NANODEGREE ENGENHEIRO MACHINE LEARNING
UDACITY BRASIL TECNOLOGIA LTDA

APLICAÇÃO DE DECISIONTREECLASSIFIER NA CLASSIFICAÇÃO DE PERFIS DE CLIENTES DE UMA EMPRESA B2C

DAVID MANOEL VIDAL

Sertãozinho
2019

SUMÁRIO

1. INTRODUÇÃO

A história da humanidade é marcada por grandes mudanças, mudanças estas que vão da colheita ao emprego do aço. O homem mudou, a sociedade não muito diferente também. Antigamente era preciso produzir, com o tempo produzir não se tornou o bastante, foi preciso adequar a produção ao uso, depois a qualidade, mais tardar ao custo, ao mercado e, por fim a globalização.

A globalização – por definição: ação de tornar algo global – rompeu diferentes fronteiras: da distância (transportes) à comunicação (internet). O mundo mudou, se tornou ágil, a concorrência acirrada, enquanto diferentes ideias e modelos de negócio surgiram. A forma de produzir e consumir também evoluiu.

Em meio a tantas transições e mudanças de paradigmas as empresas buscam se adequar as novas realidades. O marketing nunca foi tão valorizado como nos tempos de hoje, o que dizer dos departamentos de vendas então? Produzir já não é o bastante, é preciso saber para quem se produz, trabalhar a comunicação e claro segmentar perfis de clientes, bem como direcionar as vendas, de modo a torna-las assertivas.

A empresa que conhece o cliente, posiciona seu produto ao mercado, bem como deve ser, direcionando a melhor comunicação para com ele. Enquanto que a empresa que gerencia as etapas da venda, aplica funis, qualifica *leads* de negócio, mapeia e analisa perfis de cliente; formula e aplica as melhores estratégias para venda, aumentando as chances de negócio, consequentemente de sucesso.

O presente trabalho tem por objetivo aplicar algoritmo DecisionTreeClassifier de Machine Learning na base de dados de uma empresa afim de mapear perfis de clientes. Maiores detalhes serão abordados posteriormente sobre o contexto do negócio.

2. CONTEXTO DO NEGÓCIO

A empresa em questão, localiza-se na região de Ribeirão Preto, seu nome será mantido em anonimato em respeito a mesma. Para preservar detalhes e estratégias adotados em seu negócio.

A empresa atua no setor automobilístico, como uma B2C – *Business to Consumer,* em português “Negócios para o Consumidor” – em outras palavras, é uma empresa que comercializa produtos e presta serviços direto para o consumidor final. Encontra-se consolidada no mercado há mais de 10 anos, possuindo mais de 1.000 clientes na região ativos e 3.000 clientes inativos. Todos os dias formula e transmite mais de 100 propostas comerciais, das quais grande parte não é fechada, enquanto outra parte o é (dados obtidos com base em reunião de negócios).

Atua de forma estratégica para lidar com a alta demanda, conciliar as metas de vendas, sem perder na qualidade ou no atendimento prestado. Velocidade e transparência são alguns dos valores que fazem parte de sua cultura organizacional.

Dentre os departamentos, convêm destacar: a empresa possui uma equipe de vendas de mais de 50 funcionários classificando as propostas comerciais conforme suas categorias de urgência, importância e valor agregado. Grandes propostas passam por mais de um profissional para serem formuladas, avaliadas e aprovadas, enquanto propostas de valores menores ficam a cargo do responsável que as emite. Embora os preços sejam tabelados em sistema, o desconto dado a cada cliente não o é.

O desconto segue tratativas internas, levando em consideração o tipo de produto ou serviço em negociação, a fidelidade do cliente (há quanto tempo faz negócios/se mantêm fiel a empresa) e claro a margem de negociação que a empresa dispõe para lidar com as tratativas. Via de regra quanto maior o valor da proposta comercial, maior tende a ser a margem de desconto a qual podem trabalhar, podendo atingir patamares de até 15% em casos específicos.

Um fator interessante a nível de curiosidade é que via Regra de Negócio, os preços são sempre arredondados para casa decimal zero, uma vez que a empresa não atua com propostas ou descontos que lidem com centavos. Tal premissa busca tornar prático, cômodo e usual o dia a dia do trabalho.

São considerados clientes somente as pessoas que fecham propostas no período de 6 meses. Período o qual fazem balanço geral para classificarem e avaliarem cada carteira de clientes, por equipe de trabalho.

Desta forma avaliam a base de dados e atuam de forma estratégica para averiguarem mudanças no perfil de consumo (hora para mais, hora para menos) e quando algum cliente se torna inativo poderem atuar de forma a relembrá-lo de que existem e estão dispostos a atendê-los da melhor forma possível; assim como realizarem eventuais pesquisas de mercado, satisfação, analisarem dúvidas e/ou sugestões.

A classificação dos clientes segue a ideia a qual originalmente o autor deste trabalho sugeriu anos atrás. A ideia é classificar clientes conforme 3 critérios. São eles: faturamento, desconto e aceitabilidade (apresentados em Anexo A). Cada cliente é classificado em uma letra que vai de “A” até “L”, contemplando ao todo 12 perfis de clientes.

É importante ressaltar novamente que são considerados clientes somente as pessoas que fecham algum tipo de proposta comercial. As demais pessoas listadas na base de dados são denominadas por “Z”; de “Zero”, uma pessoa que não contratou nenhum tipo de produto ou serviço no período.

O departamento de *Data Science* realiza a classificação usando o Excel, utilizando-se das seguintes premissas:

- **Critério 1**: Quão considerável é o gasto da pessoa (considerável = acima ou igual a R$ 7.500, não considerável = abaixo de R$ 7.500). O que resulta em 2 possibilidades ao todo. Veja anexo B.

- **Critério 2**: Se o cliente pede desconto (qualquer desconto negociado) ou, se não pede desconto (quando o desconto é nulo, o cliente não pede nenhum tipo de desconto no momento de fechar uma proposta comercial). O que resulta 2 possibilidades ao todo. Veja anexo C.

