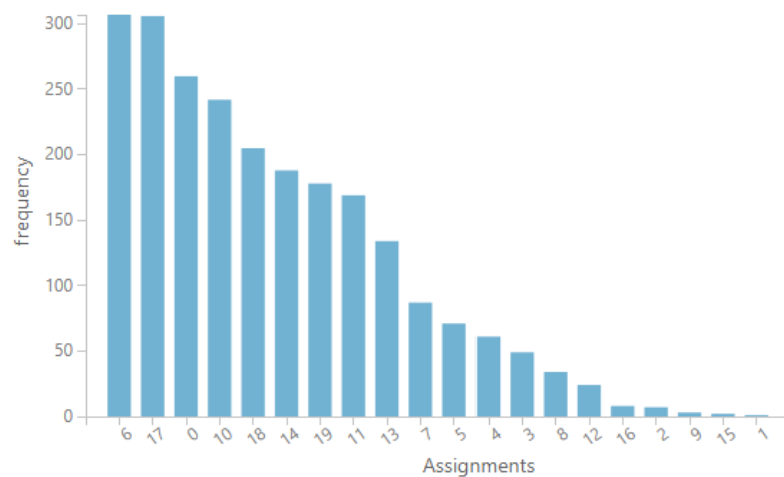


Figura 5.13: Histograma dos *clusters* da análise RFM (*Z-Score*) com a métrica *Average Deviation*.

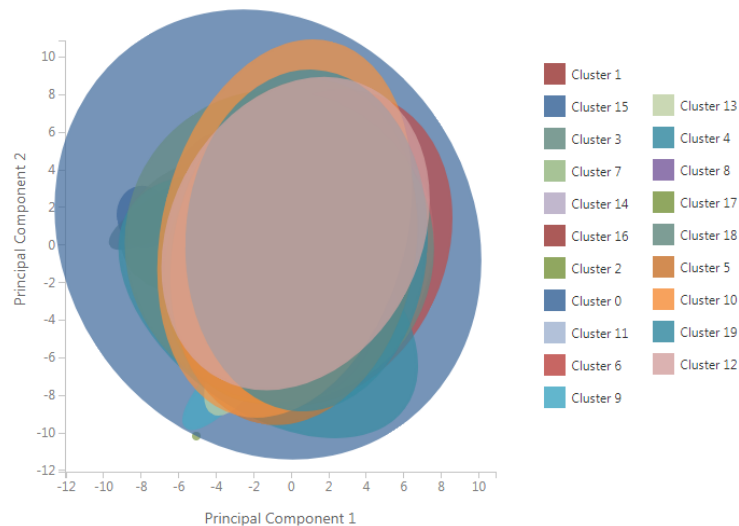


Figura 5.14: Gráfico da visualização dos *clusters* da análise RFM (*Z-Score*) com a métrica *Average Deviation*.

Em seguida é detalhado o comportamento de cada *cluster* de acordo com os resultados apresentados na Tabela 5.26.

De início o maior de todos, o *cluster* 10, com 348 registros (14,90%) e classificação R+ e FM-, representa um grupo de clientes recentes, mas não frequentes e nem monetários. Este comportamento é semelhante ao encontrado no maior *cluster* da análise RFM discutida no tópico 5.1.4, só que com médias de 41,64 em R; 1 em F e 132,90 em M. Já a taxa de acertos RFM foi de 99,71%, um resultado muito positivo principalmente para um grupo com tantos clientes, o que justifica a classificação RFM atribuída ao *cluster*.

Na sequência o *cluster* 19, com 335 registros (14,34%) e classificação RFM-, ou seja, clientes não frequentes, não recentes e nem monetários. Análogo ao comportamento encontrado no segundo maior *cluster* da análise RFM, porém com médias de R, F e M de: 223,89; 1; e 116,52. Quanto à taxa de acertos RFM o resultado obtido foi de 100%. Um resultado excepcional para um *cluster* com tantos registros, e que fundamenta a classificação RFM associada.

Em seguida o *cluster* 12, com 304 registros (13,01%) e classificação R+ e FM- (clientes recentes, mas não frequentes e nem monetários), e que possui as seguintes médias de R, F e M: 128,98; 1 e 116,32. Este comportamento é bastante distinto se comparado ao terceiro maior *cluster* da análise RFM. Já em relação à taxa de acertos RFM, o resultado obtido foi de 80,92%.

O próximo foi o *cluster* 18, com 302 registros (12,93%) e classificação RM+ e F- (recentes, frequentes, mas não monetários), com respectivas médias de R, F e M de: 58,10; 1; e 323,33, e taxa de acertos RFM de 97,02%. Mais um resultado positivo e que confirma a categorização RFM atribuída.

Já o *cluster* 6 possui 249 registros (10,66%), classificação RFM- (comportamento semelhante ao encontrado no *cluster* 19), e médias de R, F e M de: 316,73; 1,01; e 121,15. O resultado da taxa de acertos RFM foi de 98,90%, outro bom resultado.

Dando continuidade, a análise do *cluster* 5, com 240 registros (10,27%) e classificação M+ e RF-, isto é, clientes monetários, mas não recentes e nem frequentes, e com médias de R, F e M de: 174,24; 1; e 295,47. Quanto à taxa de acertos RFM, o resultado encontrado foi de apenas 51,25%, influenciado pelas taxas de R (68,75%) e M (77,08%). Nota-se aqui que este foi o primeiro *cluster* com um resultado insatisfatório, mas chama a atenção que o atributo F não teve qualquer influência nesta causa.

Agora o *cluster* 11, com 165 registros (7,06%) e classificação M+ e RF- (comportamento semelhante ao do *cluster* anterior). Já as médias encontradas foram de: 296,36 (R); 1,01 (F); e 300,16 (M). Já a taxa de acertos RFM foi de 83,03%. Um resultado consistente à classificação RFM conferida.

Em seguida o *cluster* 0 com 146 registros (6,25%) e classificação RF+ e M-, ou seja, clientes não recentes, não frequentes e nem monetários. As médias de R, F e M obtidas foram de: 100,02 (R), 2 (F) e 218,42 (M). A taxa de acertos RFM foi de 45,21%, influenciada pelas taxas de R (78,77%) e M (58,90%). Apesar deste baixo resultado, vale ressaltar que a taxa de F obteve um acerto de 100% e que todos os registros contidos neste grupo possuem o valor do atributo F igual a 2.

Na sequência o *cluster* 17 com 76 registros (3,25%) e classificação RM+ e F- (comportamento semelhante ao do *cluster* 18). As médias de R, F e M obtidas foram de: 94,82 (R), 1 (F) e 559,38 (M). Já a taxa de acertos RFM foi de 86,84%.

Em conjunto foram analisados os *clusters* 8 (com 54 registros – 2,31%) e 4 (com 36 registros – 1,54%). Ambos possuem classificação M+ e RF- (comportamento similar ao encontrado nos *clusters* 5 e 11). Porém as médias de R, F e M do primeiro ficaram em: 268,96; 1 e 556,54; enquanto que no segundo ficaram em: 170,13; 1 e 947,21. Em

compensação a taxa de acertos RFM do *cluster* 8 foi de 100%, e a taxa do *cluster* 4 foi de 55,56%. Esta diferença se deu pela taxa de acerto do atributo R (55,56%) obtida pelo último *cluster*.

O *cluster* seguinte foi o 14, com 24 registros (1,03%) e classificação RFM+, isto é, clientes recentes, frequentes e monetários; com médias de: 93,54 (R); 3 (F) e 254,66 (M). Em compensação este *cluster* obteve a pior taxa de acertos RFM, com um valor de 33,33%, impactada pelas taxas de R (75%) e M (45,83%). Novamente foi possível observar que o atributo F não trouxe qualquer influência negativa para o resultado.

O próximo *cluster* foi o 16, com 18 registros (0,77%) e classificação FM+ e R-, ou seja, clientes frequentes, monetários, mas não recentes. Ele possui as seguintes médias: 190,50 (R); 2,05 (F) e 611,57 (M), com taxa de acertos RFM de 72,22%, impactada pela taxa de acerto de R (72,22%).

Já o *cluster* 7 possui 16 registros (0,68%), classificação RFM+ (semelhante à encontrada no *cluster* 14) e médias de R, F e M de: 62,06; 4,31 e 285,98. Com taxa de acertos RFM de 50%, impactada pelas taxas de R (87,50%) e M (50%). Este é o *cluster* que contém os clientes com valor de F entre 4 e 5. O que demonstra que este modelo está conseguindo separar os grupos de clientes pelos valores contidos no atributo F, algo não encontrado até agora pelas análises anteriores.

O *cluster* 13 ficou idêntico ao encontrado no *cluster* 2 da análise RFM apresentado no tópico 5.1.4, por isso nenhuma nova observação será feita.

Assim, chega-se ao *cluster* 3, que possui 5 registros (0,21%) e classificação RFM+ (comportamento similar aos *clusters* 7 e 14). Com médias de R, F e M de: 45,04; 6,40 e 246,90. Porém, com uma taxa de acertos RFM de apenas 40%, impactada pela taxa de M (40%).

Em seguida o *cluster* 9, também com 5 registros (0,21%), classificação RM+ e F- (semelhante aos *clusters* 18 e 17) e médias de R, F e M de: 111,60; 1 e 1.942,41. A taxa de acertos RFM foi de 60%, impactada pela taxa de R (60%).

Já ao fim, os *clusters* 15, 1 e 2. Ambos possuem as mesmas quantidades de registros apresentadas pelos 3 menores *clusters* da análise 5.1.4. Porém, o penúltimo e o antepenúltimo não representam os mesmos *clusters* da análise RFM anterior. No *cluster*

5 (com 3 registros – 0,13%) foram encontradas as seguintes médias de R, F e M: 126,66; 8 e 376,57. Enquanto que no *cluster* 1 as médias ficaram em: 10,50; 10,50 e 644,81. Já a taxa de acertos RFM do *cluster* 15 foi de 66,67%, impactada pela taxa de R (66,67%); enquanto que a taxa de acertos RFM do *cluster* 1 foi de 100%. Quanto ao *cluster* 2, ele ficou exatamente igual ao o *cluster* 1 da análise RFM anterior, por isso nenhum detalhamento será feito aqui. Portanto, é possível não só verificar que os clientes contidos nos *clusters* 15 e 1 são diferentes dos que foram produzidos pela outra análise, como também que o modelo está mais sensível ao atributo F do que ao atributo M.

Com esta breve avaliação, já foi possível observar que os resultados obtidos até aqui ficaram bem melhores do que os que foram encontrados na análise RFM inicial. De todo modo, o estudo segue adiante para avaliar o modelo gerado e identificar o comportamento de separação entre os *clusters*.

De imediato nota-se o comportamento dos *clusters* com base nos valores contidos no atributo F. Pois como pode ser visto, os *clusters* 10, 19, 12, 18, 5, 17, 8, 4, 13 e 9 são aqueles formados apenas por clientes com a frequência de compra (atributo F) igual a 1. Algo quase não retratado nas análises anteriores. Já os *clusters* 0, 2, 16, 14, 7, 3 15 e 1 – que representam juntos 9,20% dos dados – possuem somente os registros de clientes cujos valores do atributo F são diferentes de 1.

Já os *clusters* 6 e 11 foram os únicos que misturaram clientes com atributo F igual a 1 ou 2, e produziram médias de F idênticas para ambos, com um valor de 1,01. Porém, eles se diferenciam pelos valores do atributo R, que ficou em 316,73 para o primeiro e 296,36 para o segundo, assim como pelos valores do atributo M, que foi de 121,15 para o *cluster* 6, e 300,16 para o *cluster* 11. Ainda assim, as taxas de acertos do atributo F obtidas por estes *clusters* foram excelentes, com valores que superaram 98%.

Foi possível constatar também a separação de *clusters* pelo atributo R quando comparados com os valores de M. Este foi o caso da relação entre os *clusters* 12, 19, 6 e 10, já que ambos ficaram com médias de M muito próximas: 116,32; 116,52; 121,15 e 132,90. Porém suas separações foram feitas através de uma escala do valor de R, uma vez que o *cluster* 10 ficou com valores entre 1 e 88; o *cluster* 12 entre 85 e 176; o *cluster* 19 entre 177 e 270; enquanto que o *cluster* 6 ficou com valores entre 269 e 365.

O mesmo ocorreu com os *clusters* 8 e 17, já que possuem as seguintes médias para o atributo M: 556,54 e 559,38. Porém, neste caso, ambos foram separados pelo valor de R, que no *cluster* 17 ficou com valores entre 3 e 178 (média de 94,82), enquanto que no *cluster* 8 ficaram os registros com valores entre 191 e 364 (média de 268,96).

Comportamento semelhante foi encontrado entre os *clusters* 11 e 18, com médias de M de: 300,16 e 323,33. Porém, com médias de R de: 296,36 e 58,10. É que no *cluster* 18 encontram-se os valores entre 1 e 119, enquanto que no *cluster* 11 os que possuem valor entre 238 e 365.

Houve também a separação de *clusters* feitos pelo atributo F quando comparados aos valores do atributo M. Neste caso, basta observar os *clusters* 16 e 1 com médias de M de: 611,57 e 644,85. Entretanto, é possível analisar que eles foram separados pelo atributo F, uma vez que o *cluster* 16 ficou com os registros com valores entre 2 e 3, enquanto que o *cluster* 1 ficou com valores entre 10 e 11. O mesmo pode ser observado entre os *clusters* 7 e 5, com médias de M de 285,98 e 295,47, mas com médias de F de 4,31 e 1.

Outros *clusters* separados pelo atributo F foram os: 0, 3, 14 e 7; já que o *cluster* 0 possui somente registros com valor de F igual a 2; o *cluster* 14 somente com valor de F igual a 3; o *cluster* 7 com valores entre 4 e 5; e o *cluster* 3 com valores entre 6 e 7. Porém, esta separação pelo atributo F trouxe uma influência negativa para a assertividade destes *clusters*, visto que ambos possuem taxa de acertos com valor menor ou igual a 50%.

Já a separação dos *clusters* 15 e 12, que possuem média de R de: 126,66 e 128,98, pode ser vista tanto pelo atributo F (com média de 8 para o *cluster* 15 e 1 para o *cluster* 12), quanto pelo atributo M (médias de 376,57 e 116,32 respectivamente). Em compensação, a separação entre os *clusters* 4, 13, 9 e 2 foi feita basicamente pelo valor de M, com médias de 947,21, 1.433,59, 1.942,41 e 2.480,00 respectivamente.

Vale ressaltar também o destaque para o registro do *cluster* 2, afinal ele aparece isolado como um grupo à parte em quase todos os cenários.

Por último, avaliando a qualidade do modelo produzido pela métrica *Average Deviation*, com os dados padronizados pelo método *Z-Score*, observa-se que a taxa média de acertos dos *clusters* foi bastante positiva, com um resultado de 73,17%. Entretanto, a

taxa de acertos RFM do modelo foi muito melhor, chegando a um surpreendente resultado de 84,55%.

Desta forma, o estudo conclui a última avaliação realizada por esta análise e segue adiante para uma apreciação geral sobre os resultados encontrados.

### 5.3.5 Síntese dos Resultados – RFM (*Z-Score*)

As três primeiras métricas indicaram como melhor valor de  $K$  um mesmo resultado, o que fez com que os modelos fossem separados em apenas 2 *clusters*. De qualquer forma, vale ressaltar que as pontuações encontradas por cada índice ficaram diferentes daquelas encontradas na análise RFM inicial. Além disso, as ordenações sugeridas do número de  $K$  subsequentes por cada uma das métricas também apresentaram variações.

No entanto, o modelo que apresentou um melhor desempenho foi justamente aquele que ficou com um valor de  $K$  diferente dos demais, que foi separado por 20 agrupamentos. Este modelo foi produzido pela métrica *Average Deviation*.

Isto porque com a normalização dos dados através do método *Z-Score* foi possível perceber uma melhora na performance do modelo, pois logo foi observado que o atributo F passou a ser considerado como regra de separação entre alguns *clusters*, o que reduziu as distorções apresentadas nas análises anteriores.

Isto fez com que a qualidade dos clusters melhorassem. Afinal, a assimetria deste atributo estava impactando diretamente nos resultados obtidos. Neste caso, foi possível verificar um salto de 58,32% para 73,17% na taxa média dos *clusters*, se comparado à primeira análise RFM (com o uso da métrica *Average Deviation*). Além disso, a taxa de acertos RFM passou de 55,14% para incríveis 84,55% – um desempenho espetacular.

Contudo, quando um determinado *cluster* não conseguiu uma taxa de acertos satisfatória, foi possível perceber que o atributo F não teve qualquer influência na causa. Por isso, ao utilizar o atributo F como critério de separação o modelo passou a errar menos. Algo totalmente diferente do que foi visto nas demais análises.

De todo modo, pelos resultados obtidos, é possível afirmar que o modelo gerado nesta avaliação conseguiu trazer uma qualidade maior aos *clusters*, já que tornou consistente à classificação RFM atribuída a cada um deles. Assim, foi encontrado um

resultado favorável que justifica a aplicação do método *Z-Score*, mas que ainda não atende à premissa estabelecida do estudo.

Na próxima etapa será avaliado se a normalização *Z-Score* trará algum impacto na criação dos modelos RFMP.

### 5.3.6 Tabelas Auxiliares – RFM (Z-Score)

Tabela 5.25: Tabela auxiliar com os resultados das métricas *Simplified Silhouette*, *Davies Bouldin* e *Dunn* da análise RFM (Z-Score).

G.	Qtd.	%	RFM+	RFM-	Media R	Min R	Max R	Media F	Min F	Max F	Media M	Min M	Max M	% Acertos R	% Acertos F	% Acertos M	Qtd. Acertos RFM	% Acertos RFM
0	2.310	98,89%		RFM	158,40	1	365	1,09	1	3	236,71	9,10	2.480	48,14%	91,56%	61,60%	706	30,56%
1	26	1,11%	RFM		62,34	1	219	5,61	4	11	316,52	96,26	853,12	88,46%	100,00%	57,69%	14	53,85%

Tabela 5.26: Tabela auxiliar com os resultados da métrica *Average Deviation* da análise RFM (Z-Score).

G.	Qtd.	%	RFM+	RFM-	Media R	Min R	Max R	Media F	Min F	Max F	Media M	Min M	Max M	% Acertos R	% Acertos F	% Acertos M	Qtd. Acertos RFM	% Acertos RFM
10	348	14,90%	R	FM	41,64	1	88	1	1	1	132,90	24,99	238,9	100%	100%	99,71%	347	99,71%
19	335	14,34%		RFM	223,89	177	270	1	1	1	116,52	11,9	231,8	100%	100%	100%	335	100%
12	304	13,01%	R	FM	128,98	85	176	1	1	1	116,32	9,1	229,9	80,92%	100%	100%	246	80,92%
18	302	12,93%	RM	F	58,10	1	119	1	1	1	323,33	218,98	470,09	100%	100%	97,02%	293	97,02%
6	249	10,66%		RFM	316,73	269	365	1,01	1	2	121,15	36,4	219,99	100%	98,80%	100%	246	98,80%
5	240	10,27%	M	RF	174,24	116	235	1	1	1	295,47	187,9	462,7	68,75%	100%	77,08%	123	51,25%
11	165	7,06%	M	RF	296,36	238	365	1,01	1	2	300,16	198,5	419	100%	98,18%	84,85%	137	83,03%
0	146	6,25%	RF	M	100,02	3	263	2	2	2	218,42	53,13	483,24	78,77%	100%	58,90%	66	45,21%
17	76	3,25%	RM	F	94,82	3	178	1	1	1	559,38	443,09	783,9	86,84%	100%	100%	66	86,84%
8	54	2,31%	M	RF	268,96	191	364	1	1	1	556,54	435,99	793,99	100%	100%	100%	54	100%
4	36	1,54%	M	RF	170,13	12	354	1	1	1	947,21	753,9	1.119,99	55,56%	100%	100%	20	55,56%
14	24	1,03%	RFM		93,54	11	235	3	3	3	254,66	116	605,79	75,00%	100%	45,83%	8	33,33%
16	18	0,77%	FM	R	190,50	51	325	2,05	2	3	611,57	362	828,8	72,22%	100%	100%	13	72,22%
7	16	0,68%	RFM		62,06	1	219	4,31	4	5	285,98	166,78	438,47	87,50%	100%	50,00%	8	50,00%

13	7	0,30%	M	RF	163,14	50	313	1	1	1	1.433,59	1.260,05	1545,4	42,86%	100%	100%	3	42,86%
3	5	0,21%	RFM		45,40	15	87	6,40	6	7	246,9	96,26	446,06	100%	100%	40,00%	2	40,00%
9	5	0,21%	RM	F	111,60	4	206	1	1	1	1.942,41	1.896,75	2.049,24	60,00%	100%	100%	3	60,00%
15	3	0,13%	RFM		126,66	74	206	8	8	8	376,57	281,63	491,86	66,67%	100%	100%	2	66,67%
1	2	0,09%	RFM		10,50	4	17	10,50	10	11	644,81	436,51	853,12	100%	100%	100%	2	100%
2	1	0,04%	FM	R	235	235	235	2	2	2	2.480	2.480	2.480	100%	100%	100%	1	100%

## 5.4 Análise RFMP (Z-SCORE)

Seguindo o processo anterior, o estudo pretende agora avaliar se a normalização dos dados através do método *Z-Score* trará algum impacto nos resultados gerados pelos modelos RFMP. Por isso é apresentado a seguir os resultados encontrados.

### 5.4.1 *Simplified Silhouette*

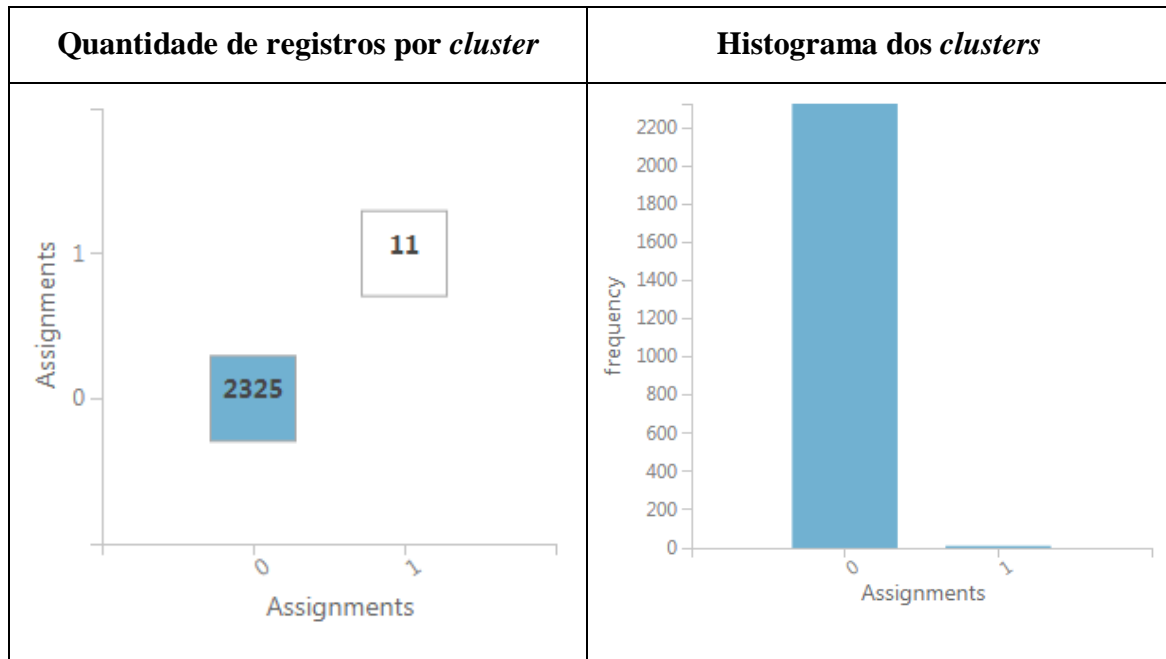
Da mesma forma que na análise anterior realizada por esta métrica, o melhor modelo sugerido obteve um resultado com apenas dois *clusters*, conforme pode ser visto na Tabela 5.27. Porém, tanto a pontuação máxima quanto a ordem subsequente com os demais valores de *K* apresentaram variações.

Tabela 5.27: Pontuação da métrica *Simplified Silhouette* da análise RFMP (Z-Score).

Pontuação	Número de <i>Clusters</i>		Pontuação	Número de <i>Clusters</i>
0,789563	2		0,567561	15
0,723134	3		0,565927	5
0,614385	11		0,56012	6
0,607813	10		0,559826	18
0,594885	12		0,55668	17
0,585107	9		0,55648	7
0,582022	4		0,555619	19
0,57895	8		0,554898	13
0,573136	16		0,546351	20
0,572215	14			

Apesar da quantidade de *clusters* ter ficado igual, foi possível observar pela Tabela 5.28 que a quantidade de registros contidos em cada *cluster* ficou diferente, já que o maior possui 2.325 registros (99,53%) e o menor apenas 11 registros (0,47%). E da mesma forma que na análise anterior, este é um resultado que atende ao critério estabelecido. Porém, como os dados ficaram concentrados em apenas dois *clusters*, a qualidade obtida pelo modelo ficou novamente comprometida.

Tabela 5.28: Gráficos dos resultados da métrica *Simplified Silhouette* da análise RFMP (Z-Score).



Quanto ao gráfico da Figura 5.15, nele é possível observar a representação destes *clusters*.

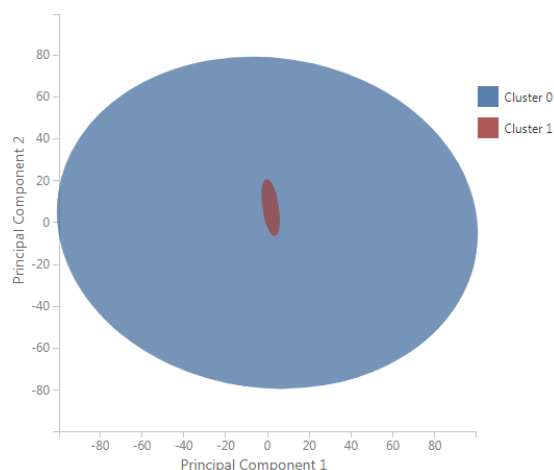


Figura 5.15: Gráfico da visualização dos *clusters* da análise RFMP (Z-Score) com a métrica *Simplified Silhouette*.

Em seguida é descrito o detalhamento de cada *cluster* conforme exibido na Tabela 5.33 (tabela auxiliar).

De início o *cluster* 0, com 2.325 registros (99,35%) e classificação RFM+ e P-, isto é, clientes recentes, frequentes, monetários, mas não lucrativos. As médias de R, F, M e P ficaram em: 157,13; 1,14; 237,85 e 44,30, com taxa de acertos RFMP de apenas

2,19%, impactada pelas taxas de R (52,34%), F (9,51%), M (38,62%) e P (60,13%). Um resultado muito ruim e que não corresponde com a classificação RFMP atribuída.

Já o *cluster* 1 possui 11 registros (0,74%) e tem classificação P+ e RFM-, ou seja, clientes lucrativos, mas não recentes, não frequentes e nem monetários. Justamente o oposto do *cluster* anterior. As médias deste grupo ficaram em: 199,09 (R); 1 (F); 183,41 (M) e 81,77 (P), com taxa de acertos RFMP de 45,45%, impactada pelas taxas de R (63,64%) e M (63,64%). Apesar de apresentar um resultado melhor do que a do *cluster* 0, esta taxa ainda não é suficiente para justificar a classificação RFMP atribuída ao *cluster*.