- **Critério 3**: A aceitabilidade do cliente para com as propostas emitidas pela empresa no período. Em outras palavras aceitabilidade alta = clientes que fecham mais propostas do que as recusam; aceitabilidade média = clientes que fecham o mesmo número de propostas que recusam, aceitabilidade baixa = clientes que fecham menos propostas do que a quantidade recusada. O que resulta em 3 possibilidades ao todo. Veja anexo D.

Os 3 critérios juntos, quando combinados, resultam nos 12 perfis de clientes possíveis que o negócio pode ter. O anexo E apresenta o cálculo, seguindo o Princípio Fundamental da Contagem para determinação da variação de perfis.

A ideia basicamente consiste em separar o joio do trigo, determinar quem são os clientes que gastam mais, os que gastam menos, quem são os que pedem desconto, os que não pedem e sua aceitabilidade para com as propostas do negócio. Conforme apresentado no Anexo F.

Sua combinação traz informações cruciais no mapeamento de cada cliente, na sua forma de tomarem decisões, agir, possivelmente pensar e cotarem produtos e/ou serviços. Não vem ao caso descriminar o que cada perfil de cliente significa, muito embora seja perfeitamente sugestivo a tomada de conclusões.

A classificação permite a empresa direcionar não apenas a melhor comunicação para cada perfil e ocasião, como também direcionar cada cliente a uma equipe especializada no tipo de atendimento, onde possam reter o máximo de clientes possíveis, abstraírem o máximo de lucro provável - em uma relação ganha a ganha - e claro, tornarem as vendas fluídas.

Em outras palavras a estratégia da empresa consiste em compreender a forma de cada cliente lidar com o negócio para então poder se ajustar a eles, ao passo em que nos modelos tradicionais de negócio as empresas pouco conhecem sobre os clientes e esperam que a sociedade se adapte a elas, o que não para menos resulta em uma baixa conversão de vendas ou retenção de clientes.

A premissa como elucidada anteriormente é desenvolver um modelo de Machine Learning o qual seja capaz de classificar os clientes sem que seja necessário usar o Excel, minimizar eventuais margens para erros e tornar mais fácil o trabalho fluir durante a classificação semestral.

3. CONJUNTO DE DADOS

Quando a análise de clientes ocorre, o departamento de *Data Science* da empresa se encarrega de emitir relatórios a diretoria, onde trazem como apontamento por cliente no período:

- A identificação do cliente;
- O total orçado;
- O total negociado;
- O desconto aplicado;
- O percentual de desconto sobre o total orçado;
- O total de propostas fechadas;
- O total de propostas recusadas;
- A classificação do perfil do cliente;

O conjunto de dados disponibilizado pela empresa possui ao todo 1.744 registros de dados, apresentados sobre os 08 atributos. Como é possível observar o atributo tempo é omitido e a classificação é gerada internamente por eles, uma vez que os dados são transmitidos e processados posteriormente em planilha eletrônica.

Maiores detalhes sobre a arquitetura do conjunto de dados podem ser observados no Anexo G, enquanto as estatísticas gerais do conjunto de dados a nível de curiosidade são apresentadas no Anexo H, eventuais premissas e análises a priori. Enquanto o Anexo I apresenta o *PrintScreen* da planilha gerada pelo departamento de *Data Science* a diretoria.

4. PROPOSTA DE *MACHINE LEARNING* AO PROBLEMA

A classificação de clientes envolve dados discretos como sendo a entrada e, dados rotulados como sendo a saída. Ao todo são 12 classes de perfis de clientes. A base de dados possui mais de mil registros e 3 critérios se demonstram necessários para realizar a classificação.

De posse dessas premissas o problema consiste em Aprendizado Supervisionado, do tipo Classificação, envolvendo múltiplas classes. A proposta é aplicar *DecisionTreeClassifier*, em português Árvore de Decisão Classificatória a questão.

Um esboço simples sobre a arquitetura da solução é apresentado no Anexo J, como forma de resumir o funcionamento em questão.

5. VALIDAÇÃO FINAL DA SOLUÇÃO

A base de dados fornecida pelo cliente em questão passará por transformações durante a etapa de pré-processamento. São estas:

1. Verificação de ocorrência de registros de pessoas que não tenham comprado nada no período, ou seja, pessoas que não tenham sido clientes;
2. Exclusão dos registros encontrados em “1”;
3. Campo Aceitabilidade será criado com base na diferença entre orçamentos fechados e orçamentos recusados;
4. Eventuais campos desnecessários serão eliminados;
5. *O campo Aceitabilidade será convertido em uma categoria, afim de transformar valores negativos, nulos e positivos;*
6. *A técnica One Hot Encoding* será aplicado posteriormente ao campo Aceitabilidade para adequar as categorias a dados binários, tornando melhor o aprendizado ao algoritmo *Machine Learning*;

Após o pré-processamento ser realizado na base de dados, os registros restantes serão divididos em: treinamento (70% dos registros) e testes (30% dos registros). Conforme descriminado no Anexo K.

Não haverá implementação de validação cruzada durante o treinamento, somente aplicação de GridSearch para seleção dos melhores parâmetros de configuração.

Serão utilizados para validar o modelo proposto:

- O tempo total, compreendendo do treinamento a previsão;
- F1 Score durante o treinamento;
- F1 Score durante a previsão;
- A Matriz Confusão para dados da previsão;

A Matriz Confusão e F1 Score serão as métricas determinantes para validação do modelo. Onde a Matriz Confusão servirá de forma visual a questão, devendo apresentar a maior parte dos dados classificados na diagonal da esquerda para a direita, enquanto o F1 Score deverá ser superior a 80% via premissa de análise.

6. BENCHMARK

O conjunto de dados em questão, como dito anteriormente pertence a uma empresa privada. Para tanto será adotado como Benchmark uma DecisionTreeClassifier sem o uso de GridSearch na escolha de hiperparâmetros, tendo como base os parâmetros principais: max_depth, min_samples_leaf e min_samples_split, iguais a 2. Ou seja, o Benchmark consistirá em uma Árvore inicial simples a ser tomada como base sua performance F1 Score durante desenvolvimento do modelo. Bem como a matriz confusão que está gerará durante a classificação dos dados em teste (predição).

É importante ressaltar que o Benchmark utilizará a mesma ideia na divisão dos dados descriminado em tópico anterior, vide Anexo K. Isso permitirá uma comparação verossímil, uma vez que consistirá em dois modelos de igual natureza, a utilizarem o mesmo conjunto de dados, porém configurados diferentes e os dados tendo algumas diferenciações.

O Benchmark utilizará o conjunto de dados limpo, sem registros que atrapalhem a classificação dos perfis de clientes, todavia não terá efeito da técnica *One Hot-Encoding*, de modo a criar extremos entre o modelo final a ser comparado com o modelo inicial (Benchmark). Já o modelo final apresentará o uso das técnicas GridSearch na escolha de melhores parâmetros e técnica *One Hot-Encoding* afim de favorecer o aprendizado máquina, assimilação dos dados.

O modelo Benchmark pode ser visualizado no arquivo benchmark_model.ipynb atingido o seguinte desempenho como referência:

- Menos de 1 segundo, compreendendo do treinamento a previsão;
- 58% de F1 Score no treinamento;
- 57% de F1 Score na previsão;
- Matriz Confusão apresentando 296 registros classificados corretamente e 225 registros incorretamente, de um total de 521 registros utilizados. O resultado pode ser conferido também no Anexo L e M.

7. DESENVOLVIMENTO

A base de dados passou por processo de limpeza (eliminação de registros inválidos) e processo de normalização, afim de adequar os dados para o aprendizado máquina. Posteriormente o modelo DecisionTreeClassifier foi implementada.

Dentre os parâmetros foram especificados a seguinte grade para o algoritmo GridSearch na escolha de melhores parâmetros:

- max_depth : [2, 3, 4, 5, 6, 7];
- min_samples_leaf : [2, 3, 4, 5, 6, 7];
- min_samples_split : [2, 3, 4, 5, 6, 7];

A melhor seleção de parâmetros demonstrou ser:

- max_depth : 4;
- min_samples_leaf : 2;
- min_samples_split : 2;

O resultado da implementação do algoritmo pode ser visualizado no arquivo DecisionTreeClassifier.ipynb atingido o seguinte desempenho como referência:

- 6 segundos, compreendendo do treinamento a previsão;
- 100% de F1 Score no treinamento;
- 100% de F1 Score na previsão;
- Matriz Confusão apresentando 520 registros classificados corretamente e somente 1 registro incorretamente (pertencente a classificação de clientes “F”, de um total de 521 registros utilizados. O resultado pode ser conferido também no Anexo N e O.

O anexo P apresenta a estrutura interna da DecisionTreeClassifier, algo interessante pode ser observado quando comparamos o anexo com a premissa do Anexo F mapeamento de perfis de cliente. Embora haja certa similaridade na estrutura a DecisionTreeClassifier se estrutura de forma diferente da que nós humanos desenharíamos/desenhamos, é um ponto interessante a se observar para fins de curiosidade acerca do aprendizado máquina.

8. CONCLUSÃO

O algoritmo DecisionTreeClassifier atendeu perfeitamente a situação. Demonstrando ser capaz de mapear os critérios implícitos na base de dados, muito embora conhecidos como regras e no modelo de negócio.

A solução apresentou uma performance aceitável, sendo ágil no treinamento e na previsão, bem como assertivo na determinação dos perfis de clientes. Algo interessante a se observar é que apesar de o modelo ter apresentado 100% de F1 Score no treinamento levantando suspeitas para uma possível ocorrência de *Overfitting* - em português Superajustado ou Sobreajustado - o modelo demonstrou uma pequena margem de erro na previsão.

No futuro caso a empresa adote novas regras para classificar clientes a DecisionTreeClassifier pode ser capaz de prevê-las, ou até mesmo ser novamente treinada para se adequar a mudanças no paradigma do negócio. O fato é que hoje o algoritmo atende perfeitamente bem a implementação da solução e sana o contexto do negócio:

- Não apresenta margens consideráveis para erros (sejam computacionais e/ou principalmente humanas);
- Torna fácil o trabalho fluir futuramente entre departamentos;
- Permite automatizar o processo de classificação;
- Não apresenta letargia durante o processo de classificação (atrasos de tempo ou perda de performance);

REFERÊNCIAS BIBLIOGRÁFICAS

SKLEARN.TREE.DECISIONTREECLASSIFIER. [*S. l.*], 2 jul. 2019. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html. Acesso em: 2 jul. 2019.

SKLEARN.METRICS.F1_SCORE. [*S. l.*], 2 jul. 2019. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html. Acesso em: 2 jul. 2019.

SKLEARN.METRICS.CONFUSION_MATRIX. [*S. l.*], 2 jul. 2019. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html. Acesso em: 2 jul. 2019.