Quanto à comparação entre os *clusters*, é possível observar que diferente da análise anterior com esta métrica, o modelo atual não levou em consideração o atributo F para separar os agrupamentos. Neste caso, percebe-se que o principal campo de separação foi o atributo P, já que o *cluster* 1 possui os registros com intervalos de P entre 7,28 e 60,40; enquanto que o *cluster* 0 possui os registros com intervalos entre 65 e 100.

Com estes resultados a taxa média de acertos dos *clusters* foi de apenas 23,82% e a taxa de acertos RFMP do modelo foi de 2,40%. Por tudo isso, o estudo classifica como não satisfatória a performance obtida por este modelo, e segue adiante em busca de outro melhor.

#### **5.4.2 *Davies-Bouldin***

Igualmente à análise anterior, esta métrica obteve como melhor valor de *K* sugerido o mesmo valor encontrado pela métrica *Simplified Silhouette* (2 *clusters*). Assim, a pesquisa não seguirá com uma avaliação detalhada deste modelo, já que os resultados foram os mesmos. De qualquer forma, vale ressaltar que a pontuação encontrada pelo índice foi diferente, assim como a ordenação do número de *K*, conforme pode ser visto na Tabela 5.29.

Tabela 5.29: Pontuação da métrica *Davies-Bouldin* da análise RFMP (*Z-Score*).

<b>Pontuação</b>	<b>Número de <i>Clusters</i></b>		<b>Pontuação</b>	<b>Número de <i>Clusters</i></b>
0,425436	2		0,824212	7
0,530382	3		0,83399	13
0,711888	11		0,837512	4
0,734799	10		0,842594	6
0,741337	12		0,842649	17
0,772129	9		0,847512	18
0,781362	8		0,852071	19
0,790463	16		0,899744	20
0,801499	14		0,903929	5
0,813928	15			

### 5.4.3 *Dunn*

Com esta métrica o resultado foi um pouco diferente das outras duas, com um valor de *K* sugerido igual a 3 *clusters*, conforme pode ser visto na Tabela 5.30.

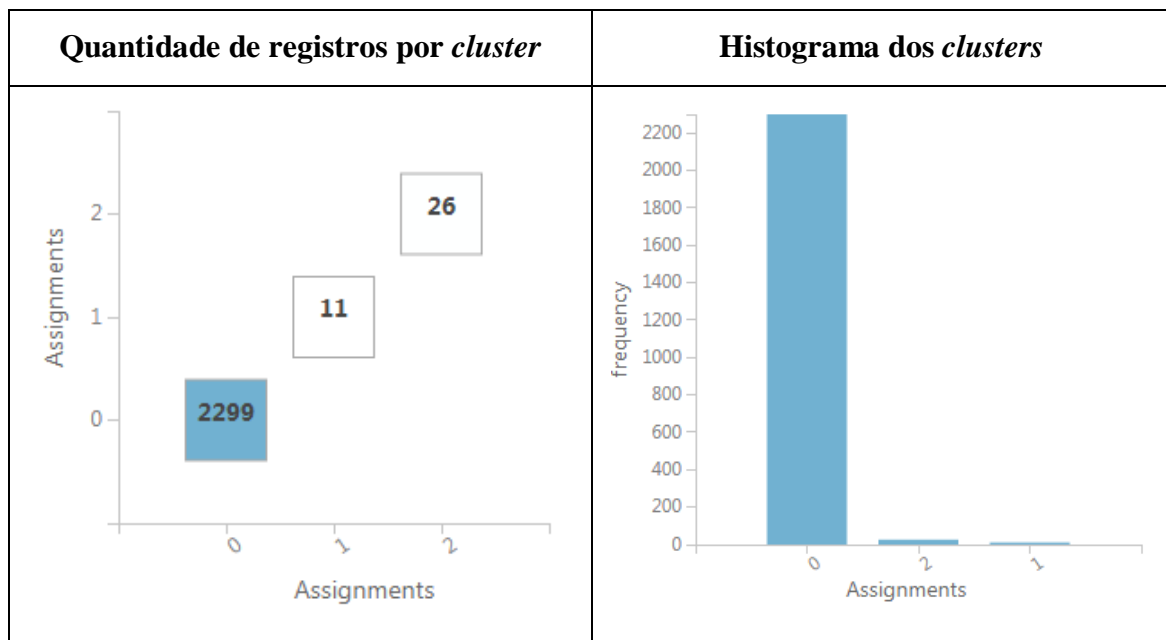
Tabela 5.30: Pontuação da métrica *Dunn* da análise RFMP (*Z-Score*).

<b>Pontuação</b>	<b>Número de <i>Clusters</i></b>		<b>Pontuação</b>	<b>Número de <i>Clusters</i></b>
0,641574	3		0,191193	10
0,581841	2		0,191193	9
0,434533	20		0,191163	8
0,425847	19		0,190561	6
0,40531	18		0,166788	11
0,379681	17		0,147209	12
0,351962	16		0,122418	13
0,283177	4		0,11841	15
0,207209	5		0,11841	14
0,191237	7			

O maior dos *clusters* ficou com 2.229 registros (98,42%) e o menor com apenas 11 registros (0,47%), conforme pode ser visto na Tabela 5.31. A seguir a pesquisa detalha o comportamento de cada *cluster*, como descrito na Tabela 5.34 (tabela auxiliar).

O maior agrupamento ficou com o *cluster* 0, com 2.229 registros (98,42%) e classificação RFMP-, ou seja, clientes não recentes, não frequentes, não monetários e nem lucrativos. As médias encontradas foram de: 158,20 (R); 1,09 (F); 236,96 (M); e 44,31 (P). Já a taxa de acertos RFMP foi de 18,70% – impactada pelas taxas de R (48,06%), 61,59% (M) e 59,98% (P). Um ponto positivo foi que a taxa de F ficou com um bom resultado, apresentando um valor de 91,52%. De todo modo, a classificação RFMP atribuída não ficou consistente com o conteúdo do *cluster*.

Tabela 5.31: Gráficos dos resultados da métrica *Dunn* da análise RFMP (*Z-Score*).



No gráfico da Figura 5.16 é possível observar que a representação dos *clusters* continua com pouca significância para a interpretação do modelo.

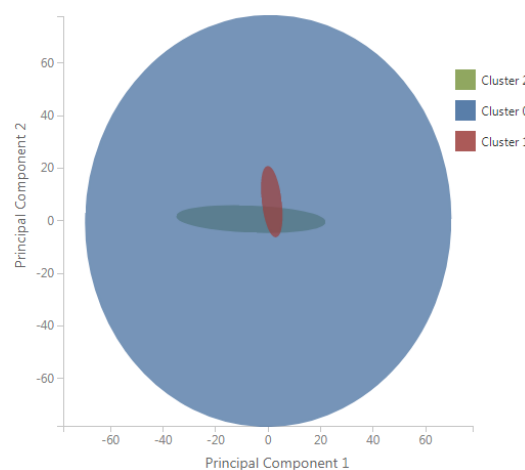


Figura 5.16: Gráfico da visualização dos *clusters* da análise RFMP (*Z-Score*) com a métrica *Simplified Silhouette*.

Já o *cluster* 2 ficou com 26 registros (1,11%) e classificação RFM+ e P-, o que significa um grupo de clientes recentes, frequentes, monetários, mas não lucrativos. As médias obtidas foram de: 62,34 (R); 5,61 (F); 316,52 (M); e 43,77 (P). Já a taxa de acertos RFMP foi de 34,62%, com impacto das taxas de R (88,46%), M (57,69%) e P (73,08%). Um resultado muito fraco e que também não justifica a classificação RFMP associada ao *cluster*.

Por último o *cluster* 1, com 11 registros (0,47%) e classificação P+ e RFM- (clientes lucrativos, mas não recentes, não frequentes e nem monetários). As médias de R, F, M e P encontradas foram de: 199,09; 1; 183,41; e 81,77. A taxa de acertos de RFMP foi de 45,45%, impactada pelas taxas de R (63,64%) e M (63,64%).

Apesar do resultado encontrado não ter sido satisfatório, já foi possível perceber que este modelo levou em consideração o atributo F para a separação do *cluster* 2. Afinal os registros encontrados dentro dele ficaram com valores entre 4 e 11, enquanto que os registros do *cluster* 0 ficaram entre 1 e 3. Em compensação, o *cluster* 1 foi separado dos demais justamente pelo atributo P, uma vez que os registros contidos possuem os valores entre 65 e 100, isto é, com os clientes mais rentáveis.

Todavia, a taxa média de acertos dos *clusters* foi de 32,92% e a taxa de acertos RFMP do modelo foi de apenas 19,01%. Mesmo assim, a pesquisa já conseguiu observar que o modelo produzido por esta métrica – que utilizou os dados normalizados pelo método *Z-Score* – foi capaz de separar os *clusters* tanto pelo atributo F, quanto pelo atributo P.

Na sequência, a avaliação do último modelo.

#### **5.4.4 *Average Deviation***

Mais uma vez esta foi a métrica que produziu o modelo com a maior quantidade de *clusters* – 20 no total – um a menos que na análise RFMP anterior, como pode ser visto na Tabela 5.32. O maior *cluster* possui 455 registros (19,48%), enquanto que outros cinco *clusters* representam os menores, com apenas 2 registros (0,09%), conforme pode ser visto na Figura 5.17 e Figura 5.18.

Tabela 5.32: Pontuação da métrica *Average Deviation* da análise RFMP (Z-Score).

Pontuação	Número de <i>Clusters</i>	Pontuação	Número de <i>Clusters</i>
0,704582	20	1,009983	10
0,717465	19	1,013367	9
0,733708	18	1,02975	8
0,738595	17	1,033227	7
0,760012	16	1,105753	6
0,764147	15	1,361926	5
0,7663	14	1,393461	4
0,823019	12	1,494072	3
0,843762	13	1,552877	2
0,902775	11		

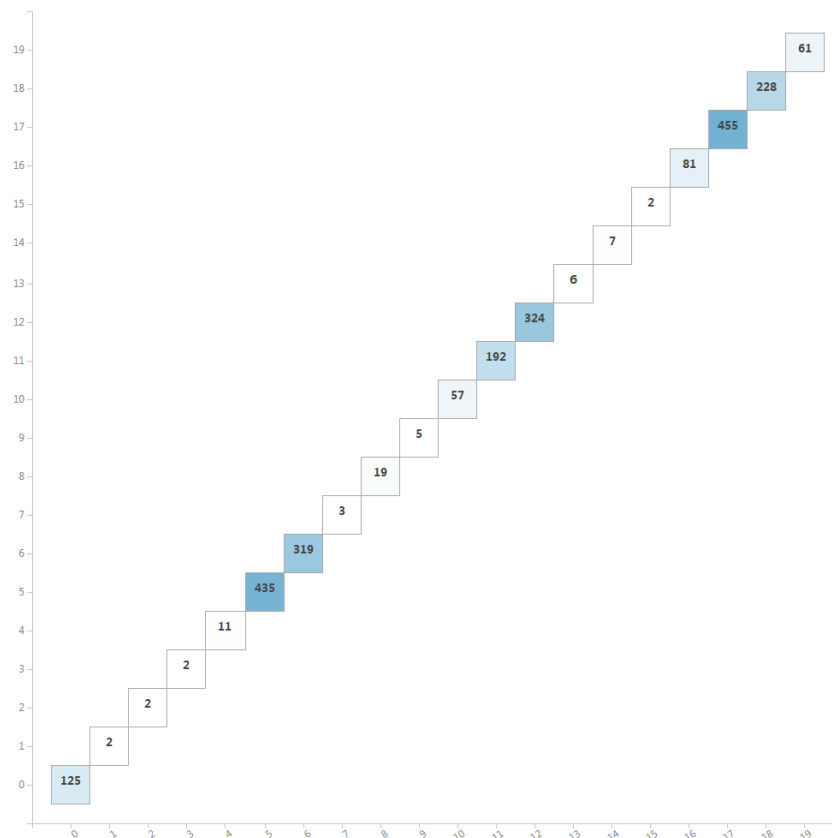


Figura 5.17: Quantidade de registros por *cluster* da análise RFMP (Z-Score) com a métrica *Average Deviation*.

Além disso, um total de 10 *clusters* ficaram com menos de 1% dos dados em seus respectivos conteúdos, o que aumentou o número de grupos que não atendem ao

critério de performance do estudo. Juntos, estes clusters representam um total de 2,53% do volume de dados. De todo modo, foi possível observar que este modelo não produziu o grupo que continha apenas 1 registro, o que significa dizer que ele conseguiu agrupar este registro em algum outro *cluster* (comportamento diferente dos demais modelos analisados até agora com esta métrica).

O gráfico da Figura 5.19 ilustra a representação do modelo gerado. Ainda assim, o problema na visualização por conta da sobreposição persiste, o que não permite um exame mais aprofundado através deste recurso.

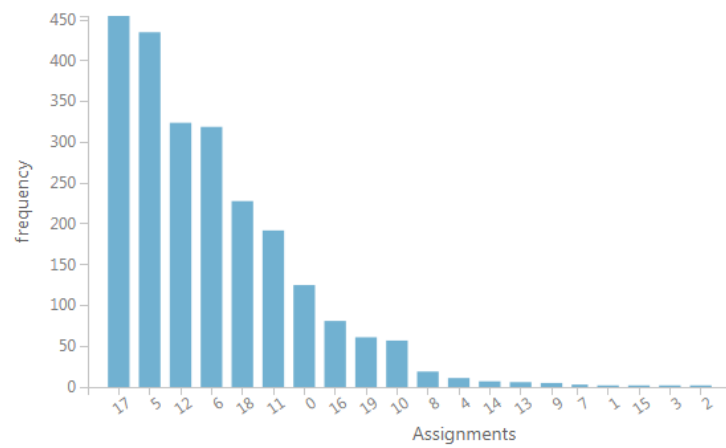


Figura 5.18: Histograma dos *clusters* da análise RFMP (*Z-Score*) com a métrica *Average Deviation*.

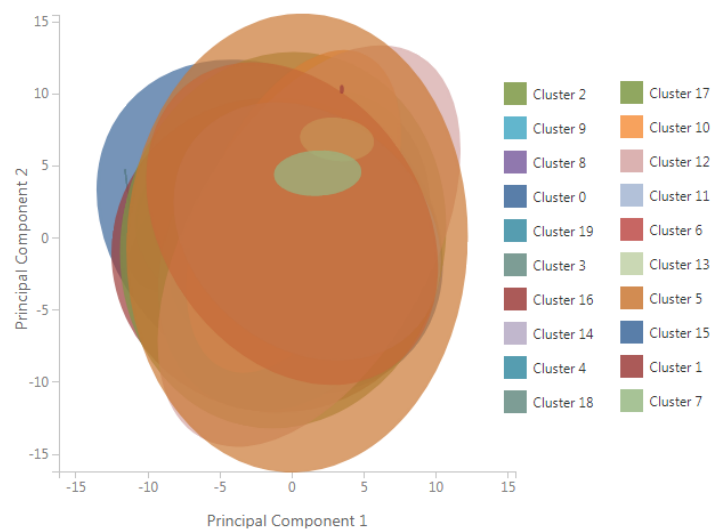


Figura 5.19: Gráfico da visualização dos *clusters* da análise RFMP (*Z-Score*) com a métrica *Average Deviation*.

De qualquer forma, a seguir é descrito o comportamento de cada *cluster* de acordo com os resultados apresentados na Tabela 5.35.

De início o *cluster* 17, com 455 registros (19,48%) e classificação RM+ e FP-, isto é, clientes recentes e monetários, mas não frequentes e nem lucrativos. As médias encontradas foram de: 63,43 (R); 1 (F); 295,69 (M) e 42,79 (P), enquanto que a taxa de acertos RFMP foi de 61,76%, impactada diretamente pela taxa de acertos do atributo M (68,79%).

Na sequência o *cluster* 5 com 435 registros (18,62%) e classificação RFMP- (clientes não recentes, não frequentes, não monetários e nem lucrativos). As médias encontradas foram de: 178,51 (R); 1 (F); 138,43 (M) e 42,18 (P), com taxa de acertos RFMP de 59,08%, impactada pela taxa de R (66,44%).

Em seguida o *cluster* 12, com 324 registros (13,87%) e classificação RFMP- (comportamento semelhante ao encontrado no *cluster* anterior), mas com médias de R, F, M e P de: 312,58; 1; 169,85 e 44,39 respectivamente. Já a taxa de acertos RFMP foi de 53,09%, impactada pelas taxas de M (79,63%) e P (57,10%).

Na continuidade o *cluster* 6, com 319 registros (13,66%) e classificação RP+ e FM-, isto é, clientes recentes e lucrativos, mas não frequentes e nem monetários. As médias encontradas neste grupo foram de: 60,41 (R); 1 (F); 115,48 (M) e 46,61 (P). Porém a taxa de acertos RFMP foi de 79,94%, com um menor desempenho no atributo P (84,64%).

O próximo foi o *cluster* 18, com 228 registros (9,76%) e classificação M+ e RFP-, ou seja, clientes monetários, mas não recentes, não frequentes e nem lucrativos. As médias encontradas foram de: 225,28 (R); 1 (F); 380,23 (M) e 43,58 (P). A taxa de acertos RFMP foi de 73,68%, impactada pela taxa do atributo P com um resultado de 78,95%.

O *cluster* 11 foi o seguinte, com 192 registros (8,22%) e classificação P+ e RFM- (clientes lucrativos, mas não recentes, não frequentes e nem monetários). As médias contidas foram de: 234,84 (R); 1 (F); 115,88 (M) e 47,43 (P), com taxa de acertos RFMP de 89,58% (a maior taxa entre os *clusters* com mais de 1%). O que significa dizer que este *cluster* ficou bastante consistente em relação à classificação RFMP atribuída.

Já o *cluster* 0 obteve 125 registros (5,35%) e classificação RF+ e MP- (clientes recentes e frequentes, mas não monetários e nem lucrativos). As médias ficaram em: 63,83 (R); 2,13 (F); 231,71 (M) e 44,35 (P). A taxa de acertos RFMP foi de apenas 25,60%, impactada pelas taxas de M (57,60%) e P (53,60%). Este baixo resultado se deu pelos valores mínimos e máximos dos atributos M e P terem ficado entre os valores das médias globais de cada atributo, o que ocasionou a imprecisão do resultado obtido. É como se houvesse um desequilíbrio causado pelo valor da média, talvez por conta de um valor muito alto ou muito baixo dentro do *cluster*. Além disso, vale destacar que a média global de M ficou em 237,60 e a de P em 44,49. Neste caso, a diferença entre a média global e a do *cluster* foi de apenas 5,89 para o atributo M e de 0,14 para o atributo P.

E aqui surge um dos pontos que será tratado mais à frente pela pesquisa: será que uma diferença tão pequena, seja por um valor maior ou menor, é o suficiente para classificar como positivo ou negativo o resultado de um *cluster*? Será mesmo que um *cluster* que possui a sua média próxima à da média global tenha a sua classificação imprecisa por conta disso? Sem dúvida nenhuma esta é uma condição que merece um melhor tratamento e deve ser abordada com a importância devida nas análises futuras. Por enquanto, a pesquisa segue com o detalhamento dos *clusters*.

Na sequência o *cluster* 16, com 81 registros (3,47%) e classificação RMP+ e F- (clientes recentes, monetários, lucrativos, mas não frequentes). As médias encontradas foram de: 154,69; 1,08; 772,04; e 45,05, com taxa de acertos RFMP de 35,80%, impactada pelas taxas de R (55,56%) e P (72,84%).

Dando continuidade, foi possível observar o *cluster* 19, com 61 registros (2,61%) e classificação FM+ e RP- (clientes frequentes, monetários, mas não recentes e nem lucrativos), com médias de: 203,70 (R); 2,13 (F); 248,28 (M) e 44,55 (P). Já a taxa de acertos RFMP foi de 16,39% (o menor resultado entre os clusters com mais de 1%), impactada pelas taxas de M (50,82%) e P (49,18%).

Já o *cluster* 10 ficou com 57 registros (2,44%) e classificação RMP+ e F- (comportamento semelhante ao encontrado no *cluster* 16), mas com médias de: 63,63 (R); 1 (F); 249,30 (M) e 52,90 (P). Em compensação a taxa de acertos RFMP foi de 54,39%, com influência da taxa do atributo M (57,89%).

O *cluster* 8 ficou com 19 registros (0,81%) e classificação RFM+ e P- (clientes recentes, frequentes, monetários, mas não lucrativos), com médias de: 58,21 (R); 4,57 (F); 265,27 (M); e 43,73 (P). Já a taxa de acertos RFMP foi de 21,05%, impactada pelas taxas de M (42,11%) e P (68,42%).

Em seguida o *cluster* 4, com 11 registros (0,47%) e classificação RM+ e FP- (comportamento semelhante ao do *cluster* 17), mas com médias de: 136,90 (R); 1 (F); 1.608,90 (M) e 42,26 (P). A taxa de acertos RFMP foi de 45,45%, impactada pelas taxas de R (63,64%) e P (63,64%).

O *cluster* 14 ficou com 7 registros (0,30%), classificação M+ e RFP- (semelhante ao *cluster* 18), e médias de: 192 (R); 1,14 (F); 832,62 (M) e 38,31 (P). Já a taxa de acertos RFMP foi de 57,41%, influenciadas pelas taxas de R (71,43%) e F (85,71%).

Em seguida o *cluster* 13, com 6 registros (0,26%) e classificação P+ e RFM- (equivalente ao *cluster* 11), mas com médias de: 242 (R); 1 (F); 183,30 (M) e 82,75 (P). Já a taxa de acertos RFMP foi de 50%, motivada pelas taxas de R (83,33%) e principalmente de M (66,67%). Como este é um *cluster* pequeno, vale ressaltar que a taxa sofre um forte impacto por qualquer erro produzido.

Já o *cluster* 9 foi o próximo a ser analisado, com 5 registros (0,21%) e classificação RFM+ e P- (comportamento similar ao *cluster* 8), mas com médias de: 98,80 (R); 7,6 (F); 379,95 (M) e 43,71 (P). Já a taxa de acertos RFMP foi de 80%, um bom resultado.

Na sequência o *cluster* 7, com 3 registros (0,13%) e classificação RP+ e FM- (conduta análoga à encontrada pelo *cluster* 6), mas com médias de: 152,66 (R); 1 (F); 139,26 (M); e 67,68 (P), taxa de acertos RFMP de 33,33%, impactada pelas taxas de R (33,33%).

Por último os *clusters* 1, 2, 3 e 15, ambos com somente 2 registros (0,09%) e com as respectivas classificações: RMP+ e F- (similar aos *clusters* 16 e 10); RFM+ e P- (similar aos *clusters* 8 e 9); FM+ e RP- (similar o *cluster* 19); e RFMP- (similar aos *clusters* 5 e 12). As médias de R encontradas foram de: 140; 10,50; 214,50; e 179. Já as médias de F foram de: 1; 10,50; 1,50; e 1; as de M: 249,99; 644,81; e 2.264,62; e as médias de P: 100; 44,33; 36,75; e 15,37. Quanto às taxas de acertos RFMP, os resultados obtidos foram de: 100% para o *cluster* 1; 50% para o *cluster* 2 (impactado pela taxa de

P); 50% para o *cluster* 3 (impactado pela taxa de F) e 0% para o *cluster* 15 (impactado pelas taxas de R e M – 50% para cada).

Com isso, o estudo encerra o detalhamento dos *clusters* gerados por esta métrica e avança com a comparação dos *clusters*. Mas de todo modo, já foi possível perceber que esta análise teve um menor desempenho em relação à análise anterior.

Porém, da mesma maneira que na análise RFM (*Z-Score*), foi possível notar que o atributo F passou a ser considerado como critério de separação entre alguns *clusters*. Isto pode ser observado na relação entre os *clusters* 6, 17, 10, 4, 1, 7, 5, 15, 18, 11, 13 e 12 (que possuem registros somente com o atributo F igual a 1) e os *clusters* 19, 0, 8, 9 e 2 (que possuem registros com o atributo F maior que 1).

Já os *clusters* 16, 14 e 3 ficaram com os clientes com frequência de pedidos entre 1 e 2, mas nestes casos eles se separaram dos demais por conta dos valores contidos no atributo M. Porém, entre eles a separação se deu tanto pelo valor de P (como é o caso dos *clusters* 16 e 14, com médias de 45,05 e 38,31 respectivamente), quanto pelo valor de M (pois os *clusters* 16 e 14 possuem médias de: 772,04 e 832,65, enquanto que o *cluster* 3 possui média de 2.264,62).

A relação entre os *clusters* 5 e 15 ficou muito semelhante no atributo R, com médias de 178,51 e 179, além das médias de M (138,43 e 150,24 respectivamente). A diferença se deu justamente pelo atributo P, já que o primeiro ficou com margem de 42,18 e o segundo com 15,37 (a menor de todas).

Os *clusters* 6 e 11 ficaram muito semelhantes nos valores de F, M e P. Porém eles foram separados pelos valores do atributo R, uma vez que o primeiro ficou com média de 60,41 e o segundo com média de 234,84.

Quanto aos *clusters* 4 e 1, ambos ficaram muito parecidos em relação às médias de R e F, mas cada *cluster* foi separado do outro não só pelo valor de M (1.608,90 para o primeiro e 249,99 para o segundo), quanto pelo valor de P (42,26 para o primeiro e 100 para o segundo).

Já os *clusters* 17 e 10 foram separados somente pelo atributo P. Afinal, os valores de R e M ficaram praticamente idênticos, assim como o valor de M que também ficou

muito próximo. Porém, P ficou com os clientes com margem de 42,79 (*cluster* 17), enquanto que o segundo com margem de 52,90 (*cluster* 10).

Comportamento parecido foi encontrado entre os *clusters* 11 e 13, uma vez que as médias dos atributos R, F e M ficaram com valores semelhantes. Entretanto, quando observadas as médias de P, pode-se notar que o *cluster* 11 ficou com média de 47,43 e o *cluster* 13 de 82,75. Esta mesma separação a partir do atributo P também pode ser constatada entre os *clusters* 5 e 7. A diferença entre eles foi praticamente a margem associada, que no primeiro ficou em 42,18 e no segundo em 67,68.

Quanto aos *clusters* 8, 9 e 2, eles se separaram dos demais pelo valor do atributo F, já que estes são os *clusters* que representam os clientes com os maiores valores de frequência. Em compensação, a separação entre os *clusters* 9 e 18 foi feita não só pelo atributo F (média 7,6 no primeiro e 1 no segundo), como também pelo atributo R (média de 98,80 no primeiro e 225,28 no segundo).

Por último, vale destacar o *cluster* 12, que ficou separado dos demais por conta do valor médio do atributo R (312,58). Sua lucratividade de 44,39 ficou muito semelhante à de outros (como é o caso dos *clusters* 2, 0, 12 e 19), assim como seu valor monetário (169,85), que ficou próximo ao valor do *cluster* 15 (150,24). Deste modo, foi possível afirmar que o atributo P também gerou um impacto na separação dos *clusters*.

Já os *clusters* 0 e 19 ficaram separados pelo atributo R, já que ambos possuem clientes com frequência entre 2 e 3 pedidos, com médias de F idênticas de 2,13, valor médio de M de 231,71 e 248,28, e P de 44,35 e 44,45 respectivamente. A diferença foi justamente nos valores encontrados pelo atributo R, que foi de 63,83 para o *cluster* 0 e 203,70 para o *cluster* 19.

Por fim, foi analisada a taxa média de acertos dos *clusters* que ficou com apenas 51,81%, enquanto que a taxa de acertos RFMP do modelo ficou com 61,30%. Este resultado foi superior ao encontrado pela primeira análise RFMP, porém ficou inferior ao que foi encontrado pela análise RFM (*Z-Score*), o que deixou as classificações RFMP não tão consistentes.

De qualquer maneira, a pesquisa finaliza a análise RMFP com os dados padronizados pelo método *Z-Score* e segue para uma breve conclusão do que foi obtido até aqui.