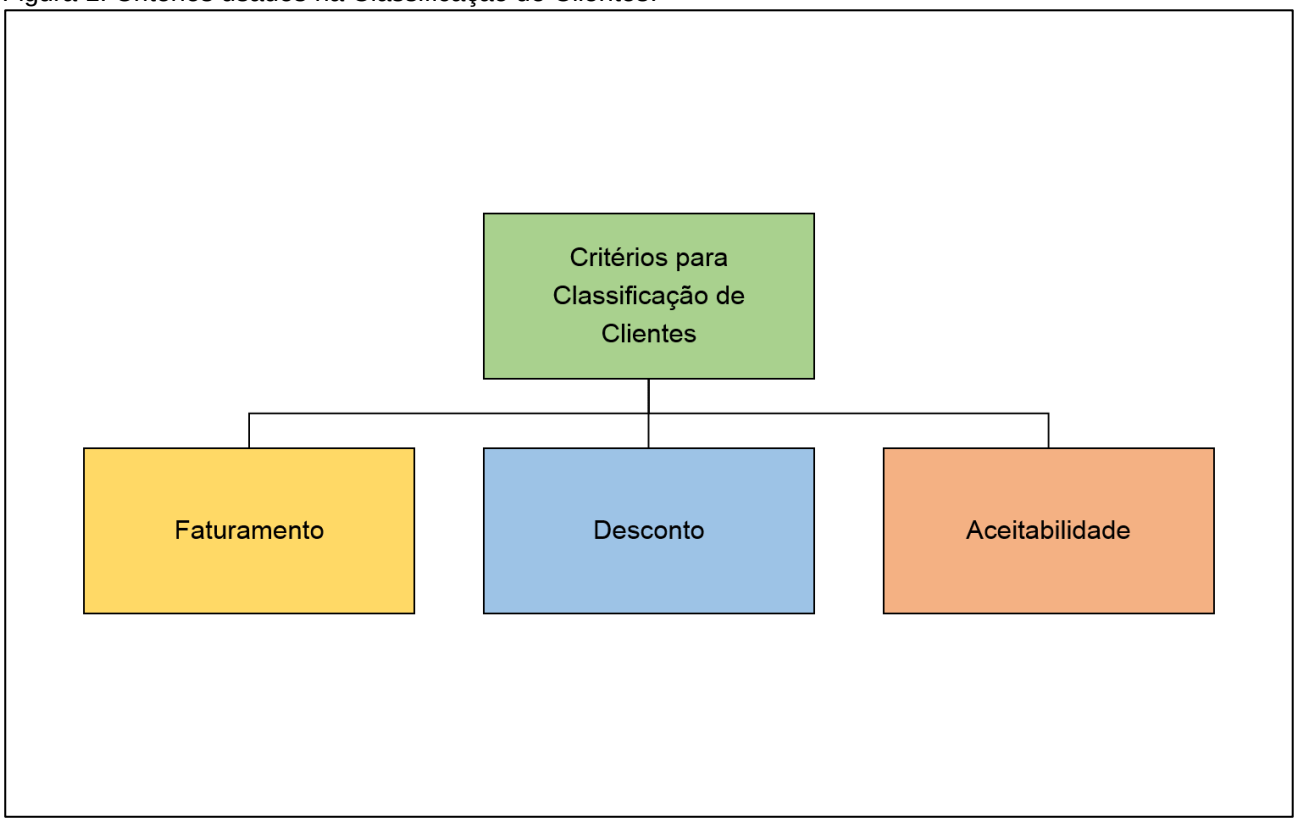
SKLEARN.METRICS.CLASSIFICATION_REPORT. [*S. l.*], 2 jul. 2019. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html. Acesso em: 2 jul. 2019.

SKLEARN.TREE.EXPORT_GRAPHVIZ. [*S. l.*], 2 jul. 2019. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.tree.export_graphviz.html. Acesso em: 2 jul. 2019.

MODEL PERSISTENCE. [*S. l.*], 2 jul. 2019. Disponível em: https://scikit-learn.org/stable/modules/model_persistence.html. Acesso em: 2 jul. 2019.

Anexo A - Critérios usados na Classificação de Clientes

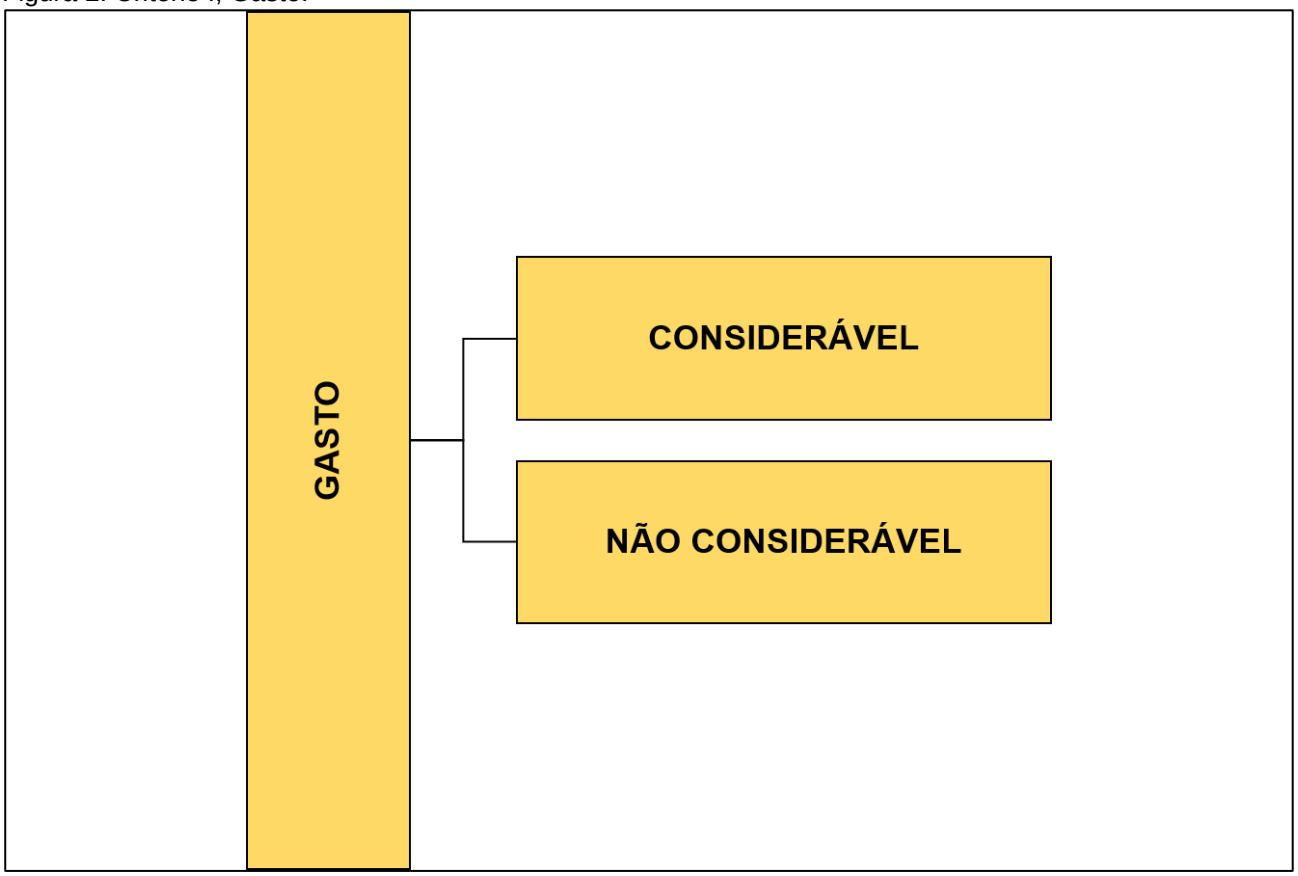
Figura 1: Critérios usados na Classificação de Clientes.