### 5.4.5 Síntese dos Resultados

Da mesma maneira que na análise RFM (*Z-Score*), também foi possível notar nesta análise que o atributo F passou a ser considerado como critério de separação entre alguns *clusters*. Afinal, os resultados das taxas de acertos deste atributo foram positivos, e o modelo passou a errar menos a classificação, minimizando assim o problema da assimetria do atributo F.

Além disso, foi possível perceber que o atributo P também impactou na formação e separação de alguns *clusters*, o que foi positivo. Afinal, saber a margem de um determinado grupo de clientes pode ser fundamental para a criação de estratégias de marketing e de precificação.

Como exemplo, a relação entre os *clusters* 19 e 10 obtida pela métrica *Average Deviation*. Em ambos, o valor médio de M ficou muito parecido (248,28 e 249,30), porém houve uma grande diferença na lucratividade (atributo P) que foi de 44,45 para o primeiro, e de 52,90 para o segundo. Ou seja, se fosse levado em consideração apenas o atributo M, seria bem provável que os clientes de ambos os *clusters* fossem tratados da mesma forma. Porém, ao analisar o atributo P, foi possível compreender que existe uma diferenciação entre eles em um item que é fundamental para as empresas, a lucratividade.

De todo modo, as taxas obtidas pela análise RFMP (*Z-Score*) ficaram abaixo daquelas encontradas pela análise RFM (*Z-Score*), o que fez com que as classificações RFMP atribuídas à cada *cluster* não ficassem tão consistentes. Contudo, isto se deve em grande parte quando as faixas de valores de um determinado *cluster* estão entre o valor médio global dos dados, o que traz uma incerteza para a classificação do *cluster*.

Este é um problema que merece atenção. Afinal, se o valor médio de um *cluster* está próximo à média global do conjunto de dados, a tendência é que os valores dos registros variem também próximos à média – tanto para cima, quanto para baixo. E foi justamente esta variação que trouxe incerteza para a classificação do modelo estudado.

Um exemplo: o *cluster* 19 da última análise ficou com a média do atributo P igual a 44,45. Já a média global deste atributo é de 44,49. Desta forma, este *cluster* foi classificado como P- (não lucrativo). Porém, há dentro deste *cluster* registros de clientes com valores de P variando entre 40,06 (valor abaixo da média) e 49,04 (valor acima da média). Entretanto, como a classificação foi associada à média do *cluster*, todos os

clientes deste grupo receberam o mesmo rótulo no atributo P. Contudo, quando analisados de forma isolada, cada registro de cliente tem o seu valor comparado à média global, e consequentemente suas respectivas classificações podem variar em relação à classificação do *cluster* (tanto positivo, quanto negativo). Portanto, foi justamente esta variação que gerou o erro de classificação.

Outro exemplo: ao avaliar o *cluster* 17, do mesmo modelo anterior, foi possível perceber que a média do atributo M ficou em 63,43, enquanto que a média global ficou em 157,33. Neste caso, todo o segmento foi classificado como R+. Entretanto, como os valores deste atributo variaram entre 1 e 145 (valores abaixo da média global), ficou fácil para o modelo acertar todos os registros contidos neste *cluster*.

Portanto, se faz necessário avaliar uma melhor maneira para classificar os registros. A ideia é evitar que a classificação seja feita apenas por um valor absoluto.

Outro ponto de destaque foi que as taxas médias de acertos globais dos *clusters* e as taxas de acertos RFMP dos modelos apresentaram resultados melhores do que aquelas obtidas sem o método *Z-Score*. Em especial a que foi encontrada pela métrica *Average Deviation*, que com 20 *clusters* obteve uma taxa média de acertos dos *clusters* de 51,81% (valor superior ao encontrado na análise RFMP – 39,12%) e taxa de acertos RFMP do modelo de 61,30% (valor superior ao encontrado na análise RFMP – 31,55%). Apesar disso, este modelo não obteve o critério mínimo de performance adotado pelo estudo.

De todo modo, esta análise possibilitou identificar que mesmo com um atributo assimétrico é possível utilizá-lo para segregação dos *clusters*, e isto se deve à utilização do método *Z-Score*. Além disso, foi possível observar também que todos os atributos, de uma forma ou de outra, foram utilizados na segmentação dos grupos. Assim, o estudo encerra esta etapa e avança para a busca de uma solução que possa atender aos critérios de qualidade e de rendimento esperados pela pesquisa.

### 5.4.6 Tabelas Auxiliares – RFMP (Z-Score)

Tabela 5.33: Tabela auxiliar com os resultados das métricas *Simplified Silhouette e Davies Bouldin* da análise RFMP (Z-Score).

G	Qtd.	%	RFMP +	RFMP -	Med R	Min R	Max R	Med F	Min F	Max F	Med M	Min M	Max M	Med P	Min P	Max P	% de Acertos				Qtd. Acertos RFMP	% Acertos RFMP
																	% R	% F	% M	% P		
0	2.325	99,53	RFM	P	157,13	1	365	1,14	1	11	237,85	9,10	2.480,00	44,30	7,28	60,4	52,34	9,51	38,62	60,13	51	2,19
1	11	0,47	P	RFM	199,09	2	278	1	1	1	183,41	30	329,94	81,77	65	100	63,64	100	63,64	100	5	45,45

Tabela 5.34: Tabela auxiliar com os resultados da métrica *Dunn* da análise RFMP (Z-Score).

G	Qtd.	%	RFMP +	RFMP -	Med R	Min R	Max R	Med F	Min F	Max F	Med M	Min M	Max M	Med P	Min P	Max P	% de Acertos				Qtd. Acertos RFMP	% Acertos RFMP
																	% R	% F	% M	% P		
0	2.299	98,42		RFMP	158,20	1	365	1,09	1	3	236,96	9,10	2480	44,31	7,28	60,4	48,06	91,52	61,59	59,98	430	18,70
2	26	1,11	RFM	P	62,34	1	219	5,61	4	11	316,52	96,26	853,12	43,77	40,22	47,45	88,46	100	57,69	73,08	9	34,62
1	11	0,47	P	RFM	199,09	2	278	1	1	1	183,41	30	329,94	81,77	65	100	63,64	100	63,64	100	5	45,45

Tabela 5.35: Tabela auxiliar com os resultados da métrica *Average Deviation* da análise RFMP (Z-Score).

G	Qtd.	%	RFMP +	RFMP -	Med R	Min R	Max R	Med F	Min F	Max F	Med M	Min M	Max M	Med P	Min P	Max P	% de Acertos				Qtd. Acertos RFMP	% Acertos RFMP
																	% R	% F	% M	% P		
17	455	19,48	RM	FP	63,43	1	145	1	1	1	295,69	39	554,98	42,79	35,23	47,36	100	100	68,79	91,21	281	61,76
5	435	18,62		RFMP	178,51	75	292	1	1	1	138,43	9,10	329,89	42,18	35,23	45,44	66,44	100	92,41	94,71	257	59,08
12	324	13,87		RFMP	312,58	241	365	1	1	1	169,85	36,4	405,70	44,39	35,23	48,85	100	100	79,63	57,10	172	53,09
6	319	13,66	RP	FM	60,41	1	146	1	1	1	115,48	24,99	340,60	46,61	43,59	50	100	100	95,30	84,64	255	79,94
18	228	9,76	M	RFP	225,28	113	364	1	1	1	380,23	224,99	621,60	43,58	37,88	48,85	91,67	100	99,56	78,95	168	73,68
11	192	8,22	P	RFM	234,84	144	359	1	1	1	115,88	24,99	356,88	47,43	44,46	52,24	94,79	100	94,79	99,48	172	89,58
0	125	5,35	RF	MP	63,83	3	137	2,13	2	3	231,71	53,13	613,24	44,35	36,84	50,72	100	100	57,60	53,60	32	25,60
16	81	3,47	RMP	F	154,69	3	354	1,08	1	2	772,04	526,3	1.119,99	45,05	41,46	48,85	55,56	91,36	100	72,84	29	35,80
19	61	2,61	FM	RP	203,70	136	359	2,13	2	3	248,28	54	580	44,45	40,06	49,04	80,33	100	50,82	49,18	10	16,39
10	57	2,44	RMP	F	63,63	1	179	1	1	1	249,30	43,6	553,19	52,90	48,85	60,4	92,98	100	57,89	100	31	54,39
8	19	0,81	RFM	P	58,21	1	219	4,57	4	6	265,27	96,26	438,47	43,73	40,22	47,45	89,47	100	42,11	68,42	4	21,05
4	11	0,47	RM	FP	136,90	4	313	1	1	1	1.608,90	1.260,05	1939,87	42,26	37,87	47,46	63,64	100	100	63,64	5	45,45
14	7	0,30	M	RFP	192	80	349	1,14	1	2	832,62	684,8	1072,8	38,31	34,15	41,45	71,43	85,71	100	100	4	57,14
13	6	0,26	P	RFM	242	65	278	1	1	1	183,30	54,99	329,94	82,75	80	83,31	83,33	100	66,67	100	3	50,00
9	5	0,21	RFM	P	98,80	27	206	7,60	7	8	379,95	281,63	491,86	43,71	42,78	44,46	80,00	100	100	100	4	80,00
7	3	0,13	RP	FM	152,66	2	277	1	1	1	139,26	30	220	67,68	65	69,93	33,33	100	100	100	1	33,33
2	2	0,09	RFM	P	10,50	4	17	10,50	10	11	644,81	436,51	853,12	44,33	43,36	45,31	100	100	100	50,00	1	50,00
15	2	0,09		RFMP	179	49	309	1	1	1	150,24	37,5	2.62,99	15,37	7,28	23,47	50,00	100	50,00	100	0	0,00
3	2	0,09	FM	RP	214,50	194	235	1,50	1	2	2.264,62	2.049,24	2480	36,77	34,68	38,87	100	50,00	100	100	1	50,00
1	2	0,09	RMP	F	140	136	144	1	1	1	249,99	249,99	249,99	100	100	100	100	100	100	100	2	100

## 5.5 Resumo

### 5.5.1 Da Análise

Esta análise teve como propósito entender o processo de geração do modelo RFM em contrapartida com o modelo RFMP proposto. O objetivo era avaliar se a inclusão do parâmetro P traria alguma influência no processo de segmentação dos clientes e na formação e separação dos seus respectivos *clusters*.

Neste intuito, o estudo produziu diversas análises para melhor compreender o processo de geração dos modelos e distinguir o comportamento de separação entre os *clusters* a partir da técnica de segmentação de dados *k-means*. Tudo isso através de uma plataforma de *Machine Learning* na nuvem.

Entretanto, como um dos grandes desafios para utilização do algoritmo *k-means* é justamente encontrar o melhor número de  $K$ , o estudo optou por definir um indicador que pudesse auxiliar na busca por este valor. Assim, a pesquisa estabeleceu como parâmetro o seguinte princípio: buscar dentro do conjunto de dados a maior quantidade possível de segmentos, mas desde que o número de registros contidos em um determinado agrupamento não fosse menor do que 1% dos registros totais.

Além disso, como não havia forma de mensurar a qualidade dos modelos e sua respectiva assertividade em relação ao processo RFM/P, o estudo propôs a criação de três novos índices de mensuração para avaliar os resultados obtidos, sendo: o primeiro associado à qualidade individual de cada *clusters* produzido – em correspondência com a classificação RFM/P atribuída; o segundo, para mensurar a qualidade média dos *clusters* gerados pelos modelos; e por último, e mais importante, para determinar a qualidade e a assertividade geral do modelo.

Estes índices foram desenvolvidos para suprir a ausência de indicadores existentes no processo de criação de modelos RFM/P. Afinal, sem isto o estudo não teria como comparar os modelos entre si, e muito menos mensurar a consistência e a qualidade dos *clusters* e modelos produzidos. Vale destacar que estes índices foram elaborados a partir da criação da tabela auxiliar gerada com base nos resultados obtidos por cada modelo.

Com estas premissas o estudo deu início a uma pesquisa empírica para avaliar os diversos resultados produzidos pela criação de vários modelos, como: RFM, RFMP, RFM (*Z-Score*) e RFMP (*Z-Score*).

Esta busca foi realizada com o uso do recurso *Sweep Clustering*, disponibilizado pela plataforma utilizada, e que tem por objetivo inferir o melhor número de  $K$  através de diferentes configurações de parâmetros. Entre eles, a de uma métrica de precisão (método matemático que estima o ajuste do modelo). Contudo, apesar da construção e testes de vários modelos, o estudo analisou somente o melhor resultado sugerido por cada métrica em cada uma das análises geradas (vale ressaltar que para cada análise, quatro métricas distintas foram utilizadas).

Em seguida, o estudo lançou mão de quatro análises para estabelecer os entendimentos necessários a respeito do comportamento da pesquisa. A primeira delas foi feita sem qualquer tratamento das variáveis R, F e M, ou seja, com os registros originais do conjunto de dados adotado. Já a segunda, seguiu o mesmo procedimento, mas com a adição do parâmetro P. A terceira análise, foi realizada com os mesmos atributos utilizados pela primeira, mas fazendo uso da normalização dos dados através do método *Z-Score*. E por último a quarta análise, que também fez uso da normalização dos dados e incluiu o atributo P na avaliação.

Porém, em todas elas o objetivo foi examinar – entre as quatro métricas adotadas através do recurso *Sweep Clustering* – qual delas criaria o melhor valor de  $K$  para a criação dos modelos RFM/P, se este valor estava dentro do indicador auxiliar definido anteriormente, e se os resultados obtidos pelas métricas de consistência e qualidade dos modelos estavam coerentes com as classificações RFM/P atribuídas. Além disso, claro, havia o propósito de se analisar o comportamento dos *clusters* – observando não só os seus respectivos conteúdos – como também os critérios adotados para a separação entre eles por cada modelo.

Desta forma, a pesquisa seguiu com o seu desenvolvimento e comparou os resultados de cada análise para averiguar o desempenho obtido entre eles através dos índices propostos e se atenderiam ao parâmetro de performance adotado pelo estudo.

Além disso, para cada análise produzida foram apresentados alguns recursos gerados pela própria plataforma utilizada. Com eles foi possível interpretar parte dos

resultados de cada modelo. Contudo eles não foram suficientes para o exame completo do mérito desta pesquisa. Por isso, o estudo complementou os resultados com a criação da tabela auxiliar que norteou o entendimento final de cada modelo. Deste modo, é apresentado a seguir um resumo do que foi alcançado por estas análises.

## 5.5.2 Dos Resultados

De um modo geral, é apresentada na Tabela 5.36 uma compilação dos resultados obtidos até aqui.

Tabela 5.36: Resumo dos resultados da primeira avaliação dos modelos.

Análise	Métrica	Qtd. Clusters	Menor Cluster	Maior Cluster	Taxa Média Acertos Clusters	Taxa Acertos RFM / P
RFM	<i>Simplified Silhouette</i>	3	(0,56%) 13 registros	(87,37%) 2.041 registros	32,43%	31,59%
	<i>Davies-Bouldin</i>	10	(0,04%) 1 registro	(29,84%) 697 registros	50,27%	46,53%
	<i>Dunn</i>	10	(0,04%) 1 registro	(29,84%) 697 registros	50,27%	46,53%
	<i>Average Deviation</i>	20	(0,04%) 1 registro	(13,14%) 307 registros	58,32%	55,14%
RFMP	<i>Simplified Silhouette</i>	3	(0,56%) 13 registros	(87,37%) 2.041 registros	21,26%	19,14%
	<i>Davies-Bouldin</i>	9	(0,04%) 1 registro	(31,04%) 725 registros	28,84%	22,47%
	<i>Dunn</i>	8	(0,04%) 1 registro	(32,75%) 765 registros	29,20%	22,00%
	<i>Average Deviation</i>	19	(0,04%) 1 registro	(13,10%) 306 registros	39,12%	31,55%
RFM (Z-Score)	<i>Simplified Silhouette</i>	2	(1,11%) 26 registros	(98,89%) 2.310 registros	42,20%	30,82%
	<i>Davies-Bouldin</i>	2	(1,11%) 26 registros	(98,89%) 2.310 registros	42,20%	30,82%
	<i>Dunn</i>	2	(1,11%) 26 registros	(98,89%) 2.310 registros	42,20%	30,82%
	<i>Average Deviation</i>	20	(0,04%) 1 registro	(14,90%) 348 registros	73,17%	84,55%
RFMP (Z-Score)	<i>Simplified Silhouette</i>	2	(0,47%) 11 registros	(99,53%) 2325 registros	23,82%	2,40%
	<i>Davies-Bouldin</i>	2	(0,47%) 11 registros	(99,53%) 2325 registros	23,82%	2,40%
	<i>Dunn</i>	3	(0,47%) 11 registros	(98,42%) 2.299 registros	39,92%	19,10%
	<i>Average Deviation</i>	20	(0,09%) 2 registros	(19,48%) 455 registros	51,81%	61,30%

A primeira análise permitiu não só avaliar o modelo RFM gerado, mas também a utilização do ambiente *Azure ML*. Assim, foi possível compreender as funcionalidades existentes e os recursos disponíveis para a geração e validação dos modelos, que de um modo geral apresentaram um bom desempenho. Entretanto, vale ressaltar que o gráfico de visualização dos *clusters* ficou com a sua funcionalidade comprometida, já que a saída produzida ficou confusa e incompleta, tornando-o – em muitos casos – incompreensível. De todo modo, o estudo utilizou uma tabela auxiliar para ajudar a complementar o processo de avaliação dos modelos.

Sendo assim, foi possível observar que todos os modelos gerados por esta análise tiveram os seus *clusters* divididos pelos atributos M e R, com o atributo M se sobrepondo ao de R. Além disso, também foi observado que quanto maior o número de *clusters*, melhor era a separação entre eles, assim como a distribuição dos registros pelas classes.

Contudo, o melhor modelo foi encontrado pela métrica *Average Deviation*, que gerou um total de 20 *clusters*, com uma taxa média de acertos dos *clusters* de 58,32% e taxa de acertos RFM do modelo de 55,14%. Porém, o menor agrupamento ficou com apenas 0,04% dos registros, não atendendo ao critério de performance estabelecido. Além disso, outros quatro *clusters* ficaram com uma quantidade de registros abaixo de 1%.

Já os outros modelos desta análise produziram os seguintes resultados: métrica *Simplified Silhouette*, 3 *clusters*, taxa de média de acertos dos *clusters* de apenas 32,43% e de acertos RFM do modelo de 31,59%; e métricas *Davies-Bouldin* e *Dunn*, com 10 *clusters*, taxa média de acertos dos *clusters* de 50,27% e de acertos RFM do modelo de 46,53%.

Todos estes resultados foram impactados pela assimetria do atributo F, o que fez com que o desempenho dos modelos, mensurados pelas taxas de acertos, ficassem com uma baixa performance.

Em relação à segunda análise, o objetivo foi avaliar se a inclusão do parâmetro P traria alguma influência no processo de segmentação dos clientes. O intuito era confrontar se haveria algum impacto entre os modelos RFM e os RFMP. Em especial, se haveria alguma diferenciação dos clientes não só pelo valor monetário – procedimento padrão dos modelos RFM – como também pela rentabilidade.

Entretanto, de um modo geral foi possível observar que a qualidade dos modelos ficou inferior à da análise RFM, já que a inclusão do atributo P trouxe uma complexidade a mais para a assertividade dos *clusters*. Todavia, também contribuíram para isto a concentração dos registros em um único valor para o atributo F e a baixa distribuição dos dados entre os *clusters*.

Novamente o melhor modelo encontrado foi produzido pela métrica *Average Deviation*, com 19 *clusters*, taxa média de acertos dos *clusters* de 39,12% e taxa de acertos RFMP do modelo de 31,55%. Além disso, não foi possível encontrar um resultado que atendesse ao critério da pesquisa, pois um total de três *clusters* ficaram com menos de 1% dos dados.

Já a métrica *Simplified Silhouette*, com 3 *clusters*, obteve um resultado médio de acertos dos *clusters* de 21,26% e de acertos RFMP do modelo de 19,14%. Em compensação, os resultados da métrica *Davies-Bouldin*, com 9 *clusters*, foram de 28,84% para a taxa média de acertos dos *clusters* e de 22,47% para taxa de acertos RFMP do modelo. Enquanto que as taxas produzidas pela métrica *Dunn* (com 8 *clusters*) foi de 29,20% para a taxa média de acertos dos *clusters* e de 22% para a taxa de acertos RFMP do modelo.

Quanto à terceira análise – RFM (*Z-Score*) – ela teve como objetivo avaliar se a criação de novos modelos, a partir da normalização dos dados através do método *Z-Score*, conseguiriam melhorar os resultados encontrados.

Neste caso, as três primeiras métricas (*Simplified Silhouette*, *Davies-Bouldin* e *Dunn*) indicaram como melhor valor de *K* um mesmo resultado – apenas 2 *clusters*, com taxa média de acertos dos *clusters* de 42,20% e taxa média de acertos RFM do modelo de 30,82%. Porém, os resultados encontrados pela métrica *Average Deviation*, separado por 20 *clusters*, apresentaram um desempenho surpreendente; com resultados de 73,17% na taxa média dos *clusters* e de 84,55% na taxa de acertos RFM, uma performance excelente se comparada às que foram apresentadas anteriormente.

Isto porque a normalização dos dados conseguiu reduzir as distorções apresentadas pelo atributo F, passando inclusive a considerar este atributo no processo de separação entre alguns *clusters*. Com isso, as qualidades dos *clusters* melhoraram e,

consequentemente, o modelo passou a errar menos, visto que a assimetria deste atributo impactava diretamente nos resultados obtidos.

Neste caso, foi possível afirmar que a normalização *Z-Score* conseguiu trazer uma qualidade maior ao modelo desenvolvido, já que tornou mais consistente à classificação RFM atribuída a cada um dos *clusters*. Entretanto, o modelo produzido pela métrica *Average Deviation* ainda não atendeu ao índice estabelecido, já que 8 agrupamentos ficaram com menos de 1%, e que juntos representaram 2,44% dos registros.

Por último, a análise RFMP (*Z-Score*), na qual foi possível notar que o atributo F também passou a ser considerado como critério de separação entre alguns *clusters*. Além disso, foi observado que o atributo P gerou um impacto na formação e separação dos agrupamentos.

Entretanto, os resultados desta análise ficaram abaixo das encontradas pela análise RFM (*Z-Score*). Contudo, isto foi causado, na maioria dos casos, quando as faixas de valores de um determinado *cluster* ficaram entre o valor médio global dos dados, o que trouxe uma incerteza para a classificação dos modelos.

De todo modo, a métrica *Average Deviation*, com 20 *clusters*, obteve uma taxa média de acertos dos *clusters* de 51,81% e taxa de acertos RFMP do modelo de 61,30%, resultados superiores aos encontrados na primeira análise RFMP. Quanto aos demais modelos, eles obtiveram os seguintes resultados: métricas *Simplified Silhouette* e *Dunn*, 2 *clusters*, com taxa média de acertos dos *clusters* de apenas 23,82% e de acertos RFM do modelo de 2,40%. Já a métrica *Davies-Bouldin*, com 3 *clusters*, obteve taxa média de acertos dos *clusters* de 39,92% e de acertos RFM do modelo de 19,10%.

Desta forma, a pesquisa finaliza por completo a primeira análise do estudo e segue adiante para tratar dos desafios encontrados e apresentar as respectivas propostas de soluções.

# 6 Desafios Encontrados

## 6.1 Problemas Identificados

Dado todo conteúdo produzido e analisado até o momento, o estudo identificou como principais desafios para o avanço desta pesquisa (com o conjunto de dados analisado) os seguintes pontos:

- a) Concentração dos registros em poucos *clusters* – todo modelo que concentrou muitos registros de clientes em uma pequena quantidade de *clusters* ficou com a sua classificação RFM/P inconsistente, ou seja, um modelo que produz poucos *clusters* tem a sua qualidade impactada negativamente;
- b) Distorção de atributo – no caso em questão, a distorção (assimetria) do atributo F interferiu no processo de separação e formação dos *clusters*;
- c) Maior complexidade com o atributo P – a inclusão do atributo P trouxe uma dificuldade a mais para a assertividade dos *clusters* e para qualidade dos modelos. Afinal, nenhum modelo RFMP teve a taxa de acertos superior às apresentadas pelas análises RFM;
- d) Comparação por um valor absoluto – a comparação das médias dos *clusters* pela média global do conjunto de dados, como forma de atribuição da classificação RFM/P, não trouxe um resultado tão adequado, produzindo em muitos casos uma distorção na classificação dos *clusters* quando comparadas às classificações RFM/P dos registros.

Todos os casos citados possuem como evidências os valores das métricas de avaliação propostas pela pesquisa, que analisaram não só os resultados obtidos por cada modelo, como também à qualidade e a assertividade individual de cada *cluster*. Portanto, tendo em mente cada um dos desafios apresentados e suas respectivas causas (analisadas e descritas ao longo desta avaliação), se faz necessário alcançar um método que permita reduzir estes impactos e aumentar a consistência da classificação RFM/P associada a cada *cluster* produzido pelos modelos gerados.

## 6.2 Soluções Recomendadas

As propostas a seguir têm por finalidade reduzir os efeitos causados por cada um dos itens citados anteriormente, além de aumentar a qualidade dos modelos gerados e melhorar a assertividade dos *clusters* no que diz respeito às classificações dos modelos RFM/P.

Quanto à solução para o primeiro problema identificado no item a do tópico anterior, o que se propõe é a criação de vários modelos, variando o número de  $K$ , para identificar aquele que irá conter o melhor valor para a métrica **Taxa de Acertos RFM ou RFMP do Modelo**. Afinal, como dito no início do estudo, o recurso *Sweep Clustering*, iria apenas auxiliar a busca por um melhor valor de  $K$ , e não definir de forma incondicional aquele que deveria ser utilizado pelo estudo. Assim, na próxima etapa, serão criados diversos modelos – com o valor de  $K$  variando entre 2 e 20 – para estabelecer qual será aquele que irá conter o melhor valor para a métrica citada acima.

Em relação à assimetria do atributo F, foi observado que a própria normalização dos dados – através do método *Z-Score* – conseguiu reduzir os impactos causados. Neste caso, não será preciso avaliar uma nova proposta para solução deste problema, mas apenas manter a utilização do método de normalização nas análises futuras.

Sobre a maior complexidade trazida pelo atributo P, este é um problema que deve ser enfrentado pelo algoritmo. Entretanto, o estudo recomenda – para ajudar no processo de segmentação dos clientes – agrupar os dados de entrada em intervalos de classes, transformando os dados brutos, provenientes das variáveis contínuas da análise, em faixas de valores que representem a estrutura do negócio modelado.

Por último, e mais importante, não basta somente introduzir meios que mitiguem os impactos dos problemas levantados anteriormente, mas sim a proposição de um novo método para a classificação dos modelos RFM/P em substituição à forma atual utilizada. E é justamente isto que o estudo sugere para solucionar o problema apresentado no item d do tópico anterior.

Esta nova proposta tem por objetivo aumentar o grau de confiança das classificações RFM/P, melhorando a consistência entre os *clusters* e seus respectivos conteúdos. Sendo assim, o estudo propõe que o melhor seria não segmentar os clientes e

os *clusters* apenas como positivos ou negativos a partir de um determinado valor absoluto (no caso o valor da média global dos atributos), mas sim classificá-los com base na criação de três faixas de valores, contendo:

- i. Clientes Neutros (°) – ou seja, aqueles que estão dentro de um intervalo de valor que contemple o comportamento padrão de um cliente para qualquer um dos atributos analisados (R, F, M ou P);
- ii. Clientes Negativos (-) – isto é, o cliente que possui um padrão de compra inferior ao observado na faixa de valores do cliente neutro;
- iii. Clientes Positivos (+) – aquele que supera o padrão de compra observado na faixa de valores que contém o cliente neutro.

A faixa com os Clientes Neutros deve estar contida, preferencialmente, entre os valores de algumas observações estatísticas, como é o caso da média, da moda e da mediana, mas isto não deve ser uma regra. Por isso, mais importante do que utilizar somente os dados provenientes das análises estatísticas deve ser a avaliação por um especialista do negócio para estabelecer ou validar esta faixa. Isto permitirá uma flexibilização no momento da definição dos valores que estabelecerão os limites mínimo e máximo que representarão este segmento de cliente. Quanto à faixa que irá representar os Clientes Negativos, ela deverá estar abaixo do valor mínimo estabelecido pela faixa de Clientes Neutros. Já a faixa com os Clientes Positivos, ela deverá estar acima do valor máximo estabelecido pela faixa de Clientes Neutros.

Deste modo, com este novo formato de classificação do modelo RFM/P, evita-se o problema apresentado no item d do tópico anterior, quando um único valor era utilizado para comparar se um cliente ou *cluster* era positivo ou negativo, com base apenas no valor obtido pela média global de cada atributo do conjunto de dados.

Sendo assim, o estudo finaliza suas propostas para os problemas identificados.

## 6.3 Reorganização dos Dados

Para atender a uma das sugestões apresentadas pela tese, o conjunto de dados utilizado pelo estudo passou por um processo de tratamento para agrupar os dados de entrada em intervalos de classes. Com isso, os valores das variáveis contínuas da análise foram definidos em faixas para melhor representar a estrutura do negócio analisado.

Assim, o estudo realizou algumas alterações e criou as seguintes classes para os atributos R, F, M e P:

- **Atributo R** – a primeira modificação realizada foi alterar o tempo do atributo R, modificando-o de dia para trimestre (uma vez que os produtos vendidos pelo portal analisado não possuem um ciclo de compra diário e nem mensal). Deste modo, o atributo R passou a ser definido pela diferença em trimestres do registro da data de compra mais recente no banco de dados para cada cliente. Conseqüentemente, 4 classes foram criadas para cada trimestre do intervalo de tempo analisado (que é de 1 ano) conforme pode ser observado na tabela a seguir:

Tabela 6.1: Divisão das classes do atributo R em trimestres e seus comportamentos.

<b>Faixas (Trimestre)</b>	<b>Comportamento</b>
1	Cliente Positivo (Recente)
2	Cliente Neutro
3	Cliente Negativo (Não Recente)
4	Cliente Negativo (Não Recente)

A primeira classe ficou com os clientes que fizeram sua última compra dentro do intervalo de tempo do 1º trimestre, e por isso os registros com este valor foram classificados como clientes recentes. Já a segunda classe ficou com os clientes que fizeram sua última compra dentro do 2º trimestre, e foram classificados como neutros. Quanto à terceira e à quarta classe, ambas foram classificadas como negativas, uma vez que os clientes contidos possuem sua última compra dentro do intervalo do 3º e 4º trimestre respectivamente, ou seja, foram considerados como clientes perdidos. Vale ressaltar que a média deste atributo computado em trimestres ficou em 2,24. Já a mediana ficou em 2 e a moda em 1. Sendo assim, a moda não entrou no cálculo da faixa neutro,

isto porque o portal está adquirindo novos clientes em uma velocidade maior do que as que foram obtidas em períodos anteriores. Este fato corrobora a necessidade de um especialista de negócio avaliar a criação das faixas de valores no momento das definições.

- **Atributo F** – não passou por nenhuma transformação. Entretanto ele foi classificado da seguinte forma, conforme descrito na tabela a seguir:

Tabela 6.2: Divisão das classes do atributo F e seus comportamentos.

<b>Faixas (Frequência)</b>	<b>Comportamento</b>
1	Cliente Neutro
2	Cliente Negativo (Não Frequente)
$\geq 3$	Cliente Positivo (Frequente)

Os clientes com 1 compra foram classificados como neutros (afinal, o cliente com uma única compra não pode ser considerado frequente, já que ele pode ser um cliente novo ou até eventual). Já os clientes com 2 pedidos foram considerados negativos, uma vez que entre os clientes frequentes a média ficou em 2,54, a mediana em 2 e a moda também em 2. Por último, a classe de clientes positivos, que ficou formada por todos aqueles que fizeram 3 ou mais compras durante o período analisado.

- **Atributo M** – o atributo M foi associado a um novo valor correspondente a cada uma das cinco faixas criadas conforme demonstrado na Tabela 6.3:

Tabela 6.3: Divisão das classes do atributo M e seus comportamentos.

<b>Faixas (Monetário)</b>	<b>Valores Contidos</b>	<b>Comportamento</b>
5	$> 500$	Cliente Positivo (Monetário)
4	$> 350$ e $\leq 500$	Cliente Positivo (Monetário)
3	$\geq 200$ e $\leq 350$	Cliente Neutro
2	$\geq 100$ e $< 200$	Cliente Negativo (Não Monetário)
1	$< 100$	Cliente Negativo (Não Monetário)

O atributo M foi dividido em 5 faixas, ficando as duas primeiras com os clientes considerados positivos (uma faixa com valores acima de R\$ 350 e menor ou igual a R\$ 500; e outra com valores acima de R\$ 500). Já o cliente considerado como neutro teve a

sua faixa de valores definida entre maior ou igual a R\$ 200 e menor ou igual a R\$ 350. E por fim, as duas últimas faixas com os clientes negativos (uma faixa com valores maior ou igual a R\$ 100 e menor que R\$ 200; e outra com valores abaixo de R\$ 100). Vale ressaltar que a média deste atributo ficou em 237,60; a mediana em 206 e a moda em 329,99.

- **Atributo P** – também foi associado a um novo valor correspondente a cada uma das cinco faixas criadas, conforme demonstrado na Tabela 6.4:

Tabela 6.4: Divisão das classes do atributo P e seus comportamentos.

<b>Faixas (Monetário)</b>	<b>Valores Contidos</b>	<b>Comportamento</b>
5	$\geq 70\%$	Cliente Positivo (Lucrativo)
4	$> 50\% \text{ e } < 70\%$	Cliente Positivo (Lucrativo)
3	$\geq 40\% \text{ e } \leq 50\%$	Neutro
2	$\geq 20\% \text{ e } < 40\%$	Cliente Negativo (Não Lucrativo)
1	$< 20\%$	Cliente Negativo (Não Lucrativo)

O atributo P (*profitable*) foi dividido em 5 faixas, ficando as lucrativas, ou seja, os clientes positivos, com uma faixa maior ou igual a 70%, e outra com valores acima de 50% e menor do que 70%. Já a faixa que corresponde ao cliente neutro ficou definida entre maior ou igual a 40% e menor ou igual a 50%. E por fim, as duas últimas faixas com os clientes negativos, com uma faixa entre maior ou igual a 20% e menor do que 40%; e outra com todos os clientes com lucratividade menor do que 20%.

Desta forma, o estudo reorganizou todo o conjunto de dados para ajudar no processo de segmentação dos clientes, com o intuito de facilitar a execução do algoritmo na busca pelo melhor critério de separação entre os *clusters*.

## 6.4 Definição das Faixas RFM/P

Um novo critério para classificação RFM/P foi definido pelo estudo a partir de uma das soluções recomendadas. Assim, um *cluster* ou registro de cliente não será classificado apenas como positivo ou negativo de acordo com o valor da média global. Mas sim como positivo, neutro ou negativo a partir da criação de três faixas de valores definidas previamente pelo conteúdo observado no conjunto de dados analisado.

Neste caso, este critério de definição das faixas levou em conta uma fundamentação baseada no modelo de negócio observado, assim a delimitação dos valores mínimo e máximo de cada intervalo, para cada atributo, ficou distribuída da seguinte forma:

- **Atributo R** – em trimestres, definidos pelas seguintes faixas:

Tabela 6.5: Faixas de valores para classificação RFM/P do atributo R.

<b>Valores das Faixas</b>	<b>Comportamento</b>
< 2	Clientes Positivos (Recente)
$\geq 2$ e $< 3$	Clientes Neutros
$\geq 3$	Clientes Negativos (Não Recente)

Deste modo, todo *cluster* que ficar com a sua média menor do que 2 será classificado de forma positiva, representando um segmento de clientes recentes. Nesta faixa encontram-se os grupos de clientes que fizeram compras com a empresa nos últimos três meses, o que configura um tipo de cliente ativo.

Já um *cluster* que tiver sua média entre maior ou igual a 2 e menor do que 3 será classificado como neutro, ou seja, clientes que não são recentes, mas que ainda não abandonaram a empresa e que estão dentro do próximo ciclo de compra. Ressalta-se que os clientes contidos nesta faixa devem ser acompanhados com atenção pelos estrategistas de marketing, pois é neste período que o cliente poderá ter a maior intenção de recompra. Por isso, é com este grupo de clientes que as campanhas de retenção devem ser realizadas.

E por fim, o *cluster* que tiver sua média maior ou igual a 3 será considerado negativo, ou seja, clientes perdidos e que já abandonaram a empresa. Neste caso, é importante acompanhar este grupo para identificar o comportamento de compra daqueles clientes que deixaram de fazer negócio com a companhia. Além disso, é com este grupo de clientes que as campanhas de reativação devem ser realizadas.

- **Atributo F** – semelhante à classificação associada a cada registro de clientes, o atributo F ficou com as seguintes faixas:

Tabela 6.6: Faixas de valores para classificação RFM/P do atributo F.

<b>Valores das Faixas</b>	<b>Comportamento</b>
$\geq 3$	Clientes Positivos (Frequente)
$< 2$	Clientes Neutros
$\geq 2$ e $< 3$	Clientes Negativos (Não Frequente)

Neste caso, todo *cluster* que tiver sua média maior ou igual a 3 será considerado como positivo, pois estes são os verdadeiros clientes frequentes, ou seja, os clientes fiéis, aqueles que já possuem um hábito de compra e que apresentam um relacionamento de confiança estabelecido com a empresa. Eles estão satisfeitos e merecem atenção em qualquer eventualidade, até porque o custo para manter estes clientes é menor do que o custo de aquisição de um novo.

Já o *cluster* que obtiver a sua média menor do que 2 será considerado como neutro. Afinal, somente o cliente que já fez mais de uma compra é que deve ser considerado como cliente frequente. Inclusive, quando computada, a média não deve levar em consideração os clientes que possuem apenas 1 pedido (isto deveria valer até mesmo para o cálculo da média no processo de classificação RFM padrão). Por isso, a média que foi considerada para a criação desta faixa levou em conta apenas os clientes com mais de 1 pedido.

E por último, o *cluster* que tiver sua média maior ou igual a 2 e for menor do que 3 será considerado como negativo. É que entre os clientes frequentes, este é o valor que indica um cliente não frequente. Porém, neste caso específico, vale uma observação: esta faixa de valor ficou reduzida por conta da assimetria do atributo F, se não fosse por isto, esta faixa teria um valor maior.

- **Atributo M** – o atributo M levou em consideração as seguintes faixas:

Tabela 6.7: Faixas de valores para classificação RFM/P do atributo M.

Valores das Faixas	Comportamento
> 350	Clientes Positivos (Monetário)
$\geq 200$ e $\leq 350$	Clientes Neutros
< 200	Clientes Negativos (Não Monetário)

O *cluster* que tiver média acima de 350 será considerado como positivo, ou seja, representa os clientes monetários, aqueles que possuem um gasto acima da média dos demais. Estes clientes devem possuir grande importância para a empresa e devem ser mantidos a qualquer custo, tendo em vista o valor desembolsado por eles.

Já o *cluster* que tiver média maior ou igual a 200 e menor ou igual a 350 será considerado neutro, pois os seus clientes estão dentro de uma faixa de gastos padrão. Este grupo expressa o cliente habitual, aquele que representa a maioria dos consumidores.

Enquanto que o *cluster* que obtiver média menor que 200 será considerado negativo, isto é, clientes não monetários, com baixo poder de compra e menor relevância para o negócio.

- **Atributo P** – por último o atributo P, que ficou com as seguintes faixas:

Tabela 6.8: Faixas de valores para classificação RFMP do atributo P.

Valores das Faixas	Comportamento
> 50%	Positivo (Lucrativo)
$\geq 40\%$ e $\leq 50\%$	Neutro
< 40%	Negativo (Não Lucrativo)

O *cluster* que ficar com média maior que 50% será considerado positivo, porque representa um grupo com clientes lucrativos e rentáveis. Este é o grupo de clientes com melhor valor para o negócio, pois uma compra efetivada por eles traz um ganho muito maior.

Já o cluster que tiver sua média maior ou igual a 40 e menor ou igual a 50 será considerado neutro, pois sua lucratividade está contida dentro da rentabilidade esperada pelo negócio.

Enquanto que o *cluster* que tiver a sua média menor que 40% será considerado negativo, ou seja, um grupo com um menor valor para a empresa e que precisa ter a sua rentabilidade melhorada.

Deste modo, foram estabelecidas as premissas – a partir das faixas de valores de cada atributo – que irão classificar os *clusters* entre positivo, neutro ou negativo. Sendo assim, o estudo avança para a criação dos modelos finais a fim de validar se as propostas sugeridas trarão de fato um melhor resultado em comparação com aqueles apresentados anteriormente.

# 7 Modelos Recomendados

## 7.1 Modelo Sugerido – RFM

Diferente do modelo adotado no tópico 5, esta etapa do estudo não utilizou o recurso *Sweep Clustering*, já que seu uso foi feito somente para auxiliar a busca por um melhor valor de  $K$ , ao fornecer apenas uma orientação para as análises anteriores.

Nesta nova fase, a busca pelo valor de  $K$  ideal foi feita a partir de uma pesquisa empírica através da construção de vários modelos. Deste modo, o estudo criou 19 modelos – com o valor de  $K$  variando entre 2 e 20 – para encontrar aquele que obtivesse o melhor valor para a métrica **Taxa de Acertos RFM Modelo**. Esta pesquisa é detalhada logo a seguir.

### 7.1.1 Análise

A execução realizada pela pesquisa computou os valores de qualidade produzidos por cada modelo – a partir da métrica **Taxa de Acertos RFM Modelo** – e os compilou na Tabela 7.1. Nela, o resultado final gerado pela Taxa de Acertos RFM Modelo e o respectivo valor de  $K$  utilizado por cada modelo (parâmetro de entrada no algoritmo *k-means*) é apresentado em ordem decrescente pela melhor taxa obtida. Neste caso, quanto maior for a taxa encontrada, melhor é a qualidade do modelo.

Tabela 7.1: Pontuação da Taxa de Acertos RFM do Modelo Sugerido.

Pontuação	Número de <i>Clusters</i>		Pontuação	Número de <i>Clusters</i>
88,23%	20		50,73%	11
79,24%	19		48,89%	7
76,41%	15		48,33%	10
70,85%	18		47,86%	12
70,25%	16		39,64%	6
66,87%	17		23,67%	3
61,99%	9		17,47%	5
56,34%	13		16,27%	4
56,12%	14		10,53%	2
54,20%	8			

E como pode ser observado, ao utilizar a Taxa de Acertos RFM Modelo como métrica de avaliação, o melhor resultado encontrado produziu 20 *clusters* distintos, no qual o maior agrupamento ficou com 270 registros e o menor apenas com 2 registros, o que representa somente 0,09% do conjunto de dados, conforme pode ser visto na Figura 7.1 e na Figura 7.2. Além disso, 5 *clusters* ficaram com menos de 1% do total de clientes (o que não atende ao critério inicial de performance), sendo que juntos estes grupos somados representam 2,18% dos registros.

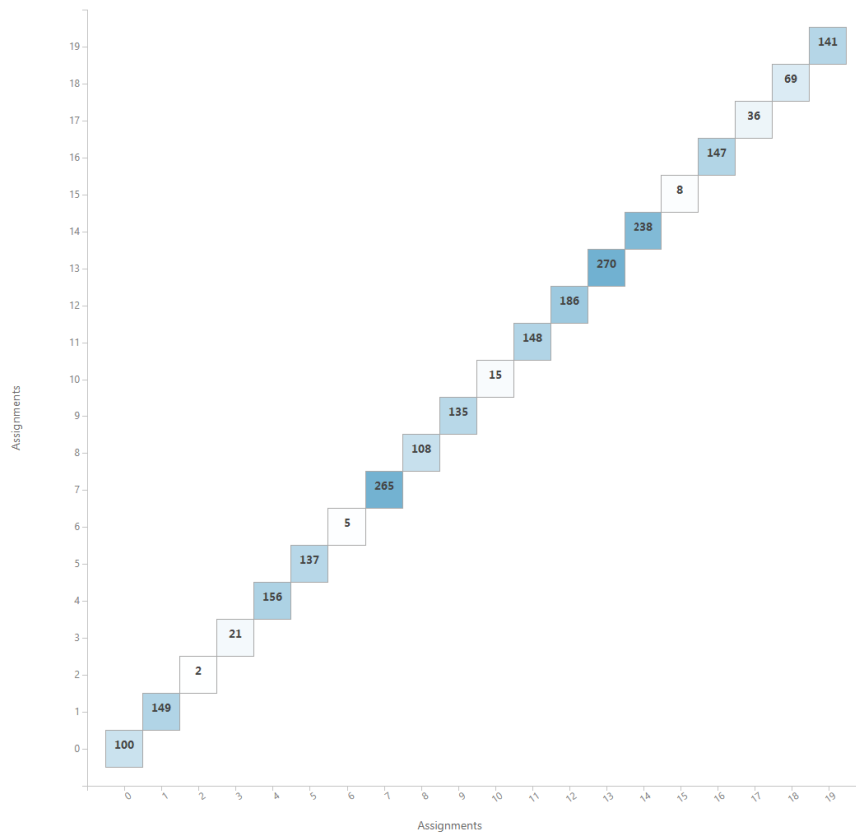


Figura 7.1: Quantidade de registros por *cluster* da análise RFM do Modelo Sugerido.

Já o gráfico da Figura 7.3 ilustra a representação do modelo gerado. E mais uma vez a visualização ficou prejudicada pela sobreposição dos *clusters*, que não permitiu um exame mais aprofundado através deste recurso. Porém, foi possível observar com mais detalhes um dos pontos já tratados nas análises anteriores, justamente sobre o problema da distribuição dos dados dentro de cada elipse que representa um determinado agrupamento. É que neste gráfico alguns *clusters* foram representados por uma faixa, sendo exibidos por uma linha estreita. Contudo, não há novamente como saber o quão distante está um registro do outro, e nem para saber se há uma concentração de dados em uma determinada área ou se estão dispersos por todo domínio do *cluster*.

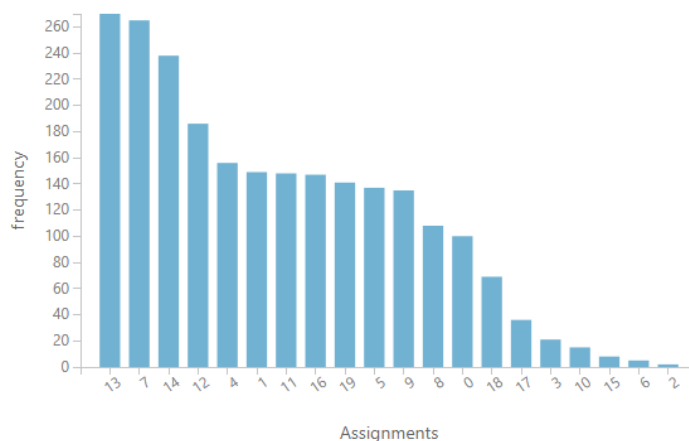


Figura 7.2: Histograma dos *clusters* da análise RFM do Modelo Sugerido.

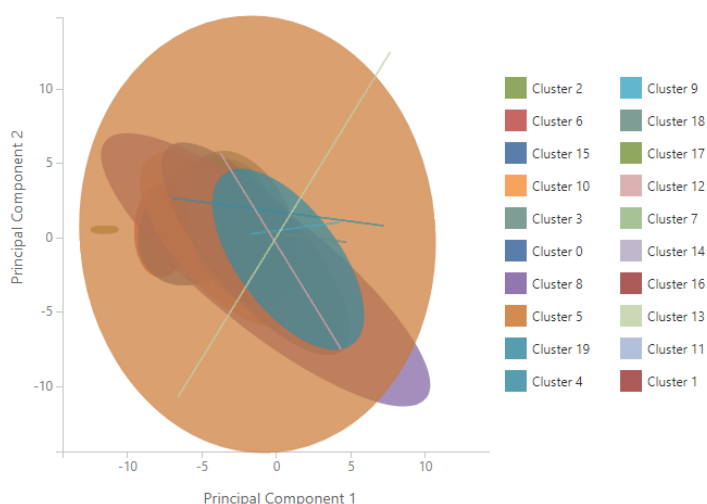


Figura 7.3: Gráfico da visualização dos *clusters* da análise RFM do Modelo Sugerido.

De qualquer forma, a seguir é descrito o comportamento de cada *cluster* de acordo com os resultados apresentados na Tabela 7.2 (tabela auxiliar).

O *cluster* 13 foi o maior de todos, com 270 registros (11,56%) e classificação RF<sup>o</sup> e M-, ou seja, clientes com comportamento padrão (neutro) nos atributos recente e frequente, mas negativo no atributo monetário. As médias encontradas foram de: 2 (R); 1 (F); e 103,82 (M), enquanto que a taxa de acertos RFM foi de 100%. Vale destacar que embora o atributo M tenha sido processado pelo algoritmo utilizando as novas faixas criadas, conforme descrito no tópico 0, as médias foram computadas pelo valor original do atributo contido em cada registro de cliente.

Em seguida o *cluster 7*, com 265 registros (11,34%) e classificação R+ e FM°, isto é, clientes recentes, porém neutros na frequência e no atributo monetário. As médias encontradas foram de: 1 (R); 1 (F); e 278,91 (M), com taxa de acertos RFM de 100%.

O próximo *cluster*, o de número 14, com 238 registros (10,19%), ficou formado pelo comportamento padrão de clientes, ou seja, atributos RFM°. Tanto que as médias encontradas foram de: 2 (R); 1 (F); e 276,52 (M). A taxa de acertos RFM deste grupo também foi de 100%.

Na sequência o *cluster 12*, com 186 registros (7,96%) e classificação FM° e R-, ou seja, clientes com comportamento padrão nos atributos frequente e monetário, mas negativo no atributo recente. As médias encontradas foram de: 3 (R); 1 (F); e 295,95 (M), porém a taxa de acertos RFM foi de 87%, influenciada pela taxa do atributo M, já que alguns registros possuem valor acima do limite superior da faixa que representa os clientes neutros. Este foi o primeiro cluster desta análise que não obteve um resultado de 100% na taxa de acerto.

Já o *cluster 4*, com 156 registros (6,68%), ficou com uma classificação R+, F° e M-, ou seja, clientes recentes, com frequência padrão (neutro) e não monetários. As médias encontradas foram de: 1 (R); 1,08 (F); e 80,20 (M), com taxa de acertos RFM de 92%, impactada pela taxa do atributo F, já que alguns registros possuem frequência com um valor igual a 2.

Os próximos dois *clusters* obtiveram comportamentos semelhantes, uma vez que o *cluster 1* – com 149 registros (6,38%) – e o *cluster 11* – com 148 registros (6,34%) – ficaram com classificação F° e RM-, ou seja, clientes com frequência padrão (neutro), mas não recentes e nem monetários. As médias encontradas foram de: 4 (R); 1 (F); e 130,66 (M) para o *cluster 1*; e 3 (R); 1 (F); e 131,51 (M) para o *cluster 11*. Ambos com taxa de acertos RFM de 100%.

Quanto ao *cluster 16*, com 147 registros (6,29%), ele ficou com classificação RFM semelhante ao do *cluster 4* (R+, F° e M-). Porém, suas médias foram de: 1 (R); 1 (F); e 138,13 (M), com taxa de acertos RFM de 100%.

O *cluster 19* ficou com 141 registros (6,04%) e classificação FM° e R-, comportamento igual ao do *cluster 12*. Entretanto, as médias encontradas foram de: 4 (R);

1,02 (F); e 286,97 (M), com taxa de acertos RFM de 84%, impactada pela taxa dos atributos F (98%) e M (86%).

Em seguida o *cluster 5*, com 137 registros (5,86%) e classificação R+, M° e F-, o que representa um grupo de clientes recentes, com comportamento monetário padrão (neutro), mas não frequentes. As médias encontradas foram de: 1,74 (R); 2 (F); e 237,71 (M), com taxa de acertos RFM de apenas 23%, impactada pela taxa dos atributos R (45%) e M (50%). Este foi o pior resultado entre todos os *clusters* deste modelo que obtiveram um total de registros superior a 1% dos dados. Com isso, a classificação RFM deste grupo ficou inconsistente.

Já o *cluster 9*, com 135 registros (5,78%) e classificação F° e RM-, obteve um comportamento semelhante aos *clusters 1* e 11. Contudo, as médias encontradas foram de: 3 (R); 1,01 (F); e 76,65 (M), com taxa de acertos RFM de 99%, impactada pela taxa do atributo F.

Na sequência o *cluster 8*, com 108 registros (4,62%) e classificação M+ e RF°, ou seja, clientes monetários, mas com um comportamento neutro nos atributos recente e frequente. Entretanto, as médias encontradas foram de: 2,86 (R); 1,10 (F); e 782,48 (M), com taxa de acertos RFM de somente 36%, impactada pelas taxas dos atributos R (40%) e F (90%). Com este resultado, a classificação RFM associada a este grupo ficou bastante prejudicada.

Quanto ao *cluster 0*, com 100 registros (4,28%), ele ficou com classificação RM+ e F°, isto é, um grupo de clientes recentes e monetários, mas com um comportamento padrão no atributo frequente. As médias encontradas foram de: 1 (R); 1,03 (F); e 574,09 (M), com taxa de acertos RFM de 97%, impactada pela taxa do atributo F.

Em relação ao *cluster 18*, com 69 registros (2,95%), ele ficou com classificação F° e RM-, comportamento semelhante aos encontrados nos *clusters 1*, 11 e 9. Em compensação, suas médias ficaram em: 4 (R); 1,03 (F); e 67,41 (M), com taxa de acertos RFM de 97%, sofrendo um baixo impacto pelo valor da taxa do atributo F.

Quanto ao *cluster 17*, com 36 registros (1,54%) e classificação M+ e RF°, ele ficou com comportamento igual ao do *cluster 8*. Porém, suas médias foram de: 2 (R); 1 (F); e 434,39, com taxa de acertos RFM de 100%.

Em compensação, os *clusters* 3 e 15, com 21 (0,90%) e 8 (0,34%) registros cada, ficaram semelhantes. Com classificação RF+ e M°, ou seja, representam clientes recentes e frequentes, mas com um comportamento monetário padrão. As médias encontradas foram de: 1,52 (R); 3,05 (F); e 220,45 para o *cluster* 3; e 1,38 (R); 5,38 (F); e 216,92 para o *cluster* 15. Já a taxa de acertos RFM foi de somente 19% no *cluster* 3 (a pior de todas neste modelo), impactada pelas taxas de R (52%) e M (48%); e de 25% para o *cluster* 15, impactada pelas taxas de R (75%) e M (38%). Estes resultados não conferem a estes grupos uma classificação RFM consistente, o que é insatisfatório.

E por último os *clusters* 10, 6 e 2, com 15 (0,64%), 5 (0,21%) e 2 (0,09%) registros cada, e classificação RFM+, ou seja, clientes recentes, frequentes e monetários, que representam os clientes VIPs do conjunto de dados. As médias encontradas foram de: 1,47 (R); 3,67 (F); e 356,63 (M) para o *cluster* 10; 1,60 (R); 7,60 (F); e 379,95 para o *cluster* 6; e 1 (R); 10,50 (F); e 644,82 (M) para o *cluster* 2. Em compensação, a taxa de acertos RFM foi de 47% no *cluster* 10, impactada pelas taxas de R (73%) e M (67%); de 20% para o *cluster* 6, impactada pelas taxas de R (60%) e M (60%); e de 100% para o *cluster* 2. Assim, pode-se dizer que os resultados da classificação RFM dos *clusters* 10 e 6 ficaram muito baixas, o que indica uma classificação com pouca qualidade.