Fonte: elaborado pelo autor.

Anexo B - Critério I: Gasto

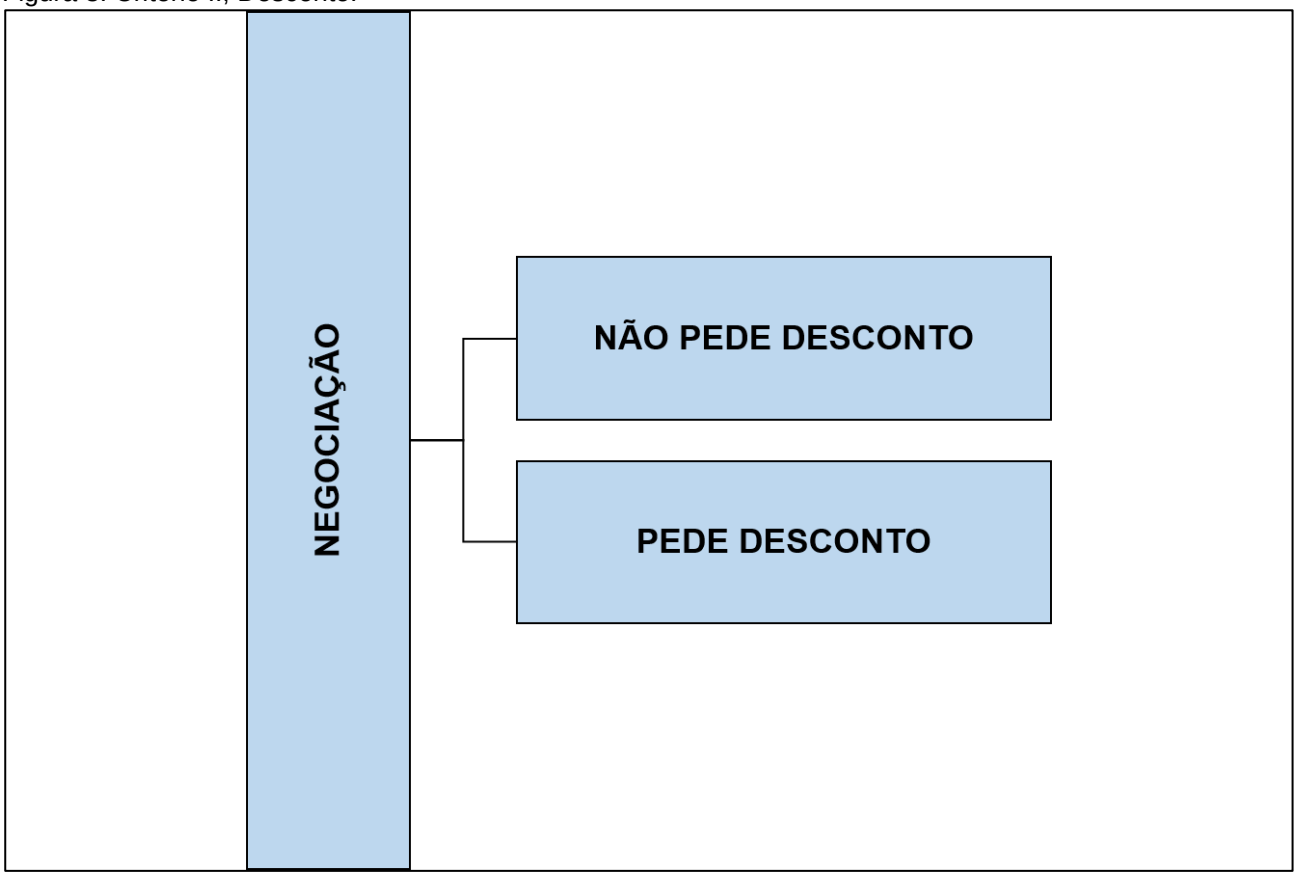
Figura 2: Critério I, Gasto.



Fonte: elaborado pelo autor.