Sendo assim, finalizado o detalhamento de cada agrupamento gerado, o estudo avança para entender o modelo produzido e distinguir o comportamento de separação entre os *clusters*. De todo modo, nesta nova fase, a comparação levou em conta a classificação RFM para melhor descrever a separação entre os *clusters*, conforme é apresentado a seguir.

Os *clusters* 17 e 8, ambos classificados como M+ e RF°, foram separados entre si pelo atributo M, já que o primeiro grupo ficou com média de 434,39 e clientes com valores entre 360,19 e 499; enquanto que o segundo ficou com média de 782,48 e clientes com valores entre 501,39 e 2.480. Tais valores representam uma escala do atributo M entre estes dois grupos. Além disso, pelo valor máximo encontrado no *cluster* 8, foi possível perceber que o registro que durante quase todo o estudo ficou segregado dos demais, representando assim um *outlier*, foi classificado dentro do agrupamento 8, como um registro comum de um cliente qualquer. De todo modo, foi possível notar também que *cluster* 17 ficou muito mais representativo, já que sua taxa de acertos foi de 100%; enquanto que o *cluster* 8 ficou com apenas 36%. A diferença entre eles foi que o primeiro

agrupou somente registros com valor do atributo R igual a 2 e valor do atributo F igual a 1, enquanto que o segundo ficou com valores de R entre 2 e 4, e de F entre 1 e 2. Mas como já informado no detalhamento do grupo 8 realizado anteriormente, o maior impacto foi causado pela taxa do atributo R, com apenas 40%.

Já os *clusters* 4 e 16, ambos classificados como R+, M° e F-, também foram separados entre eles pelo atributo M, pois o primeiro grupo ficou com a média de 80,20 e clientes com valores entre 9,10 e 93,50; enquanto que o segundo ficou com média de 138,13 e clientes com valores entre 103 e 199,99 (novamente uma representação em escala entre os dois grupos). De todo modo, o *cluster* 4 ficou com uma taxa de acertos de 92%, impactada pela taxa de F, já que alguns registros deste grupo ficaram com um valor igual a 2 neste atributo. Quanto ao atributo R, ambos ficaram somente com os clientes recentes (com valor igual a 1) e classificação RFM bastante confiável.

Comportamento semelhante – com separação pelo atributo M – foi encontrado entre os *clusters* 18 (F° e RM-), 1 (F° e RM-) e 19 (FM° e R-), já que ambos os grupos ficaram com os valores do atributo R idênticos (igual a 4), e médias de F bem próximas. Porém, a média de M do primeiro grupo ficou em 67,41, com clientes entre 36,40 e 90,40; enquanto que a média do segundo ficou em 130,66, com clientes entre 103 e 199,80; e a média do terceiro ficou em 286,97, com clientes entre 200 e 499.

O mesmo ocorreu com os *clusters* 7 (R+ e FM°) e 0 (RM+ e F°), ambos com valores do atributo R igual a 1, e média de F muito próxima de 1. Porém, o primeiro ficou com a média de M igual a 278,91, com clientes entre 206 e 348,79; enquanto que o segundo ficou com média de 574,09, com clientes entre 351,39 e 1.939,87. Esta sequência de valores de M encontradas nestes 2 grupos se complementa à escala encontrada pelos *clusters* 4 e 16.

Em compensação, o que separou o *cluster* 3 do *cluster* 15 (ambos classificados como RF+ e M°) foi o atributo F, já que o primeiro ficou com média de 3,05, e registros com valores entre 3 e 4; enquanto que o segundo ficou com média de 5,38, e registros com valores entre 5 e 6. Mais uma vez um efeito de escala entre os *clusters*; porém, neste caso, foi identificado no atributo F. Contudo, as taxas de acertos destes grupos ficaram muito baixas, conforme já descrito no detalhamento de cada *cluster*. Mas o que chamou atenção foi que em ambos os casos a taxa de acertos do atributo F foi de 100%, ou seja, o atributo que separou os *cluster* foi classificado corretamente. De qualquer maneira, as

taxas de acertos RFM obtidas por estes grupos ficaram muito baixas, deixando-os inconsistentes quanto à classificação RFM.

Da mesma forma, os *clusters* 15, 5 e 2 (classificados como RFM+) também foram separados entre si pelo atributo F, já que o primeiro ficou com média de 3,67 e clientes com frequência entre 3 e 4; enquanto que o segundo ficou com média de 7,60 e clientes com frequência entre 7 e 8; e o terceiro com média de 10,50 e clientes com frequência entre 10 e 11. Novamente, as taxas de acertos do atributo F ficaram em 100%, mas como já descrito anteriormente, as taxas de acertos RFM dos dois primeiros grupos ficaram em 47% e 20% respectivamente, ou seja, um valor não consistente para a classificação RFM dos grupos 15 e 5.

Assim, analisando os *clusters* apresentados nos dois últimos parágrafos, foi possível concluir que, apesar da taxa do atributo F não ter contribuído para as incertezas das taxas de acertos, os grupos classificados como F+ ficaram com um baixo valor nas suas respectivas taxas de acertos RFM (a exceção do grupo 2). Já que para agrupar os clientes mais frequentes, o algoritmo teve que colocar em um mesmo grupo clientes com muitas variações nos valores dos atributos R e M. Este comportamento é muito ruim, pois deixou a qualidade da classificação RFM destes grupos muito comprometida.

Esta conclusão também pode ser observada no comportamento do *cluster* 5, único grupo que possui os valores de F somente com registros iguais a 2. Nele, percebe-se que a taxa de F foi de 100%, enquanto que as demais foram de apenas 45% para R e 50% para M.

Dando continuidade, os *clusters* 1, 11, 9 e 18 (classificados como F° e RM-) foram separados entre si tanto pelo atributo R quanto pelo atributo M. É que neste caso, os *clusters* 1 e 18 ficaram com os registros de clientes com valor de R igual a 4; enquanto que os clientes dos *clusters* 11 e 9, ficaram com os registros de clientes com valor de R igual a 3. Porém, a separação entre os *clusters* 1 e 18 se deu pelo atributo M, já que o primeiro ficou com média de 130,66 e clientes com valores entre 103 e 199,80; enquanto que o segundo ficou com média de 67,41 e clientes com valores entre 36,40 e 90,40. Já a separação entre os *clusters* 11 e 9 foi feita também pelo atributo M, uma vez que o primeiro ficou com média de 131,51 e clientes com valores entre 101,18 e 199,80; enquanto que o segundo ficou com média de 76,65 e clientes com valores entre 11,90 e 95,20. De uma outra maneira, a separação entre os *clusters* 1 e 11 foi feita pelos valores

do atributo R, já que os valores de M ficaram bem semelhantes. O mesmo acontece na separação entre os *clusters* 9 e 18.

Já a separação entre os *clusters* 12 e 19 (classificação FM<sup>o</sup> e R-) se deu exclusivamente pelos valores encontrados no atributo R. Afinal, o primeiro ficou com média e registros com valores iguais a 3, e o segundo com valores iguais a 4. O valor do atributo M de ambos ficou muito semelhante, o que indica um comportamento de compra idêntico.

O mesmo comportamento foi encontrado entre os *clusters* 14 (RFM<sup>o</sup>) e 7 (R+ e FM<sup>o</sup>), pois os valores do atributo F ficaram idênticos, e os valores do atributo M ficaram muito parecidos, já que o primeiro ficou com média de 276,52 e o segundo com média de 278,91. Enquanto que os valores obtidos pelo atributo R do primeiro grupo foi de 2 e o do segundo foi de 1.

Em contrapartida, os *clusters* 17 (M+ e RF<sup>o</sup>), 13 (RF<sup>o</sup> e M-) e 14 (RFM<sup>o</sup>) ficaram com valores idênticos nos atributos R e F. Porém, o que os separa entre si é o valor do atributo M, que no primeiro grupo ficou com média de 434,39 e clientes com valores entre 360,19 e 499; enquanto que no segundo com média de 103,82 e clientes com valores entre 24,99 e 199,80; e no terceiro, média de 276,52 e clientes com valores entre 201,80 e 344,99.

Muito parecida também foi a separação entre os *clusters* 11 (F<sup>o</sup> e RM-), 9 (F<sup>o</sup> e RM-) e 12 (FM<sup>o</sup> e R-), já que ambos ficaram com valores idênticos no atributo R, mas com valores de M que se separam e se complementam entre si, uma vez que o primeiro ficou com média de 131,51 e clientes com valores entre 101,18 e 199,80; enquanto que o segundo ficou com média de 76,65 e clientes com valores entre 11,90 e 95,20; e o terceiro ficou com média de 295,95 e clientes com registros entre 201 e 499.

Por último, observa-se que a taxa média de acertos dos *clusters* foi bastante positiva, com um resultado de 76%, enquanto que a taxa de acertos RFM do modelo foi de 88,23%, conforme informado no início deste tópico.

Desta forma, o estudo conclui a análise do modelo sugerido a partir do índice criado pela Taxa de Acertos RFM Modelo.

### 7.1.2 Síntese dos Resultados e Melhorias Recomendadas

Como pôde ser visto, o modelo com a melhor taxa de acertos encontrou um total de 20 *clusters*, semelhante à quantidade produzida pelas análises RFM anteriores. Porém, os resultados encontrados foram ligeiramente maiores, já que obteve uma taxa média de acertos dos *clusters* de 76% e taxa de acertos RFM do modelo de 88,23%.

Ainda assim, o menor *cluster* ficou um total de 2 registros, o que representa somente 0,09% do conjunto de dados. Além disso, outros 5 *clusters* ficaram com menos de 1% do total de clientes. Com estas condições, o modelo não atende ao critério inicial de performance do estudo. Por isso, a tese sugere um complemento ao modelo criado para reduzir a quantidade de grupos com menos de 1% e agrupá-los em *clusters* maiores.

Sendo assim, esta pesquisa recomenda que os *clusters* 3 e 15, ambos com classificação RM+ e M°, sejam consolidados em um único grupo. Da mesma forma, recomenda que os *clusters* 10, 6 e 2, que ficaram com classificação RFM+, sejam agrupados em um só. Deste modo, um novo *cluster*, com a junção dos registros dos grupos 3 e 15, ficaria com um total de 29 registros (1,24%), enquanto que outro, criado a partir da junção dos grupos 10, 6 e 2, ficaria com um total de 22 registros (0,94%), um valor um pouco abaixo do estabelecido, mas dentro de um limite de tolerância aceitável.

Quanto à separação dos *clusters*, foi possível observar que todos os atributos R, F e M foram utilizados pelo modelo como critério de separação. Não necessariamente para a separação entre todos os *clusters*, mas na relação existente entre alguns deles, conforme descrito detalhadamente no tópico anterior. Vale destacar também que dos atributos utilizados o mais estável foi o atributo F, já que obteve menos erros do que as demais taxas.

Porém, apesar de não ter influenciado nos resultados, os *clusters* classificados como F+ ficaram com suas taxas de acertos RFM comprometidas. Isto porque ao agrupar os clientes mais frequentes, o algoritmo teve que colocar em um mesmo grupo clientes com muitas variações nos valores dos atributos R e M, o que tornou inconsistente a classificação RFM atribuída a estes grupos.

Por outro lado, é preciso destacar que o modelo poderia ser melhorado se os registros não correspondentes às classificações RFM fossem separados em um novo *cluster* ou inseridos em algum outro já existente. Por exemplo, o *cluster* 12, com 186

registros (7,96%) obteve uma taxa de acertos RFM de 87%, impactada pela taxa do atributo M. Isto porque alguns registros ficaram com valores de M acima de 350 neste grupo. Porém, se estes registros fossem separados, este *cluster* poderia ser dividido em 2, de modo que o primeiro ficaria idêntico ao cluster 12, apenas com um número menor, 162 registros (87%) dos clientes, e um novo – formado apenas pelos registros com valores de M acima de 350 – seria gerado, com um total de 24 registros (1%). Neste caso, a classificação deste novo *cluster* seria M+, F° e R-, diferente do *cluster* que lhe deu origem. De todo modo, em ambos os casos, a taxa de acertos RFM seria de 100%, o que aumentaria não só a taxa média de acertos dos *clusters*, como também a taxa de acertos RFM do modelo.

E por último, vale ressaltar que as mudanças recomendadas no tópico 0 e a reorganização dos dados, conforme descrito no tópico 6.4, surtiram efeito no resultado do modelo RFM, visto que as taxas encontradas ficaram acima daquelas que foram apresentadas nas análises anteriores. Resta saber se tais alterações trarão melhorias para os modelos RFMP.

De qualquer maneira, o estudo avança mais um pouco para colocar no modelo final RFM a sugestão recomendada neste tópico para que os *clusters* com menos de 1% sejam agrupados entre si.

### 7.1.3 Tabela Auxiliar – Modelo Sugerido RFM

Tabela 7.2: Tabela auxiliar com os resultados do Modelo Sugerido RFM.

G.	Qtd.	%	RFM+	RFM°	RFM-	Media R	Min R	Max R	Media F	Min F	Max F	Media M	Min M	Max M	% de Acertos			Qtd. Acertos RFM	% Acertos RFM
															% R	% F	% M		
13	270	11,56%		RF	M	2,00	2	2	1,00	1	1	103,82	24,99	199,80	100%	100%	100%	270	100%
7	265	11,34%	R	FM		1,00	1	1	1,00	1	1	278,91	206	348,79	100%	100%	100%	265	100%
14	238	10,19%		RFM		2,00	2	2	1,00	1	1	276,52	201,80	344,99	100%	100%	100%	238	100%
12	186	7,96%		FM	R	3,00	3	3	1,00	1	1	295,95	201	499,00	100%	100%	87%	162	87%
4	156	6,68%	R	F	M	1,00	1	1	1,08	1	2	80,20	9,10	93,50	100%	92%	100%	144	92%
1	149	6,38%		F	RM	4,00	4	4	1,00	1	1	130,66	103	199,80	100%	100%	100%	149	100%
11	148	6,34%		F	RM	3,00	3	3	1,00	1	1	131,51	101,18	199,80	100%	100%	100%	148	100%
16	147	6,29%	R	F	M	1,00	1	1	1,00	1	1	138,13	103	199,99	100%	100%	100%	147	100%
19	141	6,04%		FM	R	4,00	4	4	1,02	1	2	286,97	200	499,00	100%	98%	86%	118	84%
5	137	5,86%	R	M	F	1,74	1	3	2,00	2	2	237,71	54	494,99	45%	100%	50%	32	23%
9	135	5,78%		F	RM	3,00	3	3	1,01	1	2	76,65	11,90	95,20	100%	99%	100%	133	99%
8	108	4,62%	M	RF		2,86	2	4	1,10	1	2	782,48	501,39	2.480,00	40%	90%	100%	39	36%
0	100	4,28%	RM	F		1,00	1	1	1,03	1	2	574,09	351,39	1.939,87	100%	97%	100%	97	97%
18	69	2,95%		F	RM	4,00	4	4	1,03	1	2	67,41	36,40	90,40	100%	97%	100%	67	97%
17	36	1,54%	M	RF		2,00	2	2	1,00	1	1	434,39	360,19	499,00	100%	100%	100%	36	100%
3	21	0,90%	RF	M		1,52	1	3	3,05	3	4	220,45	116	343,32	52%	100%	48%	4	19%
10	15	0,64%	RFM			1,47	1	3	3,67	3	4	356,83	206,50	605,79	73%	100%	67%	7	47%
15	8	0,34%	RF	M		1,38	1	3	5,38	5	6	216,92	96,26	413,99	75%	100%	38%	2	25%
6	5	0,21%	RFM			1,60	1	3	7,60	7	8	379,95	281,63	491,86	60%	100%	60%	1	20%
2	2	0,09%	RFM			1,00	1	1	10,50	10	11	644,82	436,51	853,12	100%	100%	100%	2	100%

## 7.2 Modelo Final – RFM

### 7.2.1 Análise

O modelo final concentrou os registros dos grupos 3 e 15 em um único *cluster*, de modo que todos os clientes RF+ e M° ficassem em único agrupamento, formando assim um *cluster* com 29 registros, o que representa 1,24% do conjunto de dados, e que satisfaz à premissa de: buscar a maior quantidade possível de segmentos, mas desde que o menor *cluster* não possua menos do que 1% dos registros totais.

Além disso, este modelo também concentrou os clusters 10, 6 e 2, formando um único agrupamento com todos os clientes classificados como RFM+. Este grupo ficou com um total de 22 registros, o que representa 0,94% do total de clientes analisados. De todo modo, apesar de conter valor um pouco abaixo da premissa estabelecida, o estudo admite que este valor está dentro de um limite de tolerância aceitável.

Sendo assim, o estudo refaz o detalhamento do modelo anterior com a entrada destes novos *clusters* para descrever o comportamento de cada um de acordo com os resultados apresentados na Tabela 7.3.

A junção dos *clusters* 3 e 15 deu origem ao *cluster* A, que ficou com 29 registros (1,24%) e classificação RF+ e M°, com médias de: 1,48 (R); 3,69 (F); e 219,48 (M). Entretanto, a taxa de acertos RFM foi de apenas 21%, o menor valor entre todos os agrupamentos deste modelo, o que não justifica a classificação RFM atribuída. Já a união entre os *clusters* 10, 6 e 2 produziu um resultado melhor, com um total de 22 registros (0,94%) e classificação RFM+, com médias de: 1,45 (R); 5,18 (F) e 853,12 (M). A taxa de acertos foi de 45%, o que torna a sua classificação RFM não consistente.

Como os demais *clusters* ficaram idênticos ao do modelo anterior, nenhuma nova observação precisou ser gerada.

Sendo assim, com esta transformação o modelo final ficou com um total de 17 *clusters*, e com a mesma taxa de acertos RFM do modelo anterior, ou seja, 88,23%, e taxa média de acertos dos *clusters* em 81%. Um resultado ligeiramente melhor do que os 76% do modelo antecedente. Assim, o estudo dá por encerrada a busca por um modelo RFM e segue adiante para avaliar a criação dos modelos RFMP.

## 7.2.2 Tabela Auxiliar – Modelo Final RFM

Tabela 7.3: Tabela auxiliar com os resultados do Modelo Final RFM.

G.	Qtd.	%	RFM+	RFM°	RFM-	Media R	Min R	Max R	Media F	Min F	Max F	Media M	Min M	Max M	% de Acertos			Qtd. Acertos RFM	% Acertos RFM
															% R	% F	% M		
13	270	11,56%		RF	M	2,00	2	2	1,00	1	1	103,82	24,99	199,80	100%	100%	100%	270	100%
7	265	11,34%	R	FM		1,00	1	1	1,00	1	1	278,91	206	348,79	100%	100%	100%	265	100%
14	238	10,19%		RFM		2,00	2	2	1,00	1	1	276,52	201,80	344,99	100%	100%	100%	238	100%
12	186	7,96%		FM	R	3,00	3	3	1,00	1	1	295,95	201	499,00	100%	100%	87%	162	87%
4	156	6,68%	R	F	M	1,00	1	1	1,08	1	2	80,20	9,10	93,50	100%	92%	100%	144	92%
1	149	6,38%		F	RM	4,00	4	4	1,00	1	1	130,66	103	199,80	100%	100%	100%	149	100%
11	148	6,34%		F	RM	3,00	3	3	1,00	1	1	131,51	101,18	199,80	100%	100%	100%	148	100%
16	147	6,29%	R	F	M	1,00	1	1	1,00	1	1	138,13	103	199,99	100%	100%	100%	147	100%
19	141	6,04%		FM	R	4,00	4	4	1,02	1	2	286,97	200	499,00	100%	98%	86%	118	84%
5	137	5,86%	R	M	F	1,74	1	3	2,00	2	2	237,71	54	494,99	45%	100%	50%	32	23%
9	135	5,78%		F	RM	3,00	3	3	1,01	1	2	76,65	11,90	95,20	100%	99%	100%	133	99%
8	108	4,62%	M	RF		2,86	2	4	1,10	1	2	782,48	501,39	2.480,00	40%	90%	100%	39	36%
0	100	4,28%	RM	F		1,00	1	1	1,03	1	2	574,09	351,39	1.939,87	100%	97%	100%	97	97%
18	69	2,95%		F	RM	4,00	4	4	1,03	1	2	67,41	36,40	90,40	100%	97%	100%	67	97%
17	36	1,54%	M	RF		2,00	2	2	1,00	1	1	434,39	360,19	499,00	100%	100%	100%	36	100%
A	29	1,24%	RF	M		1,48	1	3	3,69	3	6	219,48	96,26	413,99	59%	100%	45%	6	21%
B	22	0,94%	RFM			1,45	1	3	5,18	3	11	388,27	206,5	853,12	73%	100%	68%	10	45%

## 7.3 Modelo Sugerido – RFMP

Da mesma forma como foi feita na análise anterior, esta avaliação também não fez uso do recurso *Sweep Clustering*. Assim, a execução foi realizada seguindo o mesmo modelo da última análise RFM, ou seja, computou os valores de qualidade produzidos por cada modelo RFMP – a partir da métrica **Taxa de Acertos RFMP Modelo** – e os ordenou de forma decrescente para definir o que obteve o melhor resultado como o modelo vencedor.

### 7.3.1 Análise

Conforme observado na Tabela 7.4, o melhor modelo produziu um total de 20 *clusters*, semelhante à análise RFM. Porém, o maior grupo ficou com 376 registros (16,10%) e o menor com 2 registros (0,09%), conforme pode ser visto na Figura 7.4 e Figura 7.5. Além disso, a Figura 7.6 ilustra de forma gráfica a representação do modelo gerado.

Entretanto, vale ressaltar que 7 grupos ficaram com menos de 1% do total de clientes, e que juntos estes segmentos representam 3,42% dos registros da base de dados analisada. Deste modo, este modelo também não atendeu à premissa estabelecida no início do estudo.

Tabela 7.4: Pontuação da Taxa de Acertos RFMP do Modelo Sugerido.

Pontuação	Número de <i>Clusters</i>		Pontuação	Número de <i>Clusters</i>
73,24%	20		43,07%	15
59,03%	16		38,91%	12
57,45%	18		37,11%	9
54,92%	14		36,43%	6
52,70%	19		27,91%	7
49,61%	10		15,28%	4
49,36%	13		10,87%	5
48,20%	8		10,79%	3
46,58%	17		10,27%	2
44,09%	11			

Ainda assim, como ele obteve o melhor resultado, o estudo detalhou o comportamento de cada *cluster* e as suas respectivas características para entender suas formações e os critérios de separação utilizados para segregá-los uns dos outros. Todo este detalhamento é descrito logo a seguir.

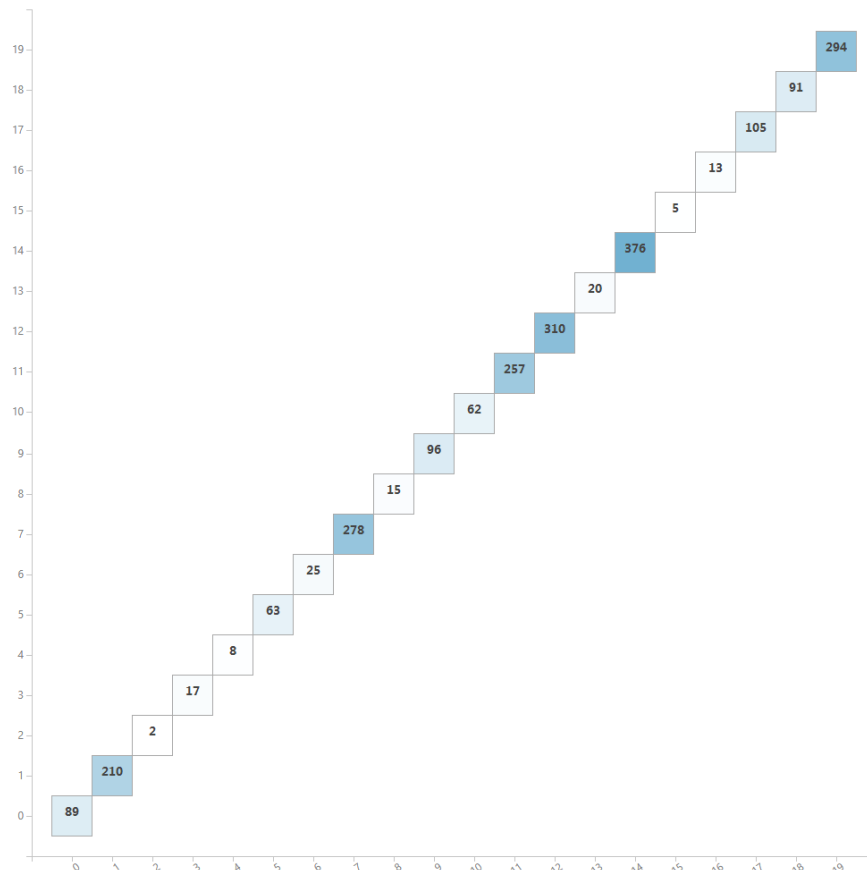


Figura 7.4: Quantidade de registros por *cluster* da análise RFMP do Modelo Sugerido.

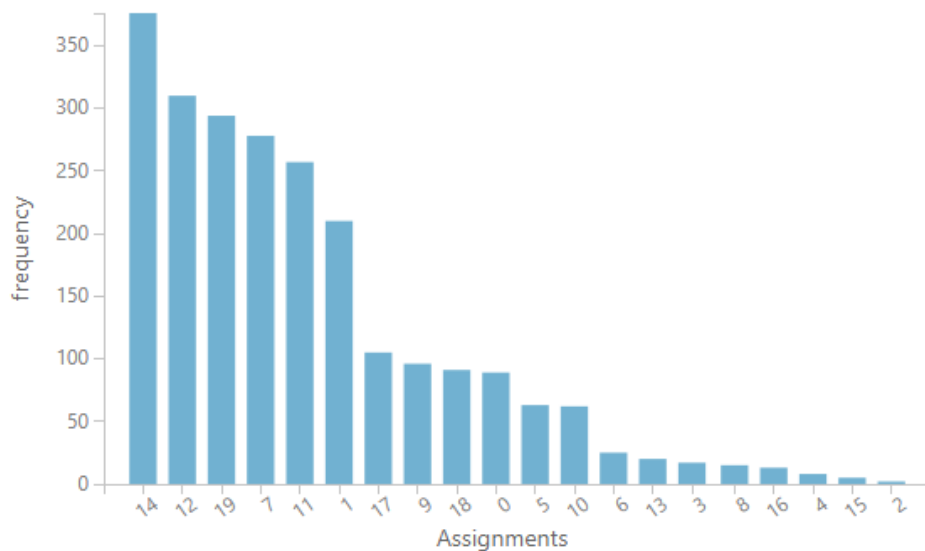


Figura 7.5: Histograma dos *clusters* da análise RFMP do Modelo Sugerido.

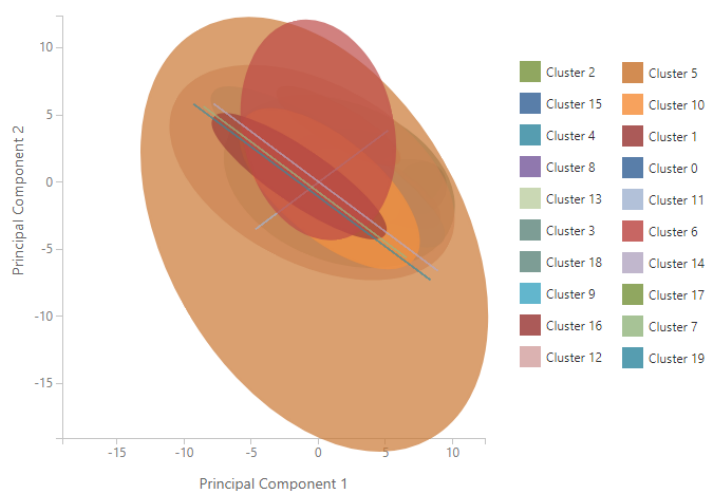


Figura 7.6: Gráfico da visualização dos *clusters* da análise RFMP do Modelo Sugerido.

O *cluster* 14 foi o maior de todos, com 376 registros (16,10%) e classificação R+ e FMP°, ou seja, clientes recentes, mas com comportamento padrão (neutro) nos atributos frequente, monetário e lucrativo. As médias encontradas foram de: 1 (R); 1 (F); 231,75 (M); e 43,77 (P), enquanto que a taxa de acertos RFMP foi de 65,95%, impactado pela taxa de acertos de M. Um resultado mediano que pouco justifica a classificação RFMP atribuída.

Em seguida o *cluster* 12, que ficou com 310 registros (13,27%) e classificação FMP° e R-, isto é, clientes com comportamento padrão nos atributos frequente, monetário e lucrativo, porém não recente. As médias encontradas foram de: 3,42 (R); 1,01 (F); 292,43 (M); e 44,18 (P), enquanto que a taxa de acertos RFMP foi de 85,48%, impactada pelas taxas de F (99,03%) e M (86,45%). Este é um resultado positivo e que confirma a classificação RFMP associada ao agrupamento.

Contudo, o *cluster* de número 19, com 294 registros (12,59%) e classificação R+, FP° e M-, representa um conjunto de clientes recentes, com comportamento neutro nos atributos frequente e lucrativo, porém não monetário. As médias encontradas foram de: 1,53 (R); 1 (F); 79,88 (M) e 45,34 (P), enquanto que a taxa de acertos RFMP foi de apenas 46,60%, impactada pela taxa de R, o que tornou inconsistente a classificação RFMP deste grupo.

Na sequência o *cluster* 7, com 218 registros (11,90%) e classificação FP° e RM-, ou seja, clientes com comportamento neutro nos atributos frequente e lucrativo, mas não

recente e nem monetário. As médias encontradas foram de: 3 (R); 1 (F); 105,65 (M); e 43,98 (P), enquanto que a taxa de acertos RFMP foi de 100%.

Já o *cluster* 11, com 257 registro (11%), ficou classificado como RFMP<sup>o</sup>, isto é, clientes que possuem um comportamento padrão (neutro) em todos os atributos. As médias encontradas foram de: 2 (R); 1 (F); 296,85 (M); e 43,55 (P), enquanto que a taxa de acertos RFMP ficou em 86,77%, impactada pela taxa do atributo M. Este é um bom resultado e que justifica a classificação RFMP.

O próximo *cluster*, o de número 1, ficou com 210 registros (8,99%) e classificação FP<sup>o</sup> e RM-, isto é, clientes com comportamento neutro nos atributos frequente e lucrativo, porém não recente e nem monetário. As médias encontradas foram de: 4 (R); 1,01 (F); 109,52 (M); e 44,64 (P), enquanto que a taxa de acertos RFMP ficou em 99,05, impactada pela taxa do atributo F.

O *cluster* 17, com 105 registros (4,49%) e classificação RFP<sup>o</sup> e M-, representa os clientes com comportamento padrão nos atributos recente, frequente e lucrativo, porém não monetário. As médias encontradas foram de: 2 (R); 1 (F); 141,77 (M); e 44,33 (P), enquanto que a taxa de acertos RFMP ficou em 100%.

Em contrapartida, o *cluster* 9, com 96 registros (4,11%), ficou com classificação M+ e RFP<sup>o</sup>, ou seja, clientes monetários, mas com comportamento padrão nos atributos recente, frequente e lucrativo. As médias encontradas foram de: 2,89 (R); 1,06 (F); 729,25 (M); e 45,03 (P), enquanto que a taxa de acertos RFMP de apenas 37,50%, impactada pelas taxas de R (37,50%) e F (93,75%) o que tornou a classificação RFMP inconsistente.

Da mesma forma, o *cluster* 18, com 91 registros (3,90%) e classificação R+, MP<sup>o</sup> e F-, isto é, clientes recentes e com comportamento padrão nos atributos monetário e lucrativo, porém não frequentes, também obteve um baixo resultado. Já que a taxa de acertos RFMP foi de apenas 35,16%, impactada pelas taxas de R (45,05%) e M (75,82%), outro resultado que não justifica a classificação RFMP atribuída. Entretanto, as médias encontradas foram de: 1,68 (R); 2 (F); 315,53 (M); e 44,22 (P).

Já o *cluster* 0, com 89 registros (3,81%), ficou com classificação RM+ e FP<sup>o</sup>, ou seja, clientes recentes e frequentes, mas com comportamento padrão nos atributos frequente e lucrativo. As médias encontradas foram de: 1 (R); 1 (F); 560,36 (M); e 44,48 (P), enquanto que a taxa de acertos foi de 100%.

Em seguida o *cluster 5*, com 63 registros (2,70%) e classificação RP+ e FM°, ou seja, clientes recentes e lucrativos, mas com comportamento padrão nos atributos frequente e monetário. As médias encontradas foram de: 1,77 (R); 1,03 (F); 209,19 (M); 58,39 (P), enquanto que a taxa de acertos foi de somente 26,98%, impactada pelas taxas de R (55,56%) e M (44,44%).

Na sequência o *cluster 10*, com 62 registros (2,65%) e classificação R+, P° e FM-, ou seja, clientes recentes, mas com comportamento padrão no atributo lucrativo e não recente e nem frequente. As médias encontradas foram de: 1,74 (R); 2 (F); 135,53 (M); e 44,79 (P), enquanto que a taxa de acertos foi de apenas 50%, impactada pela taxa de R. Este resultado tornou inconsistente a classificação RFMP atribuída.

O próximo *cluster*, o de número 6, ficou com 25 registros (1,07%) e classificação RFM° e P-, isto é, clientes com comportamento padrão nos atributos recente, frequente e monetário, porém não lucrativo. As médias encontradas foram de: 2,96 (R); 1 (F); 263,56 (M); e 36,07 (P), enquanto que a taxa de acertos RFMP ficou em 36%, impactada pelas taxas de R (36%) e M (76%), um resultado muito fraco e que não justifica a classificação RFMP deste grupo.

O *cluster 13*, com 20 registros (0,86%) e classificação RF+ e MP-, representa os clientes recentes e frequentes, mas com comportamento neutro nos atributos monetário e lucrativo. As médias encontradas foram de: 1,55 (R); 3,05 (F); 223,39 (M); e 44,01 (P), enquanto que a taxa de acertos RFMP foi de apenas 20%, impactada pelas taxas de R (50%) e M (50%) um péssimo resultado e que não justifica a classificação deste grupo.

Em compensação, o *cluster 3* ficou com 17 registros (0,73%) e classificação R+, F° e MP-, representando um conjunto de clientes recentes, com comportamento neutro no atributo frequente, porém não monetário e nem lucrativo. As médias encontradas foram de: 1 (R); 1,23 (F); 130,96 (M); e 36,82 (P), enquanto que a taxa de acertos RFMP foi de 82,35%, impactada pela taxa do atributo F.

Já o *cluster 8*, com 15 registros (0,64%), ficou classificado como RFM+ e P°, o que representa os clientes recentes, frequentes e monetários, mas com comportamento padrão na lucratividade. As médias encontradas foram de 1,46 (R); 3,66 (F); 356,83 (M); e 44,14 (P), enquanto que a taxa de acertos RFMP foi de 46,67%, impactada pelas taxas de R (73,33%) e M (66,67%).

Em seguida o *cluster* 16, com 13 registros (0,56%) e classificação M+, RF° e P+, ou seja, clientes monetários, com comportamento neutro nos atributos recente e frequente, porém não lucrativos. As médias foram de: 2,23 (R); 1,15 (F); 1.267,01 (M) e 37,67 (P), com taxa de acertos de 23,08%, impactada pelas taxas de R (30,77%) e F (84,62%).

O próximo *cluster*, de número 4, ficou com 8 registros (0,34%) e classificação RF+ e MP°, um comportamento semelhante ao do *cluster* 13. Porém as médias foram de: 1,37 (R); 5,37 (F); 216,92 (M); e 43,19 (P), enquanto que a taxa de acertos foi de apenas 25%, impactada pelas taxas de R (75%) e M (37,50%).

E por último os *clusters* 15 e 2, com 5 registros (0,21%) e 2 registros (0,09%) cada. Ambos ficaram com classificação RFM+ e P°, comportamento semelhante ao encontrado no *cluster* 8. Já as médias encontradas no *cluster* 15 foram de: 1,6 (R); 7,6 (F); 379,95 (M); e 43,71 (P). Enquanto que as médias do *cluster* 2 foram de: 1 (R); 10,5 (F); 644,81 (M); e 44,33 (P). A taxa de acertos RFMP do primeiro *cluster* foi de apenas 20%, impactada pelas taxas de R (60%) e M (60%), enquanto que a taxa do segundo foi de 100%.

Com isso, o estudo finaliza o detalhamento de cada *cluster* e segue adiante para entender o critério de separação utilizado pelo modelo. Contudo, da mesma forma que foi feita na análise RFM anterior, a comparação levou em conta a classificação RFMP para descrever o comportamento de separação entre os *clusters*, conforme é apresentado a seguir.

Os *clusters* 7 e 1, classificados como FP° e RM-, foram separados entre si pelo valor contido no atributo R, já que o primeiro ficou apenas com os registros do atributo R igual a 3, enquanto que o segundo ficou com os registros de R igual a 4. Já o que diferenciou os *clusters* 19 (R+, FP° e M-) e 0 (RM+, FP°) foi o valor do atributo M, que no primeiro ficou negativo (média de 79,88), enquanto que no segundo ficou positivo (média de 560,36).

Entretanto, é interessante notar que estes *clusters* (7, 1, 19 e 0) ficaram com uma relação comportamental semelhante nos atributos F e P, já que em todos eles estes atributos foram considerados neutros. Porém, é possível analisar que os *clusters* 19 e 0 ficaram com o atributo R classificado como positivo, enquanto que os *clusters* 7 e 1

ficaram como o atributo R classificado como negativo. Ou seja, ao observar estes quatro agrupamentos nota-se que o algoritmo utilizou tanto o atributo M quanto o atributo R como critério de separação entre os *clusters*.

Comportamento semelhante, cuja separação entre os *clusters* se deu pelo atributo R, também pode ser encontrado entre os grupos 14 (R+ e FMP<sup>o</sup>) e 12 (FMP<sup>o</sup> e R-), já que o primeiro ficou apenas com os registros de R igual a 1, enquanto que o segundo ficou com os registros de R entre 3 e 4.

Já a separação entre os *clusters* 14 (R+ e FMP<sup>o</sup>) e 5 (RP+ e FM<sup>o</sup>) se deu pelo atributo P, já que no primeiro ficaram os registros com o valor de P dentro da faixa dos clientes com lucratividade padrão (neutro), enquanto que no segundo ficaram os clientes contidos dentro da faixa dos clientes lucrativos.

O mesmo ocorreu com a separação entre os *clusters* 6 (RFM<sup>o</sup> e P-) e 11 (RFMP<sup>o</sup>), pois no primeiro ficaram os registros de clientes não lucrativos, enquanto que no segundo ficaram os clientes com lucratividade padrão. Da mesma forma, os *clusters* 16 (M+, RF<sup>o</sup> e P-) e 9 (M+ e RFP<sup>o</sup>) também foram distinguidos pelo valor contido no atributo P, já que o primeiro ficou com os clientes não lucrativos, enquanto que o segundo ficou com os clientes com lucratividade padrão.

Em contrapartida, o critério de separação entre os *clusters* 15, 8 e 2 (ambos com classificação RFM+ e P<sup>o</sup>) foi o valor contido no atributo F, que no primeiro grupo ficou com variação dos registros entre 7 e 8, no segundo entre 3 e 4, e no terceiro entre 10 e 11. Este comportamento também pode ser visto no critério de separação entre os *clusters* 4 e 13 (ambos classificados como RF<sup>o</sup> e MP-), já que o primeiro ficou com os valores de F entre 5 e 6, enquanto que o segundo ficou com os valores entre 3 e 4.

Relação interessante foi encontrada entre os *clusters* 3 (R+, F<sup>o</sup> e MP-) e 18 (R+, MP<sup>o</sup> e F-), pois ambos possuem clientes recentes, mas que se invertem na comparação entre comportamento contido nos atributos F, M e P, já que o primeiro ficou com os clientes com comportamento padrão no atributo F e negativo nos atributos M e P, enquanto que o segundo ficou com os clientes com comportamento neutro nos atributos M e P, enquanto que o atributo F ficou negativo.

Distinção pelo atributo M foi encontrada entre os *clusters* 17 (RFP° e M-) e 9 (M+ e RFP°), já que o primeiro ficou com os clientes não monetários, enquanto que o segundo com os clientes monetários.

Já a relação entre os *clusters* 3 (R+, F° e MP-) e 10 (R+, P° e FM-), a separação se deu pelos atributos F e P, pois enquanto o primeiro ficou com os clientes com comportamento padrão no atributo F, o segundo ficou com os clientes não frequentes. Em compensação, o *cluster* 3 ficou com os clientes não lucrativos, enquanto que o *cluster* 10 ficou com os clientes com comportamento padrão na lucratividade.

Por fim, observa-se que a taxa média de acertos dos *clusters* foi de apenas 59,33%, enquanto que a taxa de acertos RFMP do modelo foi de 73,24%, conforme informado no início deste tópico, resultados inferiores aos que foram encontrados na análise RFM.

Desta forma, o estudo dá por concluído a análise do modelo sugerido a partir do índice criado pela Taxa de Acertos RFMP Modelo.

### **7.3.2 Síntese dos Resultados e Melhorias Recomendadas**

Conforme observado, o valor encontrado pela taxa de acertos RFMP do modelo foi de 73,24%, bem acima dos 61,30% da análise RFMP (*Z-Score*). Ainda assim, este valor ficou abaixo do que foi produzido pela última análise RFM, que ficou em 88,23%. Por outro lado, a taxa média de acertos dos *clusters* ficou em 59,33%, resultado ligeiramente maior do que o encontrado na análise RFMP (*Z-Score*) que foi de 51,81% e bem menor do que os 76% da análise RFM anterior.

Entretanto, este modelo produziu um total de 20 *clusters*, semelhante à análise RFM anterior e à análise RFMP (*Z-Score*), sendo que o maior agrupamento ficou com 376 registros e o menor ficou com apenas 2 registros. Além disso, outros 7 segmentos ficaram com menos de 1% (e juntos representam 3,42% dos clientes), o que não atende ao critério estabelecida no início do estudo. Por isso, da mesma forma que foi realizado na análise anterior, a pesquisa sugere que este modelo seja reajustado para tentar reduzir a quantidade de grupos com menos de 1% e agrupá-los em *clusters* maiores.

Desta forma, o estudo recomenda que os *clusters* 13 e 4, ambos com classificação RF+ e MP° sejam consolidados em um único grupo. Da mesma maneira,

sugere que os *clusters* 15, 8 e 2, todos com classificação RFM+ e P°, sejam agrupados em um só. Deste modo, um novo *cluster*, com a junção dos registros dos grupos 13 e 4, ficaria com um total de 28 registros (1,20%), enquanto que um outro, criado a partir da junção dos grupos 15, 8 e 2, ficaria com um total de 22 registros (0,94%), um valor um pouco abaixo do estabelecido, mas como na análise anterior, dentro de um limite de tolerância aceitável.

Entretanto, apenas a junção destes *clusters* não reduz a quantidade de grupos com menos de 1%. Afinal, os grupos 3 (R+, F° e MP-) e 16 (M+, RF° e P-) continuariam sem atender aos critérios de performance, e como possuem um comportamento único no modelo (classificação RFMP atribuída é exclusiva), não há como uni-los a qualquer outro de forma direta. Por isso, a tese deve abordar uma forma de associá-los a outros *clusters*, a partir de um novo critério, na tentativa de se ajustar à quantidade mínima estabelecida pelo estudo.

Já em relação à separação dos *clusters*, foi possível perceber que todos os atributos R, F, M e P foram utilizados como critério pelo modelo, e que os atributos F e P foram os mais estáveis quanto à classificação RFMP atribuída. Contudo, a inclusão do atributo P continuou a trazer uma dificuldade maior para o modelo, conforme observado nas análises RFMP anteriores.

Além disso, destaca-se que as mudanças recomendadas no tópico 0 e no tópico 6.4 também melhoram o resultado do modelo RFMP, já que as taxas obtidas foram melhores do que aquelas apresentadas nas análises RFMP anteriores. E por fim, vale notar que a variedade de classificações encontradas neste modelo foi maior do que a dos demais, e que as escalas de valores encontradas nos outros modelos não puderam ser observadas com tanta frequência neste último.

Assim, a tese parte para uma última tentativa a fim de se obter um melhor resultado nas taxas avaliadas e reduzir a quantidade de *clusters* com menos de 1%. Tudo isso será feito de forma manual, a partir da interpretação e ajustes dos resultados obtidos por este modelo.

### 7.3.3 Tabela Auxiliar – Modelo Sugerido RFMP

Tabela 7.5: Tabela auxiliar com os resultados do Modelo Sugerido RFMP.

G	Qtd.	%	RFMP +	RFMP °	RFMP -	Med R	Min R	Max R	Med F	Min F	Max F	Med M	Min M	Max M	Med P	Min P	Max P	% de Acertos				Qtd. RFMP	% RFMP
																		% R	% F	% M	% P		
14	376	16,10	R	FMP		1	1	1	1	1	1	231,75	103	345,19	43,77	40	50	100	100	65,96	100	248	65,96
12	310	13,27		FMP	R	3,42	3	4	1,01	1	2	292,43	200	499	44,18	40	49,56	100	99,03	86,45	100	265	85,48
19	294	12,59	R	FP	M	1,53	1	2	1	1	1	79,88	9,10	94,90	45,34	40,56	50	46,60	100	100	100	137	46,60
7	278	11,90		FP	RM	3	3	3	1	1	1	105,65	11,90	199,80	43,98	40,56	49,58	100	100	100	100	278	100
11	257	11,00		RFMP		2	2	2	1	1	1	296,85	201,80	499	43,55	40,64	50	100	100	86,77	100	223	86,77
1	210	8,99		FP	RM	4	4	4	1,01	1	2	109,52	36,40	199,80	44,64	40,66	49,87	100	99,05	100	100	208	99,05
17	105	4,49		RFP	M	2	2	2	1	1	1	141,77	100,49	199,80	44,33	40	50	100	100	100	100	105	100
9	96	4,11	M	RFP		2,89	2	4	1,06	1	2	729,25	501,39	1.920,37	45,03	41,12	48,85	37,50	93,75	100	100	36	37,50
18	91	3,90	R	MP	F	1,68	1	3	2	2	2	315,53	203,80	805,10	44,22	40,37	49,16	45,05	100	75,82	100	32	35,16
0	89	3,81	RM	FP		1	1	1	1	1	1	560,36	351,39	1.939,87	44,48	41,21	49,65	100	100	100	100	89	100
5	63	2,70	RP	FM		1,77	1	4	1,03	1	2	209,19	30	462,70	58,39	50,08	100	55,56	96,83	44,44	100	17	26,98
10	62	2,65	R	P	FM	1,74	1	3	2	2	2	135,53	53,13	199,80	44,79	40,32	49,04	50,00	100	100	100	31	50,00
6	25	1,07		RFM	P	2,96	2	4	1	1	1	263,56	44,90	404,99	36,07	7,28	39,92	36,00	100	76,00	100	9	36,00
13	20	0,86	RF	MP		1,55	1	3	3,05	3	4	223,39	116	343,32	44,01	40,66	46,93	50,00	100	50,00	100	4	20,00
3	17	0,73	R	F	MP	1	1	1	1,23	1	3	130,96	37,50	184,90	36,82	23,47	39,96	100	82,35	100	100	14	82,35
8	15	0,64	RFM	P		1,46	1	3	3,66	3	4	356,83	206,50	605,79	44,14	40,06	47,45	73,33	100	66,67	100	7	46,67
16	13	0,56	M	RF	P	2,23	1	4	1,15	1	2	1.267,01	525	2.480	37,67	34,15	39,88	30,77	84,62	100	100	3	23,08
4	8	0,34	RF	MP		1,37	1	3	5,37	5	6	216,92	96,26	413,99	43,19	40,90	45,80	75,00	100	37,50	100	2	25,00
15	5	0,21	RFM	P		1,60	1	3	7,60	7	8	379,95	281,63	491,86	43,71	42,78	44,46	60,00	100	60,00	100	1	20,00
2	2	0,09	RFM	P		1	1	1	10,50	10	11	644,81	436,51	853,12	44,33	43,36	45,31	100	100	100	100	2	100

## 7.4 Modelo Final – RFMP

### 7.4.1 Análise

Tomando por base o que foi feito no tópico anterior, esta análise passou a ajustar o modelo para tentar melhorar a qualidade das taxas de acertos e reduzir a quantidade de *clusters* com menos de 1%. O objetivo é fazer com que estes *clusters* se adaptassem às premissas estabelecidas no início do estudo. A seguir os detalhes dos ajustes:

As primeiras alterações foram feitas pela junção dos *clusters* 13 e 4, ambos com classificação RF+ e MP°, gerando assim um novo *cluster* A, com um total de 28 registros (1,20%) e mesma classificação. Porém, as médias encontradas ficaram em: 1,5 (R); 3,71 (F); 221,54 (M); e 43,77 (P), enquanto que a taxa de acertos RFMP ficou em 21,43%, impactada pelas taxas de R (57,14%) e M (46,43%). Além disso, a união entre os *clusters* 15, 8 e 2, todos com classificação RFM+ e P°, gerou um novo *cluster* B, com igual classificação, 22 registros (0,94%), e médias de: 1,45 (R); 5,18 (F); 388,26 (M); e 44,06 (P), enquanto que a taxa de acertos RFMP ficou em 45,45%, impactada pelas taxas de R (72,73%) e M (68,18%). Estas alterações serviram apenas para reduzir a quantidade de *clusters* com menos de 1% e não produziu qualquer alteração significativa na qualidade do modelo.

Em seguida, o *cluster* 14 teve parte dos seus registros separados, uma vez que o atributo M foi o responsável pelo baixo desempenho deste grupo. Afinal, o atributo M ficou classificado como neutro, porém alguns registros estavam abaixo do limite desta faixa. Deste modo, a separação destes registros deu origem a 2 novos *clusters*, sendo: o primeiro o *cluster* C, com 248 registros (10,62%), classificação R+ e FMP° (semelhante ao agrupamento de origem), mas com médias de: 1 (R); 1 (F); 280,08 (M); e 43,62 (P), e taxa de acertos RFMP em 100%; enquanto que o segundo, o *cluster* D, com 128 registros (5,48%), ficou com classificação R+, FP° e M-, ou seja, clientes recentes, com comportamento neutro nos atributos frequente e lucrativo, porém não monetário. As médias encontradas ficaram em: 1 (R); 1 (F); 138,11 (M); e 44,06 (P), enquanto que a taxa de acertos RFMP ficou com 100%. Neste caso, o que separou os *clusters* foi o atributo M, pois no primeiro ficaram aqueles registros com M°, enquanto que no segundo aqueles com M-. Esta alteração produziu um bom resultado para o modelo, já que ambos os *clusters* obtiveram taxas de 100% (além de gerar um novo segmento de clientes) e com

classificação RFMP distinta, que não havia sido identificado pelo modelo sugerido. Com isso, um *cluster* que tinha taxa de acertos em 65,96% foi dividido em 2 com taxas de acertos RFMP em 100% para ambos.

Outra alteração produzida foi realizada com o *cluster* 19, já que parte dos seus registros foi separada para ajustar a classificação do atributo R, responsável pela performance negativa da classificação RFMP atribuída a este grupo. Assim, a separação deste *cluster* deu origem a 2 novos grupos, sendo: o primeiro o *cluster* E, com 157 registros (6,72%), classificação RFP<sup>o</sup> e M- (semelhante ao *cluster* 17). Já as médias ficaram em: 2 (R); 1 (F); 78,81 (M); e 43,92 (P), enquanto que a taxa de acertos RFMP foi de 100%; já o segundo, o *cluster* F, com 137 registros (5,86%), ficou com classificação R+, FP<sup>o</sup> e M- (semelhante ao agrupamento de origem e ao *cluster* D gerado pela transformação anterior). As médias encontradas ficaram em: 1 (R); 1 (F); 81,10 (M); e 46,96 (P), enquanto que a taxa de acertos RFMP também ficou com 100%. Neste caso, o que diferenciou um *cluster* do outro foi o valor contido no atributo R, em que o primeiro ficou com os registros contendo valores iguais a 2; e o segundo ficou com os registros contendo os valores iguais a 1. Novamente, a alteração gerada produziu um excelente resultado, já que os *clusters* ficaram com taxas de 100%, aumentando assim a qualidade do modelo, visto que o *cluster* original possuía uma taxa de acertos RFMP de apenas 46,60%.

Mudança semelhante foi produzida com o *cluster* 9, já que o maior impacto foi causado pelo atributo R. Assim, a separação deste *cluster* deu origem a 2 novos grupos, sendo: o primeiro o *cluster* G, com 60 registros (2,57%), classificação M+, FP<sup>o</sup> e R-, ou seja, clientes monetários, com comportamento padrão nos atributos frequente e lucrativo, mas não recente (comportamento que até então não havia sido identificado pelo modelo). Já as médias ficaram em: 3,43 (R); 1,1 (F); 729,98 (M); e 45,27 (P), enquanto que a taxa de acertos RFMP foi de 90%. Já o segundo, o *cluster* H, com 36 registros (1,54%), ficou com classificação M+, RFP<sup>o</sup> (igual ao segmento de origem). As médias encontradas ficaram em: 2 (R); 1 (F); 728,03 (M); e 44,63 (P), enquanto que a taxa de acertos RFMP ficou em 100%. Novamente, o que diferenciou um *cluster* do outro foi o valor contido no atributo R, em que o primeiro ficou com os registros contendo valores iguais a 2; e o segundo ficou com os registros contendo os valores entre 3 e 4. Esta transformação gerou um bom resultado para o modelo, ao converter um único *cluster* com taxa de acertos de 37,50% em dois, com taxas muito superiores a esta.

Outra melhoria foi encontrada também na modificação do *cluster* 18, que obteve uma taxa de acertos de apenas 35,16%, impactada diretamente pela taxa de R, já que este grupo continha registros contendo valores entre 1 e 3 neste atributo. Deste modo, houve uma separação para agrupar os registros com valores iguais a 1 em um único *cluster*, e valores entre 2 e 3 em um outro, o que deu origem então a 2 novos grupos, sendo: o primeiro o *cluster* I, com 50 registros (2,14%), classificação RMP<sup>o</sup> e F-, ou seja, clientes com comportamento padrão nos atributos recente, monetário e lucrativo, porém não frequente (comportamento que até então não havia sido identificado pelo modelo). Já as médias ficaram em: 2,24 (R); 2 (F); 317,37 (M); e 44,29 (P), enquanto que a taxa de acertos RFMP foi de 56%, um pouco melhor do que os 36% do *cluster* original. Já o segundo, o *cluster* J, com 41 registros (1,76%), ficou com classificação R+, MP<sup>o</sup> e F-, isto é, clientes recentes, com comportamento padrão nos atributos monetário e lucrativo, porém não frequente (comportamento novo dentro do modelo analisado), enquanto que a taxa de acertos RFMP foi de 78,05%, um valor muito superior do que a encontrada pelo *cluster* original. Uma singularidade sobre a separação do *cluster* 18 foi que nenhum novo grupo gerado a partir desta segregação ficou com o comportamento (classificação RFMP) igual ao do *cluster* de origem.