Anexo C - Critério II: Desconto

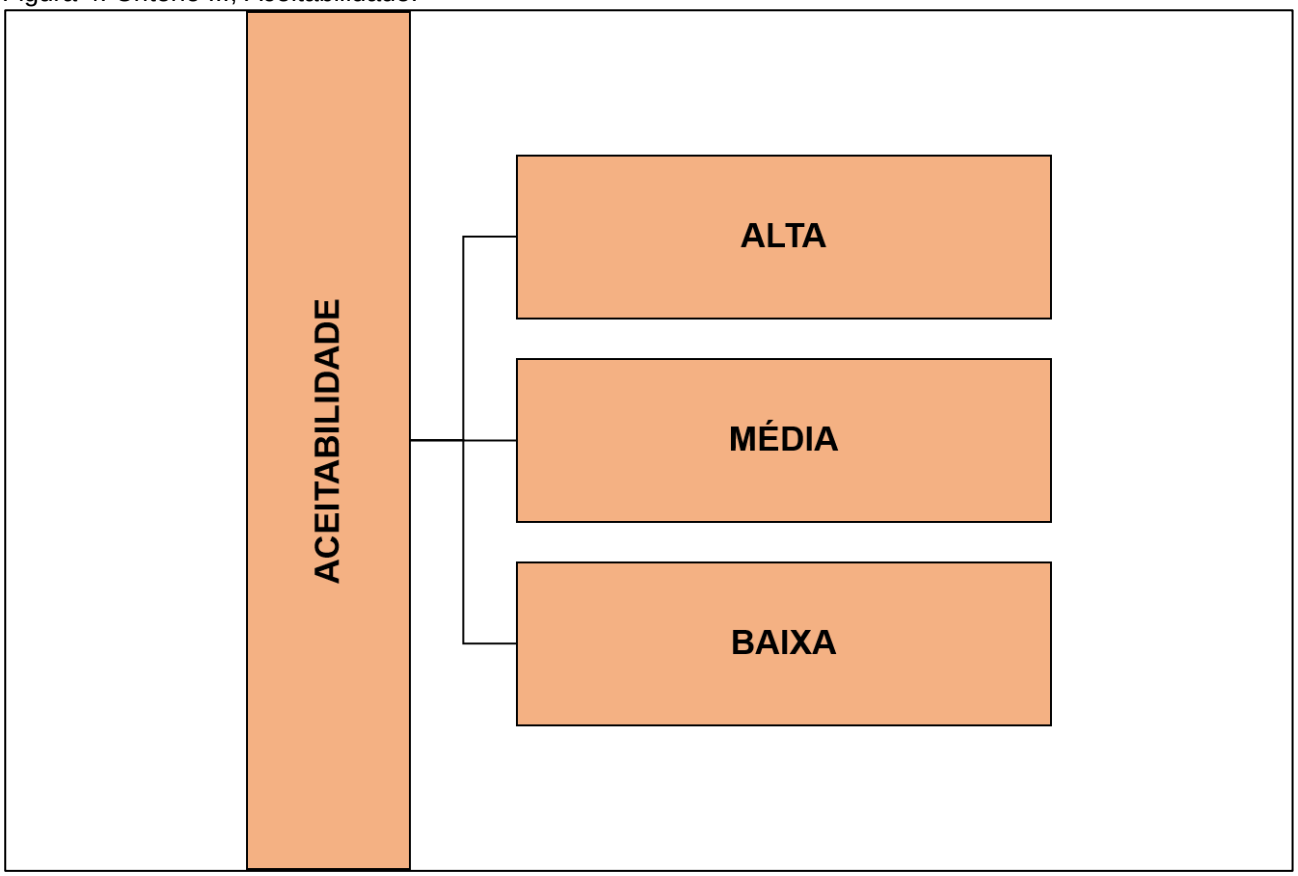
Figura 3: Critério II, Desconto.



Fonte: elaborado pelo autor.

Anexo D - Critério III: Aceitabilidade

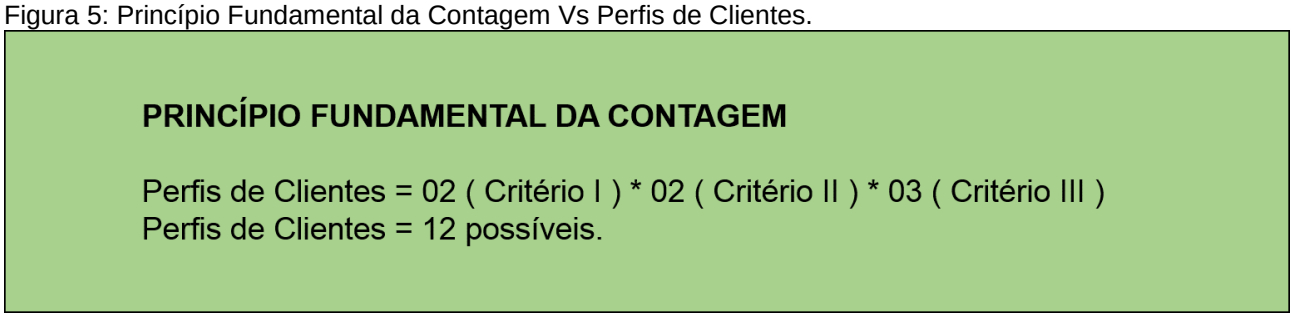
Figura 4: Critério III, Aceitabilidade.



Fonte: elaborado pelo autor.

Anexo E - Princípio Fundamental da Contagem Vs Perfis de Clientes

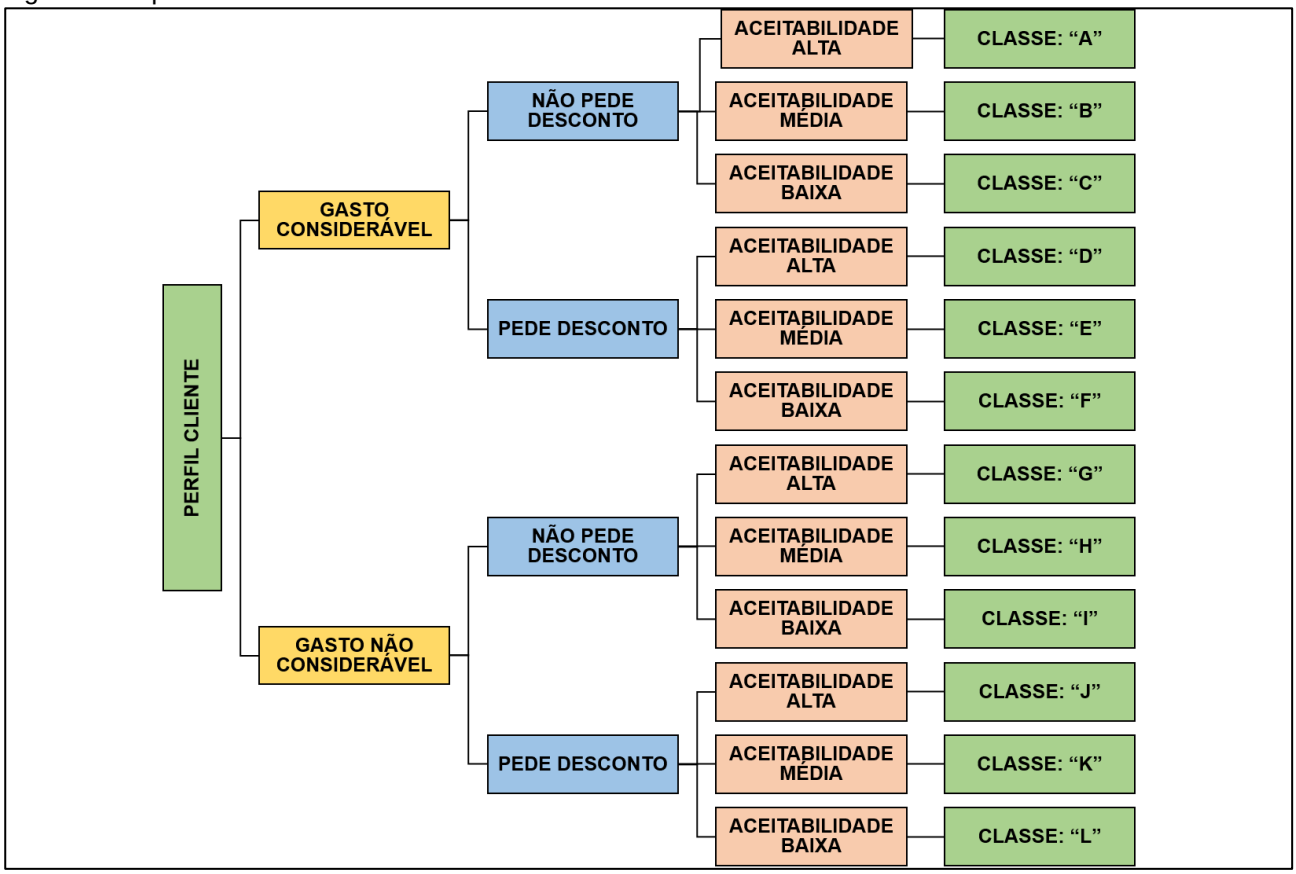
Figura 5: Princípio Fundamental da Contagem Vs Perfis de Clientes.



Fonte: elaborado pelo autor.

Anexo F - Mapeamento de Perfis de Cliente

Figura 6: Mapeamento de Perfis de Cliente.



Fonte: elaborado pelo autor.

Anexo G - Análise da Base De Dados

Quadro 1: Análise da Base de Dados.

CAMPO	TIPO	O QUE É?	SERÁ ÚTIL?	MACHINE LEARNING
id_cliente	inteiro	Código de cada cliente.	Não. Eliminar!	-
cla_cliente	string, categórico	Classificação que cada cliente recebeu.	Sim.	(y) Saída.
tot_orcamento	inteiro	Total orçado no período.	Não. Eliminar!	-
fat_cliente	inteiro	Total negociado no período.	Sim.	(X) Entrada.
des_cliente	inteiro	Desconto total negociado no período.	Sim.	(X) Entrada.
por_desconto	inteiro	Porcentagem de desconto sobre o total.	Não. Eliminar!	-
fec_orcamento	inteiro	Total de orçamentos fechados no período.	Sim. Transformar.	(X) Entrada.
rec_orcamento	inteiro	Total de orçamentos recusados no período.	Sim. Transformar.	(X) Entrada.

Fonte: elaborado pelo autor.

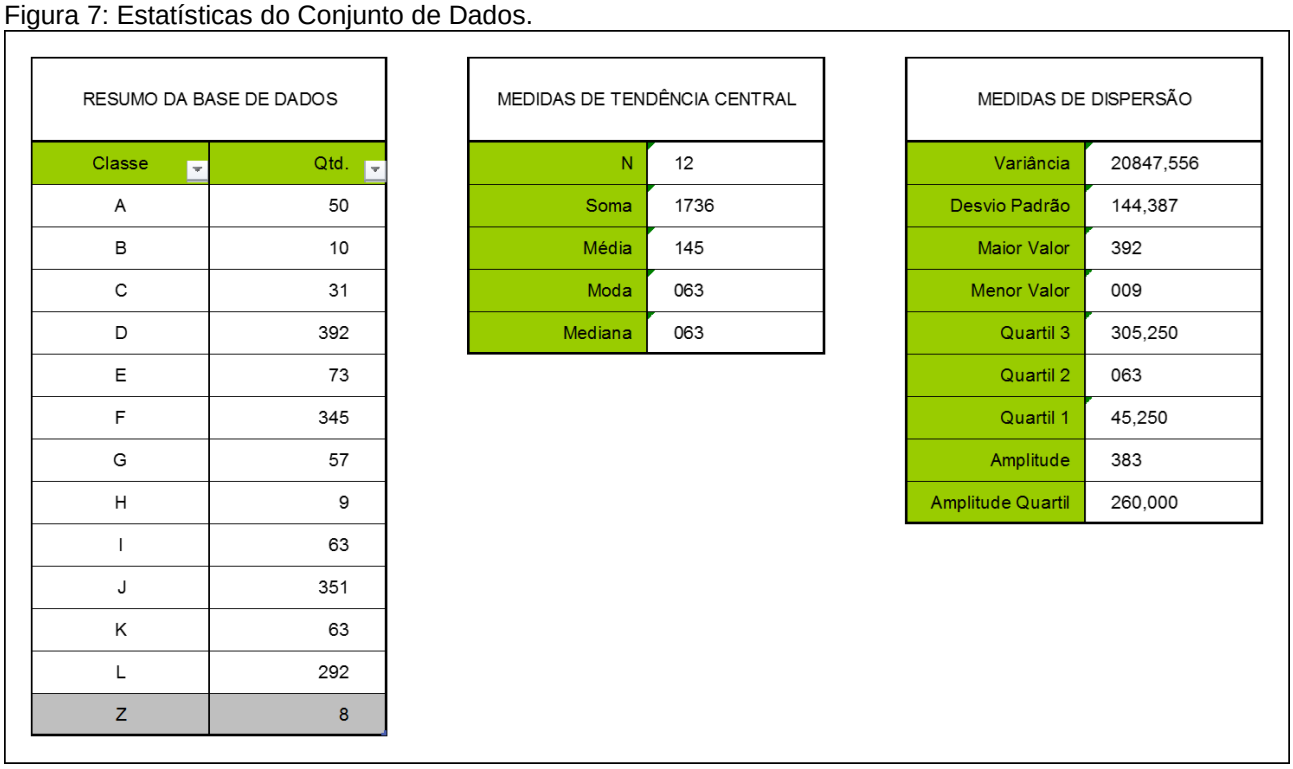
A aceitabilidade do cliente é determinada em função da categoria alta, média ou baixa, conforme dito anteriormente. Alta é quando o cliente fecha mais orçamentos do que recusa, média é quando o cliente fecha o mesmo número de orçamentos que recusa e, baixa é quando o cliente fecha menos orçamentos do que recusa.