O *cluster* 5 também passou por uma alteração para tentar melhorar o desempenho que foi de apenas 26,98%, com grande impacto do atributo M, pois semelhante à mudança do *cluster* 14, o *cluster* 5 também ficou classificado como neutro no atributo M, porém alguns registros estavam abaixo do limite desta faixa. Deste modo, a separação deste *cluster* deu origem a 2 novos *clusters*, sendo: o primeiro o *cluster* K, com 34 registros (1,46%), classificação RP+ e FM<sup>o</sup> (semelhante ao agrupamento de origem), mas com médias de: 1,61 (R); 1 (F); 293,73 (M); e 58,56 (P), e taxa de acertos RFMP em 50% (valor superior ao *cluster* de origem, mas que ainda assim não justifica a classificação RFMP atribuída); enquanto que o segundo, o *cluster* L, com 29 registros (1,24%), ficou com classificação RP+, F<sup>o</sup> e M-, ou seja, clientes recentes e lucrativos, com comportamento neutro no atributo frequente, porém não monetário. As médias encontradas ficaram em: 1,96 (R); 1,06 (F); 110,09 (M); e 59,19 (P), enquanto que a taxa de acertos RFMP ficou com 41,38%, o que melhorou a taxa original, mas que ainda assim não justifica a classificação RFMP atribuída. Neste caso, um *cluster* ficou separado do outro pelos valores do atributo M.

Apesar dos *clusters* produzidos pelas duas últimas transformações terem ficado com as taxas de acertos maiores do que as dos *clusters* originais, nota-se que os resultados começaram a ficar com um menor desempenho quando comparados às transformações iniciais. Isto demonstra um esgotamento das possibilidades de se melhorar o modelo. Por isso, o estudo deu por encerradas as tentativas de melhorias de qualidade através dos ajustes no modelo e partiu para uma última tentativa de reduzir a quantidade de *clusters* com menos de 1%.

Para isso, o primeiro passo foi juntar os *clusters* H e 16 em um único *cluster*, que ficou denominado como H1, com um total de 49 registros (2,10%), e classificação semelhante ao do *cluster* H. Porém, as médias encontradas ficaram em: 2,06 (R); 1,04 (F); 871,02 (M); e 42,78 (P), enquanto que a taxa de acertos RFMP ficou em 73,47%, menor do que a taxa de 100% do *cluster* H, porém maior que a taxa de 23,08% do *cluster* 16. Desta forma pelo menos, a tese reduziu mais um *cluster* que não atenderia à premissa inicial. Quanto à justificativa para a junção destes *clusters*, ambos possuíam os atributos M+ e RF<sup>o</sup>, com variação apenas no atributo P, ou seja, o *cluster* H era o mais próximo do *cluster* 16.

E por último, o estudo juntou os *clusters* D e 3 em um único *cluster*, denominado de D1, que ficou com um total de 154 registros (6,21%), e classificação igual ao do *cluster* D. Já as médias encontradas foram de: 1 (R); 1,02 (F); 137,27 (M); e 43,21 (P), enquanto que a taxa de acertos RFMP ficou em 88,28%, um valor menor do que os 100% do *cluster* D, porém maior que os 82,35% do *cluster* 3. E assim, a tese reduziu o último *cluster* com menos de 1% que restava no modelo final (a exceção do *cluster* B, que ficou com 0,94% e que foi justificado pelo estudo como aceitável). Em relação aos motivos que levaram a junção do *cluster* 3 ao D, este era o mais próximo pela classificação RFMP, visto que ambos possuíam comportamento R+, F<sup>o</sup> e M-, variando apenas o valor da classificação do atributo P.

Quanto aos demais *clusters* do modelo final, eles ficaram iguais ao do modelo RFMP sugerido, por isso nenhuma nova observação precisou ser feita, o que pode ser observado na Tabela 7.6.

Deste modo, com todos estes ajustes, o modelo final continuou com um total de 20 *clusters*, porém passou a ter uma taxa de acertos RFMP de 88,74% e taxa média de acertos dos *clusters* em 75,07%, valor muito superior aos encontrados pelo modelo

sugerido, que foi de 73,24% na taxa de acertos RFMP e dos 59,33% da taxa média de acertos dos *clusters*.

Além disso, se comparado ao modelo RFM final, este modelo também ganhou por uma ligeira vantagem, já que a taxa de acertos RFM ficou em 88,23%. Porém, vale ressaltar que nenhum ajuste foi feito para aumentar a qualidade do modelo RFM.

Por fim, vale destacar que, apesar das dificuldades impostas pela adição do atributo P, foi possível encontrar um bom resultado que fosse capaz de gerar uma taxa de acerto muito confiável para o estudo, o que justifica – em quase todo o modelo – as classificações RFMP atribuídas aos *clusters*.

Ainda assim, durante todas estas transformações, foi possível perceber que os *clusters* com classificação positiva ou negativa do atributo P obtiveram taxas de acertos muito baixas, mas não pela assertividade deste atributo. É que para acertar as taxas de P, o modelo acabou sacrificando os demais atributos, em especial os atributos R e M.

De todo modo, o estudo dá por encerrada a busca por um modelo RFMP e segue para um breve resumo do que foi encontrado no processo de desenvolvimento dos modelos finais.

## 7.4.2 Tabela Auxiliar – Modelo Final RFMP

Tabela 7.6: Tabela auxiliar com os resultados do Modelo Final RFMP.

G	Qtd.	%	RFMP +	RFMP °	RFMP -	Med R	Min R	Max R	Med F	Min F	Max F	Med M	Min M	Max M	Med P	Min P	Max P	% de Acertos				Qtd. RFMP	% RFMP
																		% R	% F	% M	% P		
12	310	13,27		FMP	R	3,42	3	4	1,01	1	2	292,43	200	499	44,18	40	49,56	100	99,03	86,45	100	265	85,48
7	278	11,90		FP	RM	3	3	3	1	1	1	105,65	11,9	199,8	43,98	40,56	49,58	100	100	100	100	278	100
11	257	11,00		RFMP		2	2	2	1	1	1	296,85	201,8	499	43,55	40,64	50	100	100	86,77	100	223	86,77
C	248	10,62	R	FMP		1	1	1	1	1	1	280,08	206	345,19	43,62	40,32	50	100	100	100	100	248	100
1	210	8,99		FP	RM	4	4	4	1,01	1	2	109,52	36,4	199,8	44,64	40,66	49,87	100	99,05	100	100	208	99,05
E	157	6,72		RFP	M	2	2	2	1	1	1	78,81	24,99	94,9	43,92	40,56	49,94	100	100	100	100	157	100
D1	145	6,21	R	FP	M	1	1	1	1,02	1	3	137,27	37,5	199,99	43,21	23,47	50	100	97,93	100	88,28	128	88,28
F	137	5,86	R	FP	M	1	1	1	1	1	1	81,10	9,1	93,5	46,96	40,56	50	100	100	100	100	137	100
17	105	4,49		RFP	M	2	2	2	1	1	1	141,77	100,49	199,8	44,33	40	50	100	100	100	100	105	100
0	89	3,81	RM	FP		1	1	1	1	1	1	560,36	351,39	1.939,87	44,48	41,21	49,65	100	100	100	100	89	100
10	62	2,65	R	P	FM	1,74	1	3	2	2	2	135,53	53,13	199,8	44,79	40,32	49,04	50,00	100	100	100	31	50,00
G	60	2,57	M	FP	R	3,43	3	4	1,10	1	2	729,98	508,1	1.920,37	45,27	41,12	48,85	100	90,00	100	100	54	90,00
I	50	2,14		RMP	F	2,24	2	3	2	2	2	317,37	203,8	805,1	44,29	40,37	48,16	76,00	100	74,00	100	28	56,00
H1	49	2,10	M	RFP		2,06	1	4	1,04	1	2	871,02	501,39	2.480,00	42,78	34,15	47,21	81,63	95,92	100	73,47	36	73,47
J	41	1,76	R	MP	F	1	1	1	2	2	2	313,29	204,75	613,24	44,14	42,16	49,16	100	100	78,05	100	32	78,05
K	34	1,46	RP	FM		1,61	1	4	1	1	1	293,73	206,97	462,7	58,56	50,08	100	61,76	100	82,35	100	17	50,00
L	29	1,24	RP	F	M	1,96	1	4	1,06	1	2	110,09	30	180	58,19	50,38	83,31	48,28	93,10	100	100	12	41,38
A	28	1,20	RF	MP		1,5	1	3	3,71	3	6	221,54	96,26	413,99	43,77	40,66	46,93	57,14	100	46,43	100	6	21,43
6	25	1,07		RFM	P	2,96	2	4	1	1	1	263,56	44,9	404,99	36,07	7,28	39,92	36,00	100	76,00	100	9	36,00
B	22	0,94	RFM	P		1,45	1	3	5,18	3	11	388,26	206,5	853,12	44,06	40,06	47,45	72,73	100	68,18	100	10	45,45
12	310	13,27		FMP	R	3,42	3	4	1,01	1	2	292,43	200	499	44,18	40	49,56	100	99,03	86,45	100	265	85,48
7	278	11,90		FP	RM	3	3	3	1	1	1	105,65	11,9	199,8	43,98	40,56	49,58	100	100	100	100	278	100

## 7.5 Resumo

### 7.5.1 Da Análise

O propósito geral continuou sendo a análise do processo de geração de modelos RFMP, em comparação com os modelos RFM, para avaliar se a introdução do parâmetro P traria algum impacto no processo de segmentação de clientes e na formação e separação dos seus respectivos *clusters*.

Contudo, a análise atual teve por objetivo examinar os resultados dos modelos sugeridos a partir da identificação dos desafios encontrados (conforme descritos no tópico 6.1) e avaliar se as soluções propostas (detalhadas no tópico 6.2) obtiveram algum êxito.

Assim, a partir das recomendações sugeridas, o conjunto de dados analisado passou por uma transformação, ao agrupar os dados de entrada em intervalos de classes, transformando os dados provenientes das variáveis contínuas em faixas de valores. Além disso, uma nova proposta foi utilizada para aumentar o grau de confiança das classificações RFM/P, ao classificá-los com base na criação de três faixas de valores, contendo os clientes Positivos (+), Neutros (°) e Negativos (-).

Ademais, diferente do modelo adotado no tópico 5, esta análise não utilizou o recurso *Sweep Clustering*, já que seu uso foi feito somente como auxílio para fornecer uma orientação sobre o melhor valor de  $K$ . Deste modo, nesta etapa do estudo, a busca pelo melhor valor de  $K$  foi feita a partir de uma pesquisa empírica – através da construção de diversos modelos – para identificar aquele que iria obter o melhor valor para a métrica **Taxa de Acertos RFM ou RFMP do Modelo**. Por isso, o estudo criou 19 modelos por análise (com o valor de  $K$  variando entre 2 e 20) para definir como vencedor aquele que obtivesse o melhor resultado nesta métrica.

E então, a partir destas premissas, o estudo realizou uma pesquisa para avaliar os diversos resultados produzidos pela criação de vários modelos a partir das recomendações sugeridas. De tal modo que estas avaliações foram realizadas pelas comparações dos resultados dos modelos RFM/P das análises feitas no capítulo 5 com os modelos gerados pelas análises deste capítulo.

Sendo assim, a primeira análise teve por objetivo a criação de modelos com os atributos R, F e M a partir das soluções recomendadas. Já a segunda teve por finalidade

melhorar o desempenho do modelo RFM anterior, através de pequenos ajustes, para que os *clusters* encontrados ficassem dentro do limite de performance estabelecido no início desta pesquisa. A terceira teve como propósito a criação de modelos com os atributos R, F, M e P a partir das soluções recomendadas. Enquanto que a quarta – e última análise – teve como intuito melhorar o desempenho do modelo RFMP anterior, ao ajustar manualmente os *clusters* gerados para que eles produzissem melhores resultados.

Além disso, o estudo também avaliou o comportamento de cada *clusters*, ao observar tanto os seus respectivos conteúdos quanto os critérios adotados para a separação entre eles.

Por tudo isso, é possível afirmar que a geração de cada modelo desta análise teve por finalidade reduzir os efeitos causados por cada um dos desafios identificados, aumentar a qualidade dos modelos e melhorar a assertividade dos *clusters* no que diz respeito às classificações RFM/P a partir das soluções recomendadas.

Desta forma, é descrito em seguida um resumo dos resultados que foram alcançados.

### 7.5.2 Dos Resultados

De um modo geral, é apresentada na Tabela 7.7 uma compilação dos resultados obtidos neste capítulo da tese.

Tabela 7.7: Resumo dos resultados da avaliação final dos modelos.

Análise	Métrica	Qtd. <i>Clusters</i>	Menor <i>Cluster</i>	Maior <i>Cluster</i>	Taxa Média Acertos <i>Clusters</i>	Taxa Acertos RFM / P
RFM	Modelo Sugerido	20	(0,09%) 2 registros	(11,56%) 270 registros	76,00%	88,23%
	Modelo Final	17	(0,94%) 22 registros	(11,56%) 270 registros	81,00%	88,23%
RFMP	Modelo Sugerido	20	(0,09%) 2 registros	(16,10%) 376 registros	59,33%	73,24%
	Modelo Final	20	(0,94%) 22 registros	(13,27%) 310 registros	75,07%	88,74%

Conforme observado, a primeira análise (Modelo Sugerido – RFM) gerou um total de 20 *clusters* ao utilizar a Taxa de Acertos RFM Modelo como métrica para a busca do melhor valor de *K*, no qual o maior agrupamento ficou com 270 registros e o menor com apenas 2 registros (0,09% do conjunto de dados). Esta quantidade de *clusters* ficou

semelhante às obtidas pelas análises RFM anteriores. Entretanto, os resultados das taxas qualitativas que mensuram o modelo ficaram maiores, já que a taxa média de acertos dos *clusters* ficou em 76% e a taxa de acertos RFM do modelo ficou em 88,23%.

Contudo, como o menor *cluster* ficou com menos de 1% dos registros, assim como outros quatro, o resultado final do modelo não atendeu ao critério inicial estabelecido. Por isso, a pesquisa recomendou um ajuste no modelo para que fosse reduzida a quantidade de grupos com menos de 1%. Assim, os *clusters* com menos de 1% foram consolidados de acordo com as suas respectivas classificações RFM, ou seja, a regra para que um *cluster* fosse unificado a outro levou em conta a classificação RFM atribuída a cada agrupamento.

Com estas premissas o estudo deu início aos ajustes necessários que produziram a segunda análise (Modelo Final – RFM). Desta forma, com este modelo a quantidade total de *clusters* ficou em 17, com o maior contendo os mesmos 270 registros (11,56%) do modelo sugerido e o menor com um total de 22 registros (0,94%), um valor um pouco abaixo do estabelecido, já que por uma diferença mínima de registros (apenas 1,36) este *cluster* não ficou com 1%.

Neste ponto, vale destacar que o estudo poderia ter manualmente realocado 2 registros de um *cluster* qualquer para compor o menor agrupamento. Assim, faria com que o critério de performance estabelecido fosse atendido. Porém, a tese cairia no problema de *overfitting* para o modelo, o que poderia se mostrar ineficaz para efeito de comparação em alguma análise futura. Por isso, o estudo optou por acolher o resultado com o menor *cluster* com apenas 0,94%, já que ficou dentro de um limite de tolerância aceitável e justificável pela tese.

Com isso, o Modelo Final RFM ficou com apenas 1 *cluster* com menos de 1%, enquanto que o Modelo Sugerido RFM ficou com cinco. Além disso, a taxa média de acertos dos *clusters* ficou um pouco maior, variando dos 76% do Modelo Sugerido RFM para 81% no Modelo Final RFM, enquanto que a taxa de acertos RFM do modelo permaneceu nos 88,23%.

Já em relação aos critérios de separação dos *clusters*, foi possível observar que todos os atributos R, F e M foram utilizados, com destaque para o atributo F, que obteve uma taxa de acerto melhor do que os demais. Entretanto, os *clusters* classificados como

F+ ficaram com suas taxas de acertos RFM comprometidas, pois ao agrupar os clientes mais frequentes, o modelo teve que colocar em um mesmo *cluster* clientes com muitas variações nos conteúdos de R e M, o que de certo modo é uma distorção do atributo F.

Ainda assim, vale destacar que o modelo poderia ter sua performance aperfeiçoada se os registros não correspondentes às classificações RFM fossem separados em novos *clusters* ou inseridos em algum outro com a classificação correlata a dos registros. Mas esta execução ficou para um estudo futuro, já que o objetivo da pesquisa era comparar os modelos RFM com os modelos RFMP.

De todo modo, fica o registro de que as soluções recomendadas no tópico 0 e a reorganização dos dados descritas no tópico 6.4 geraram um efeito positivo nos resultados dos modelos RFM produzidos nesta análise, uma vez que as taxas encontradas ficaram acima daquelas que foram apresentadas nas análises anteriores.

Já em relação aos modelos RFMP, o primeiro deles (Modelo Sugerido RFMP) obteve um total de 20 *clusters*, com taxa de acertos RFMP em 73,24%, acima dos 61,30% da análise RFMP (*Z-Score*). Enquanto que a taxa média de acertos dos *clusters* ficou em 59,33%, resultado ligeiramente maior do que o encontrado na análise RFMP (*Z-Score*), que foi de 51,81%. Contudo, estes valores ficaram menores dos que os encontrados pelo Modelo RFM Final.

De todo modo, o maior agrupamento ficou com 376 registros (16,10%) e o menor com 2 registros (0,09%), sendo que outros 6 *clusters* ficaram com menos de 1% do total de clientes, ou seja, uma performance que não atendia ao critério estabelecido pelo estudo. Por isso, da mesma maneira que na análise RFM, a pesquisa recomendou que o modelo fosse reajustado para reduzir a quantidade de grupos com menos de 1%, ao agrupá-los em *clusters* maiores. Além disso, os ajustes sugeridos também tiveram por objetivo buscar um aumento nas taxas produzidas pelo modelo, o que acabou por produzir o quarto e último modelo (Modelo RFMP – Final).

Sendo assim, uma série de adaptações foram realizadas, não só para que os *clusters* com menos de 1% fossem consolidados em um único grupo, mas também para que o modelo tivesse um melhor resultado. E isto foi feito a partir de um conjunto de operações, que permitiram ajustar os dados classificados erroneamente para a produção de *clusters* mais precisos através das classificações RFMP atribuídas aos *clusters*.

Todos estes ajustes fizeram com que a taxa de acertos RFMP subisse para 88,74%, e a taxa média de acertos dos *clusters* subisse para 75,07%, ou seja, valores superiores aos que foram encontrados pelo Modelo Sugerido. Entretanto, se comparado ao modelo RFM final, este modelo levou ligeira vantagem já que a taxa de acertos RFM ficou em 88,23%. Porém, vale ressaltar que nenhum ajuste foi feito para aumentar a qualidade do modelo RFM.

Entretanto, apesar de todas as mudanças realizadas, o Modelo Final – RFMP continuou com um total de 20 *clusters*, de modo que o maior ficou com 310 registros (13,27%) e o menor com apenas de 22 registros (0,94%). Novamente, um valor abaixo do que foi estabelecido, mas dentro de um limite aceitável pela tese. Além disso, vale notar que a variedade de classificações encontradas neste modelo foi maior do que a dos demais.

Já em relação aos atributos R, F, M e P, todos foram utilizados como critério de separação pelo modelo, sendo que os atributos F e P foram os mais estáveis em relação à classificação RFMP. Entretanto, a inclusão do atributo P continuou a trazer uma dificuldade a mais para o modelo, pois durante todas estas transformações foi possível observar que os *clusters* com classificação positiva ou negativa do atributo P obtiveram taxas de acertos mais baixas. Porém, isto não foi ocasionado pela assertividade deste atributo, mas sim pelo fato de que para acertar as taxas de P, o modelo teve que sacrificar a assertividade dos demais atributos, em especial os atributos R e M.

Por fim, destaca-se que as sugestões recomendadas nos tópicos 0 e 6.4 também melhoram o resultado do modelo RFMP, já que que as taxas obtidas foram superiores às aquelas encontradas nas análises RFMP anteriores.

## 8 Conclusões

### 8.1 Avaliação das Análises e dos Resultados

Este trabalho teve como propósito apresentar a criação de um novo modelo – denominado RFMP (*Recency – Frequency – Monetary – Profitability*) – como alternativa à criação dos tradicionais modelos RFM sugeridos por (Hughes, 1994), quando utilizados em conjunto com as técnicas de mineração de dados. O objetivo foi avaliar se a inclusão de um novo parâmetro P (da palavra inglesa *Profitability*), associado à lucratividade dos consumidores, traria algum impacto na segmentação dos clientes contidos em uma base de dados de um site de *e-commerce*. A ideia era tentar diferenciar os grupos de clientes não só pelo valor monetário – comportamento padrão dos modelos RFM – mas também através da rentabilidade. Já que pela proposta original, os modelos RFM ignoram uma das principais motivações das empresas: o lucro.

Assim, o estudo deu início às avaliações que constituíram o domínio da tese, ao desenvolver diversas análises para compreender a geração dos modelos RFM/P e o comportamento de separação entre os *clusters* a partir dos padrões de compras dos clientes. Tudo isto foi feito com o algoritmo de classificação não supervisionada *k-means*, através de uma plataforma de *Machine Learning* na nuvem.

Porém, logo surgiu a primeira dificuldade. Pois como se sabe, um dos maiores desafios do método *k-means* é justamente encontrar o valor ideal para o parâmetro *K*. Por isso, como uma tentativa de solução para este problema, o estudo utilizou como critério inicial pela busca do melhor valor de *K* uma abordagem orientada ao negócio. E definiu como ponto de partida a seguinte premissa: encontrar dentro do conjunto de dados a maior quantidade possível de segmentos, desde que o número de registros contidos no menor *cluster* não fosse inferior a 1% dos registros totais.

Vale destacar que a motivação para tal premissa foi a criação da maior quantidade de segmentos de clientes, a partir dos seus respectivos padrões de compras, para permitir um trabalho futuro de marketing pelos estrategistas de negócio da empresa proprietária do conjunto de dados analisado. Contudo, este limite de 1% se deu apenas pelo tamanho da base de dados, já que um *cluster* com um volume menor do que este

seria muito pequeno. Porém, a tese deixou claro em seu conteúdo que tal valor percentual poderia variar de acordo com o conjunto de dados analisado.

Apesar disso, a definição desta premissa para guiar a busca pelo melhor valor de  $K$  não foi suficiente para a continuidade do estudo, pois logo em seguida foi identificada a ausência de indicadores capazes de mensurar a qualidade dos modelos e suas respectivas assertividades em relação às classificações RFM/P. Pois do contrário, o estudo não teria como comparar os modelos entre si e muito menos determinar a consistência e a qualidade dos *clusters* e modelos produzidos.

Por isso – como contribuição da tese para o desenvolvimento de modelos RFM/P mais apurados – a pesquisa propôs a criação de três novos índices de mensuração para avaliar os resultados obtidos, sendo: o primeiro atrelado à qualidade individual de cada *cluster* produzido (em correspondência com a classificação RFM/P associada); o segundo, para mensurar a qualidade média dos *clusters* gerados pelos modelos; e por último, e mais importante, para determinar a qualidade e a assertividade geral do modelo.

Desta forma, ao utilizar estes índices em seu processo de avaliação, o estudo conseguiu dar continuidade a sua execução. E assim desenvolveu as quatro primeiras análises – RFM, RFMP, RFM (*Z-Score*) e RFMP (*Z-Score*) – que estabeleceram os entendimentos essenciais sobre os modelos RFM/P a partir do conjunto de dados utilizado. Neste ponto, vale destacar que o uso do recurso *Sweep Clustering*, disponibilizado pela plataforma utilizada, foi de extrema relevância para a produtividade da pesquisa. Afinal, com ele foi possível a construção e testes de vários modelos, através do uso de diversas métricas de precisão, que facilitaram o processo de análise. Entretanto, a tese avaliou somente o melhor resultado sugerido por cada métrica.

E a partir dos resultados destas análises foi observado que os modelos gerados com o uso da normalização *Z-Score* conseguiram obter uma qualidade maior em relação aos demais, já que às classificações RFM/P atribuídas ficaram mais consistentes. Isto porque a normalização dos dados conseguiu reduzir as distorções apresentadas pelo atributo F, o que permitiu inclusive que este atributo fosse considerado no processo de separação entre alguns *clusters*. Com isso, a qualidade dos agrupamentos melhorou e conseqüentemente o modelo passou a errar menos, visto que a assimetria deste atributo estava impactando diretamente nos resultados obtidos.

Também foi possível notar que a qualidade dos modelos RFMP ficou inferior às apresentadas pelos modelos RFM, já que nenhum modelo RFMP teve a taxa de acerto superior às encontradas pelos modelos RFM. Isto porque a inclusão do novo atributo acabou trazendo ao processo de classificação uma maior complexidade, o que impactou negativamente nos resultados gerados. Além disso, a maioria das médias dos *clusters* ficaram muito próximas à da média global nos modelos RFMP, o que aumentou a inconsistência da classificação pelo atributo P.

Ademais, estas análises propiciaram avaliar o comportamento dos *clusters*, observando não só os seus respectivos conteúdos, como também os critérios adotados para a separação entre eles. E para isso, o estudo teve que complementar os resultados das análises com a criação de uma tabela auxiliar que ajudou no entendimento final de cada modelo.

Contudo, a tese conseguiu identificar os principais desafios encontrados para o avanço da pesquisa, como: a concentração dos registros em poucos *clusters*; a assimetria do atributo F; uma maior complexidade na assertividade dos modelos a partir da adição do atributo P; e a comparação por um valor absoluto (média) como forma de atribuição da classificação RFM/P. Porém, vale destacar que todos estes desafios constatados tiveram como fundamento os valores dos índices de mensuração propostos, o que demonstrou a importância da criação e utilização das métricas recomendadas pela tese.

Entretanto, a partir destes desafios e de suas respectivas causas, a pesquisa propôs um conjunto de soluções para tentar reduzir os impactos dos problemas levantados, aumentar a qualidade dos modelos gerados e melhorar a assertividade dos *clusters* em relação às classificações RFM/P. E como principais pontos para as soluções destes desafios foram sugeridos os seguintes itens:

- Variar o valor do parâmetro  $K$  do algoritmo *k-means* para identificar aquele que iria produzir o melhor resultado para o índice Taxa de Acertos RFM/P;
- Gerar os modelos apenas com a normalização dos dados através do método *Z-Score* para reduzir o impacto da assimetria dos atributos; e
- Agrupar os dados de entrada em intervalos de classes para melhor representar a estrutura do negócio.

Além disso, e mais importante, a pesquisa também precisou desenvolver um novo método para a classificação dos modelos RFM/P – em substituição à forma tradicional utilizada pela metodologia padrão dos modelos RFM, quando realizados em conjunto com os métodos de *Machine Learning*. Esta nova proposta teve por objetivo aumentar o grau de confiança das classificações RFM/P e melhorar a consistência entre os *clusters* e seus respectivos conteúdos.

Desta forma, a tese passou a recomendar uma nova metodologia para que tanto os *clusters* quanto os clientes não fossem mais segmentados apenas como positivos ou negativos a partir de um determinado valor absoluto (no caso, o valor da média global dos atributos), mas sim pela classificação a partir da criação de três faixas de valores, contendo: os Clientes Positivos (+); os Clientes Neutros (°); e os Clientes Negativos (-). Em contrapartida, antes de testar a efetividade destas novas propostas, o estudo teve que reorganizar o conjunto de dados e definir as faixas de valores que iriam classificar os *clusters* e registros dos clientes a respeito desta nova metodologia.

Assim, a pesquisa seguiu adiante com uma nova rodada de avaliações para examinar se as recomendações sugeridas trariam algum aperfeiçoamento nos resultados dos modelos. Neste caso, os resultados dos índices de mensuração dos novos modelos foram comparados com aqueles gerados a partir da primeira análise. Além disso, o estudo continuou avaliando o comportamento de cada *clusters*, ao observar tanto os seus conteúdos quanto os critérios adotados para a separação entre eles.

Nesta nova etapa de avaliação, a primeira análise (Modelo Sugerido – RFM) teve por objetivo a criação de um modelo RFM a partir das soluções recomendadas, e utilizou como critério de definição para o melhor valor de  $K$  a métrica Taxa de Acertos RFM Modelo. Como resposta, o modelo produzido encontrou um resultado bem positivo, mas ainda com uma quantidade de *clusters* com menos de 1% do total dos clientes. Por isso, este modelo foi logo complementado por uma segunda análise (Modelo Final – RFM), que teve por finalidade melhorar o desempenho através de pequenos ajustes para que estes *clusters* ficassem dentro do limite de performance estabelecido no início desta pesquisa.

A partir destes ajustes, o modelo gerado produziu um total de 17 *clusters*, com o menor deles possuindo somente 22 registros (0,94%). Porém, para evitar o problema de *overfitting*, a tese acolheu este resultado. Afinal, foi por uma diferença de apenas 2

registros (ou melhor, 1,36) que este *cluster* não ficou com 1% dos dados, ou seja, com um total de 0,94% este *cluster* ficou dentro de um limite de tolerância aceitável e justificável pela tese (pois não faria sentido adotar um outro modelo com taxas de acertos inferiores só para que o menor *cluster* ficasse dentro da premissa acordada). Isto mostra que é preciso ter flexibilidade no momento de escolha do melhor modelo, e que apenas a automação do processo não é suficiente para a sua criação, pois mais importante do que o modelo obtido é saber interpretar os seus resultados.

Já em relação aos critérios de separação, foi possível observar que todos os atributos foram utilizados. No entanto, os *clusters* classificados como F+ ficaram com suas taxas de acertos comprometidas, pois ao agrupar os clientes mais frequentes o modelo teve que colocar em um mesmo *cluster* clientes com muitas variações nos demais atributos, o que pode ser considerado como uma pequena imperfeição ou ponto de melhoria para o modelo. Apesar disso, a aplicação das soluções recomendadas trouxe um resultado positivo na qualidade do Modelo RFM – Final, já que as taxas obtidas foram superiores às encontradas nos modelos da primeira análise.

Na sequência foi criada uma outra análise (Modelo Sugerido – RFMP) para avaliar se as soluções recomendadas também trariam uma evolução nos resultados dos modelos RFMP. Contudo, este modelo também precisou passar por uma nova análise (Modelo Final – RFMP) para ajustar manualmente os *clusters* gerados. Porém, neste caso os ajustes não foram feitos apenas para reduzir a quantidade de *clusters* com menos de 1%, mas também para aperfeiçoar os registros não correspondentes às classificações RFMP (através de uma reconfiguração do conteúdo), ao fazer com que eles fossem separados em novos *clusters* ou inseridos em algum outro com a classificação correlata a dos registros.

Como resultado, o Modelo Final – RFMP ficou com 20 *clusters* e obteve uma taxa de acertos RFMP de 88,74%, um excelente desempenho, principalmente se comparado com a melhor taxa da primeira análise RFMP, que foi de apenas 61,30%. Entretanto, quando comparado ao Modelo RFM – Final o ganho foi por uma estreita vantagem, já que a taxa de acertos ficou em 88,23% (e sem nenhum ajuste de performance neste modelo). Já em relação ao menor *cluster*, ele também ficou com apenas de 22 registros (0,94%). Portanto, dentro do limite aceitável que foi estabelecido pela tese no Modelo Final – RFM.

Quanto aos critérios de separação, todos os atributos R, F, M e P foram utilizados pelo modelo, de modo que os atributos F e P foram os mais estáveis nas classificações RFMP. Porém, ainda assim foi possível perceber que o atributo P continuou a trazer uma dificuldade para o modelo, já que os *clusters* com classificação positiva ou negativa neste atributo ficaram com taxas inferiores à dos demais. De todo modo, isto não foi ocasionado pela assertividade do atributo P, mas sim pelo fato de que para acertar as taxas de P o modelo teve que sacrificar o acerto dos demais atributos, em especial os atributos R e M.

Como ponto positivo, vale destacar a variedade de perfis encontrados nos *clusters* do Modelo RFMP – Final, já que as quantidades de classificações distintas foram maiores do que nos demais modelos. Em relação às sugestões recomendadas, o estudo finaliza admitindo que elas também produziram melhores resultados nas taxas dos modelos RFMP (apesar das dificuldades impostas pela adição do atributo P), pois os valores obtidos foram superiores aos encontrados nas análises RFMP da primeira avaliação.

Desta forma, o estudo reconhece que a aplicação de todas as recomendações apresentadas, como a criação das métricas de avaliações que serviram para suprir a ausência de indicadores capazes de mensurar os modelos RFM/P, assim como a recomendação de uma nova metodologia para classificar os modelos RFM/P (em substituição à forma tradicional utilizada), conseguiram reduzir os efeitos causados por cada um dos desafios identificados ao longo do desenvolvimento desta pesquisa.

Por isso a tese conclui suas alegações afirmando que suas propostas foram capazes de produzir um aumento na qualidade dos modelos gerados, assim como asseguraram uma melhor assertividade dos *clusters* em relação às classificações dos modelos RFM/P, o que tornou o processo muito mais confiável, já que o seu desempenho pôde ser constatado com êxito através dos resultados obtidos nos índices de mensuração dos modelos finais.

Entretanto, vale ressaltar que esta pesquisa foi realizada a partir de um caso real. E que, de acordo com a experiência deste autor, este é um caso muito representativo, pois as questões tratadas são comuns e se reproduzem com frequência em outras empresas e organizações. Contudo, apesar ter sido feita com apenas um único conjunto de dados, as diversas análises serviram para confirmar a eficácia do que foi proposto. Além disso, estudos futuros poderão comprovar também que os novos índices e métodos sugeridos

serão capazes de resolver os problemas identificados, o que trará com o tempo a consagração de todas as recomendações apresentadas pela tese.

## 8.2 Sobre o Ambiente Tecnológico

Este estudo também permitiu avaliar o ambiente de computação em nuvem da Microsoft, denominado *Azure Machine Learning* (ou *Azure ML*). Com ele foi possível realizar todas as tarefas intrínsecas ao processo de mineração de dados, como a limpeza e o tratamento das informações, as estatísticas básicas, as customizações através de consultas SQL e programação em linguagem *Python*, o desenvolvimento de modelos, testes e avaliação dos resultados.

Uma das principais vantagens observadas pelo uso desta plataforma em nuvem foi que não houve a necessidade de se preocupar com as típicas questões técnicas do ambiente de TI ao se iniciar um projeto, como a configuração dos servidores, instalações de softwares e manutenção, o que possibilitou focar naquilo que era mais importante: a solução de um problema real, e não em um tecnológico. Além disso, a disponibilização destes recursos, a partir de um modelo de serviço, permite a contratação de acordo com a necessidade e a capacidade de investimentos de cada projeto.

Em relação à utilização da ferramenta, o estudo conseguiu obter um bom desempenho a partir do seu uso. Principalmente pela simplicidade gerada a partir do prático recurso de arrastar e colar, o que facilitou muito o processo de criação dos modelos. Porém um ponto relevante, pois mais significativo do que apenas arrastar e colar os recursos disponíveis para a construção dos modelos é saber encadear de forma correta as atividades necessárias para a construção de um bom modelo. E muito mais importante ainda é saber interpretar os resultados obtidos por estes modelos.

Quanto ao algoritmo *k-means*, a execução foi realizada sem qualquer problema. E parte disto foi ocasionada pela quantidade de dados analisada, já que não era tão volumosa assim. Em compensação, a plataforma precisa melhorar bastante a saída dos resultados, principalmente o gráfico de visualização dos *clusters*, que durante todo o estudo manteve a sua funcionalidade comprometida, pois ficou com uma visualização confusa e incompleta. Tanto que em muitos casos o entendimento dos *clusters* através deste recurso ficou incompreensível. Apesar disso, foi possível explorar e compreender diversas outras funcionalidades existentes, principalmente aquelas que auxiliaram no

desenvolvimento do trabalho, como o recurso *Sweep Clustering*, tão discutido e detalhado durante toda a tese.

De todo modo, como um panorama geral da plataforma, o estudo conclui que: ela é rápida – principalmente pela forma utilizada para desenvolver o modelo (recurso de arrastar e colar); possui uma boa performance no treinamento dos modelos (mas fica a ressalva que o volume de dados utilizado pelo estudo foi pequeno); ela é de fácil aprendizado e adaptação; não possui um recurso para visualizar os registros contidos em cada *cluster* (fundamental para uma boa compreensão de um modelo de classificação não supervisionada); por rodar na nuvem é mais lento do que em uma máquina local (além disso, precisa de uma boa banda larga, algo que nem sempre é possível se encontrar por aqui); e permite a customização através de *scripts* em linguagem *R* ou *Python* (se um cientista de dados estiver interessado somente na execução a partir das suas rotinas e bibliotecas escritas nesta linguagem, eles irão encontrar um excelente ambiente para rodar os seus códigos, e portanto não precisarão se preocupar com o ambiente e infraestrutura, focando apenas na criação dos modelos).

Em compensação, por rodar na nuvem os dados também ficam nela. E isso acaba trazendo um risco, pois se um *hacker* invadir a nuvem ele terá acesso e controle a todo o conjunto de dados. Além disso, o estudo não conseguiu gerar um *backup* do modelo fora do ambiente utilizado, já que este é um ponto importante para proteger todo o trabalho em caso de uma eventual adversidade (como no caso de um ataque *hacker*). Outra questão que ficou sem resposta foi uma forma de voltar a um estado anterior (recurso *undo*) após a exclusão de um modelo (principalmente quando acontece de forma não intencional).

Sendo assim, o estudo conclui sua avaliação sobre a plataforma utilizada afirmando que os recursos disponíveis são suficientes para gerar e validar um modelo, e que de um modo geral estes recursos são plenamente capazes de apresentar um bom desempenho. Além disso, este é um serviço que pode ser dimensionado de forma rápida e eficiente para atender à demanda de investimentos de acordo com a necessidade das empresas.

### 8.3 Estudos Futuros

A propósito de todas as questões tratadas por esta tese – a partir do desenvolvimento de uma nova proposta e metodologia para os modelos RFM/P – algumas outras considerações merecem um estudo mais aprofundado em um trabalho futuro, conforme é descrito a seguir:

- ✓ Avaliar o histórico de cada grupo de clientes classificados através dos modelos RFM/P no intuito de se obter um entendimento da evolução de cada perfil ao longo do tempo;
- ✓ Identificar através de outras técnicas de *Machine Learning* o padrão de compras de cada perfil RFM/P atribuído pelos modelos gerados;
- ✓ Verificar se os clientes novos não precisam de um modelo próprio para avaliar qualquer alteração no padrão de compras e no perfil destes clientes;
- ✓ Usar os *clusters* e registros rotulados corretamente para criar um modelo de classificação supervisionada, com o objetivo de classificar novamente os registros que foram classificados de forma incorreta no processo de classificação não supervisionada;
- ✓ Utilizar alguma técnica de agrupamento *Fuzzy*, como o algoritmo *fuzzy c-means*, por exemplo, para avaliar o comportamento e a formação dos *clusters*;
- ✓ Revisitar os artigos publicados que fizeram uso do modelo RFM para validar a eficácia dos seus resultados quando comparados aos novos índices de mensuração sugeridos por esta tese.

Desta forma, o estudo indica os diversos caminhos que poderão ser tomados para evoluir com o trabalho desenvolvido nesta pesquisa.

## 8.4 Considerações Finais

A proposta da tese pela criação de um novo modelo RFMP, como alternativa aos modelos RFM, trouxe uma nova ótica sobre a forma pela qual se deve segmentar os clientes, já que pelo processo original os modelos RFM ignoram uma das principais motivações das empresas: a lucratividade. Afinal, como pôde ser observado, ao incluir um novo atributo P (da palavra inglesa *Profitability*), relacionado ao lucro dos consumidores, foi possível notar um impacto direto na formação dos *clusters* produzidos. Tanto que a pesquisa conseguiu diferenciar os grupos de clientes não só pelo valor monetário (comportamento padrão dos modelos RFM), mas como também a partir das suas respectivas rentabilidades.

Isto permitiu constatar que nem sempre os clientes com os maiores valores são de fato os mais lucrativos, circunstância que foi inclusive comprovada pela ausência de um *cluster* que tivesse os atributos P e M classificados simultaneamente como positivos no modelo final. É claro que esta afirmação vem do conjunto de dados analisado, mas que serve para alertar aos estrategistas de marketing sobre a importância de se conhecer de perto os diferentes perfis de consumidores que compõem a sua base de clientes. Afinal, as empresas precisam ser capazes de encontrar os clientes certos, entender quais tipos de clientes possuem e saber como manter os melhores. Pois no fim das contas, podem descobrir tardiamente que estão vendendo o produto errado, pelo preço errado, para o cliente errado.

Com isso em mente, são inúmeras as estratégias que poderão ser adotadas, uma vez que as empresas consigam segmentar corretamente as suas bases de dados a partir dos atributos definidos pelo modelo RFMP proposto. Porém, antes disso é preciso entender os problemas que podem ser resolvidos pela ciência de dados, para só depois transformá-los em modelos e ações que irão permitir às empresas tomarem decisões mais rápidas, melhores e que impactem positivamente nas suas estratégias de negócio.

Entretanto, vale ressaltar que um modelo jamais será perfeito, e por isso deverá ser avaliado constantemente por um especialista para definir o quanto de erro poderá suportar, pois sempre haverá uma imperfeição para se enfrentar (em especial nos de clusterização), como ocorrido, por exemplo, nos modelos finais produzidos por este estudo (já que nenhum modelo ficou com 100% de acerto). Este é um problema que irá

variar de acordo com o domínio de negócio de cada empresa, e dependerá da estratégia de execução das ações que serão adotadas a partir dos resultados obtidos pelos modelos desenvolvidos.

Por fim, mais importante do que criar os modelos e saber interpretá-los é dominar o que fazer com os resultados obtidos. Portanto, é necessário tirar o foco dos detalhes técnicos e ter a ciência de dados mais orientada às soluções de negócio, conectando diretamente os modelos desenvolvidos às ações estratégicas das empresas.

Tudo isso só reforça a importância de se definir previamente quem utilizará as soluções propostas e de que maneira criará valor para as companhias, já que haverá sempre um elemento humano necessário por todo o processo de desenvolvimento. Elemento este que será difícil de automatizar.

**Fim.**

*Felicidade!*  
*Passei no vestibular*  
*Mas a faculdade, é particular*  
...  
*Livros tão caros*  
*Tantas taxas pra pagar*  
*Meu dinheiro muito raro*  
*Alguém teve que emprestar*  
...  
*Morei no subúrbio, andei de trem atrasado*  
*Do trabalho ia pra aula, sem jantar e bem cansado*  
*Mas lá em casa à meia-noite tinha sempre a me esperar*  
*Um punhado de problemas e criança pra criar*  
...  
*Mas felizmente eu consegui me formar*  
*Mas da minha formatura, não cheguei participar*  
*Faltou dinheiro pra beca e também pro meu anel*  
*Nem o diretor careca entregou o meu papel*  
...  
*E depois de tantos anos*  
*Só decepções, desenganos*  
*Dizem que sou um burguês muito privilegiado*  
*Mas burgueses são vocês*  
*Eu não passo de um pobre-coitado*  
*E quem quiser ser como eu,*  
*Vai ter é que penar um bocado!*  
*Um bom bocado, vai penar um bom bocado*

Martinho da Vila – O Pequeno Burguês

# Referências

AIT DAOUD, R. et al. Combining RFM model and clustering techniques for customer value analysis of a company selling online. Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA, 2016.

AL-AQRABI, H. et al. Cloud BI: Future of business intelligence in the Cloud. **Journal of Computer and System Sciences**, v. 81, n. 1, p. 85-96, 2// 2015. ISSN 0022-0000.

BARNES, J. **Microsoft Azure Essentials Azure Machine Learning**. Pearson Education, 2015. ISBN 9780735698185.

BEZDEK, J. C. **Pattern recognition with fuzzy objective function algorithms**. New York ; London: Plenum, 1981. ISBN 0306406713 : Unpriced.

BHENSADIA, C. K.; KOSTA, Y. P. Discovering Active and Profitable Patterns with Rfm ( Recency , Frequency and Monetary ) Sequential Pattern Mining – a Constraint Based Approach. 2010.

BUYYA, R. et al. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. **Future Generation Computer Systems**, v. 25, n. 6, p. 599-616, 6// 2009. ISSN 0167-739X.

CARVALHO, L. S. **Desenvolvimento de modelos RFM com ajuda de Plataformas de Aprendizado de Máquina na Nuvem**. Proceedings of the XXXVI Iberian Latin American Congress on Computational Methods in Engineering: ABMEC Brazilian Association of Computational Methods in Engineering 2015.

CHANG, E.-C.; HUANG, S.-C.; WU, H.-H. Using K-means method and spectral clustering technique in an outfitter's value analysis. **Quality & Quantity**, v. 44, n. 4, p. 807-815, 2010// 2010. ISSN 1573-7845.

CHEN, Y.-L. et al. Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. **Electronic Commerce Research and Applications**, v. 8, n. 5, p. 241-251, 10// 2009. ISSN 1567-4223.

CHENG, C.-H.; CHEN, Y.-S. Classifying the segmentation of customer value via RFM model and RS theory. **Expert Systems with Applications**, v. 36, n. 3, Part 1, p. 4176-4184, 4// 2009. ISSN 0957-4174.

COUSSEMENT, K.; VAN DEN BOSSCHE, F. A. M.; DE BOCK, K. W. Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees. **Journal of Business Research**, v. 67, n. 1, p. 2751-2758, 1// 2014. ISSN 0148-2963.

DAVIES, D. L.; BOULDIN, D. W. A Cluster Separation Measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. PAMI-1, n. 2, p. 224-227, 1979. ISSN 0162-8828.

DELEN, D.; DEMIRKAN, H. Data, information and analytics as services. **Decision Support Systems**, v. 55, n. 1, p. 359-363, 4// 2013. ISSN 0167-9236.

DEMIRKAN, H.; DELEN, D. Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. **Decision Support Systems**, v. 55, n. 1, p. 412-421, 4// 2013. ISSN 0167-9236.

DEVI, B. N. et al. Design and Implementation of Web Usage Mining Intelligent System in the Field of e-commerce. **Procedia Engineering**, v. 30, p. 20-27, 2012/01/01 2012. ISSN 1877-7058.

DUNN†, J. C. Well-Separated Clusters and Optimal Fuzzy Partitions. **Journal of Cybernetics**, v. 4, n. 1, p. 95-104, 1974/01/01 1974. ISSN 0022-0280.

DURUN, A.; CABER, M. Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. **Tourism Management Perspectives**, v. 18, p. 153-160, 2016.

EBIT. **Webshoppers**. Ebit. São Paulo, p.1 - 45. 2018

GANNON, D. et al. **Science in the cloud: lessons from three years of research projects on microsoft azure**. Proceedings of the 5th ACM workshop on Scientific cloud computing. Vancouver, BC, Canada: ACM: 1-8 p. 2014.

GRZYMALA-BUSSE, J. W. A New Version of the Rule Induction System LERS. **Fundam. Inf.**, v. 31, n. 1, p. 27-39, 1997. ISSN 0169-2968.

HOSSEINI, S. M. S.; MALEKI, A.; GHOLAMIAN, M. R. Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. **Expert Systems with Applications**, v. 37, n. 7, p. 5259-5264, 7// 2010. ISSN 0957-4174.

HRUSCHKA, E. R.; CASTRO, L. N. D.; CAMPELLO, R. J. G. B. Evolutionary algorithms for clustering gene-expression data. Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on, 2004, 1-4 Nov. 2004. p.403-406.

HU, Y.-H.; CHEN, K.; LEE, P.-J. The effect of user-controllable filters on the prediction of online hotel reviews. **Information & Management**, v. 54, n. 6, p. 728-744, 2017/09/01/ 2017. ISSN 0378-7206.

HU, Y.-H.; YEH, T.-W. Discovering valuable frequent patterns based on RFM analysis without customer identification information. **Knowledge-Based Systems**, v. 61, p. 76-88, 5// 2014. ISSN 0950-7051.

HUGHES, A. M. **Strategic Database Marketing**. Chicago, Illinois: Probus Pub Co, 1994.

KAHAN, R. Using database marketing techniques to enhance your one-to-one marketing initiatives. **Journal of Consumer Marketing**, v. 15, n. 5, p. 491-493, 1998/10/01 1998. ISSN 0736-3761.

KASS, G. V. An Exploratory Technique for Investigating Large Quantities of Categorical Data. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, v. 29, n. 2, p. 119-127, 1980. ISSN 00359254, 14679876.

KHAJVAND, M. et al. Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. **Procedia Computer Science**, v. 3, p. 57-63, 2011/01/01 2011. ISSN 1877-0509.

KHODABANDEHLOU, S.; RAHMAN, M. Z. Providing a new approach for segmenting customers based on their purchasing behavior change over time in electronic business. **Journal of Information Technology Management**, v. 9, n. 2, p. 277-300, 2017.

KOHONEN, T. **Self-organizing maps**. Berlin: Springer 1995.

LIU, D.-R.; SHIH, Y.-Y. **Integrating AHP and data mining for product recommendation based on customer lifetime value**. 2005. 387-400.

LIU, F.; ZHAO, S.; LI, Y. How many, how often, and how new? A multivariate profiling of mobile app users. **Journal of Retailing and Consumer Services**, v. 38, p. 71-80, 2017/09/01/ 2017. ISSN 0969-6989.

MACQUEEN, J. B. Some Methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Math, Statistics, and Probability*, 1967, University of California Press. p.281-297.

MARSTON, S. et al. Cloud computing — The business perspective. **Decision Support Systems**, v. 51, n. 1, p. 176-189, 4// 2011. ISSN 0167-9236.

MCCARTY, J. A.; HASTAK, M. Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. **Journal of Business Research**, v. 60, n. 6, p. 656-662, 6// 2007. ISSN 0148-2963.

MICHAEL, J. B.; LISA, M. F.; MICHELLE, S. D. On Average Deviation Indices for Estimating Interrater Agreement. **Organizational Research Methods**, v. 2, n. 1, p. 49-68, 1999/01/01 1999. ISSN 1094-4281.

MIGUÉIS, V. L. et al. Modeling partial customer churn: On the value of first product-category purchase sequences. **Expert Systems with Applications**, v. 39, n. 12, p. 11250-11256, 9/15/ 2012. ISSN 0957-4174.

MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. **Psychometrika**, v. 50, n. 2, p. 159-179, June 01 1985. ISSN 1860-0980.

MIRKIN, B. Choosing the number of clusters. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 1, n. 3, p. 252-260, 2011. ISSN 1942-4795.

MOHAMMADZADEH, M.; HOSEINI, Z. Z.; DERAFSHI, H. A data mining approach for modeling churn behavior via RFM model in specialized clinics Case study: A public

sector hospital in Tehran. **Procedia Computer Science**, v. 120, p. 23-30, 2017/01/01/2017. ISSN 1877-0509.

OLEJNIK, R.; FORTIŞ, T.-F.; TOURSEL, B. Webservices oriented data mining in knowledge architecture. **Future Generation Computer Systems**, v. 25, n. 4, p. 436-443, 4// 2009. ISSN 0167-739X..

OLSON, D. L.; CHAE, B. Direct marketing decision support through predictive customer response modeling. **Decision Support Systems**, v. 54, n. 1, p. 443-451, 12// 2012. ISSN 0167-9236.

PEKER, S.; KOCYIGIT, A.; EREN, P. E. LRFMP model for customer segmentation in the grocery retail industry: a case study. **Marketing Intelligence & Planning**, v. 35, n. 4, p. 544-559, 2017/06/05 2017. ISSN 0263-4503.

RAO, R. V.; SELVAMANI, K. Data Security Challenges and Its Solutions in Cloud Computing. **Procedia Computer Science**, v. 48, p. 204-209, 2015/01/01 2015. ISSN 1877-0509.

SCHROEPFER, A. et al. **Secure benchmarking in the cloud**. Proceedings of the 18th ACM symposium on Access control models and technologies. Amsterdam, The Netherlands: ACM: 197-200 p. 2013.

SHIM, B.; CHOI, K.; SUH, Y. CRM strategies for a small-sized online shopping mall based on association rules and sequential patterns. **Expert Systems with Applications**, v. 39, n. 9, p. 7736-7742, 7// 2012. ISSN 0957-4174.

SINHA, N.; KHREISAT, L. Cloud computing security, data, and performance issues. 2014 23rd Wireless and Optical Communication Conference (WOCC), 2014, 9-10 May 2014. p.1-6.

SONG, M. et al. Statistics-based CRM approach via time series segmenting RFM on large scale data. **Knowledge-Based Systems**, v. 132, p. 21-29, 2017.

SUN, D. et al. Surveying and Analyzing Security, Privacy and Trust Issues in Cloud Computing Environments. **Procedia Engineering**, v. 15, p. 2852-2856, 2011/01/01 2011. ISSN 1877-7058.

TAJADOD, G.; BATTEN, L.; GOVINDA, K. Microsoft and Amazon: A comparison of approaches to cloud security. 4th IEEE International Conference on Cloud Computing Technology and Science Proceedings, 2012, 3-6 Dec. 2012. p.539-544.

WANG, C.-H. Apply robust segmentation to the service industry using kernel induced fuzzy clustering techniques. **Expert Systems with Applications**, v. 37, n. 12, p. 8395-8400, 12// 2010. ISSN 0957-4174.

WEI, J.-T. et al. Customer relationship management in the hairdressing industry: An application of data mining techniques. **Expert Systems with Applications**, v. 40, n. 18, p. 7513-7518, 12/15/ 2013. ISSN 0957-4174.

WEI, J.-T.; LIN, S.-Y.; WU, H.-H. A review of the application of RFM model. **African Journal of Business Management**, v. 4, n. 19, p. 4199-4206, 2010. ISSN 1993-8233

WEI, J. T. et al. Applying Data Mining and RFM Model to Analyze Customers' Values of a Veterinary Hospital. 2016 International Symposium on Computer, Consumer and Control (IS3C), 2016, 4-6 July 2016. p.481-484.

WONG, E.; WEI, Y. **Customer online shopping experience data analytics: Integrated customer segmentation and customised services prediction model**. 2018.

WU, H.-H.; CHANG, E.-C.; LO, C.-F. Applying RFM Model and K-Means Method in Customer Value Analysis of an Outfitter. In: CHOU, S.-Y.; TRAPPEY, A., *et al* (Ed.). **Global Perspective for Competitive Enterprise, Economy and Ecology: Proceedings of the 16th ISPE International Conference on Concurrent Engineering**. London: Springer London, 2009. p.665-672. ISBN 978-1-84882-762-2.

ZHANG, Q.; CHENG, L.; BOUTABA, R. Cloud computing: state-of-the-art and research challenges. **Journal of Internet Services and Applications**, v. 1, n. 1, p. 7-18, 2010// 2010. ISSN 1869-0238.

ZORRILLA, M.; GARCÍA-SAIZ, D. A service oriented architecture to provide data mining services for non-expert data miners. **Decision Support Systems**, v. 55, n. 1, p. 399-411, 4// 2013. ISSN 0167-9236.