Para determinar a aceitabilidade do cliente no período será subtraído do número total de orçamentos fechados o número total de orçamentos recusados, posteriormente uma classificação será dada aos valores (alta > 0; média = 0 e baixa < 0), por fim será aplicado a técnica *One Hot Encoding* afim de transformar as categorias da aceitabilidade em valores binários para melhor assimilação do aprendizado máquina.

O total de orçamentos transmitidos ao cliente no período corresponde a soma entre o número de orçamentos fechados e o número total de orçamentos recusados. Muito embora essa informação não seja tão relevante aos olhos da diretoria. A preocupação da mesma consiste em apenas classificar os clientes.

Anexo H - Análise do Conjunto de Dados

Figura 7: Estatísticas do Conjunto de Dados.



RESUMO DA BASE DE DADOS

Classe	Qtd.
A	50
B	10
C	31
D	392
E	73
F	345
G	57
H	9
I	63
J	351
K	63
L	292
Z	8

MEDIDAS DE TENDÊNCIA CENTRAL

N	12
Soma	1736
Média	145
Moda	063
Mediana	063

MEDIDAS DE DISPERSÃO

Variância	20847,556
Desvio Padrão	144,387
Maior Valor	392
Menor Valor	009
Quartil 3	305,250
Quartil 2	063
Quartil 1	45,250
Amplitude	383
Amplitude Quartil	260,000

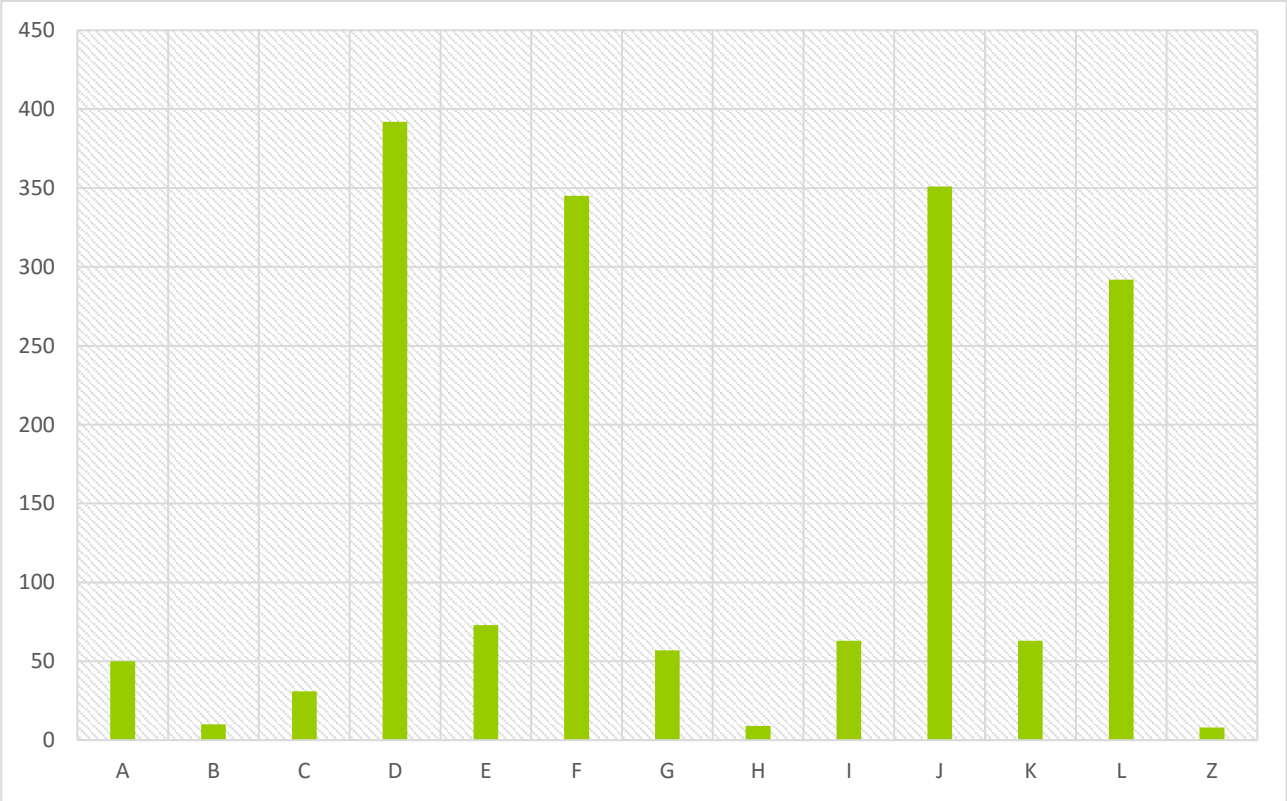
Fonte: elaborado pelo autor.

Como é possível observar no resumo da base de dados, as classes mais predominantes são: D, F, J e L, uma vez que apresentam frequência superior à média. Enquanto as demais classes de clientes apresentam menor ocorrência na base de dados.

As análises excluíram a categoria “Z” uma vez que representa um resíduo a ser eliminado na base de dados durante a etapa de pré-processamento. Pessoas que não realizaram nenhum tipo de aquisição de produto ou serviço no período de 6 meses não são clientes e para tanto devem ser eliminadas das análises / classificação via regra de negócio da empresa.

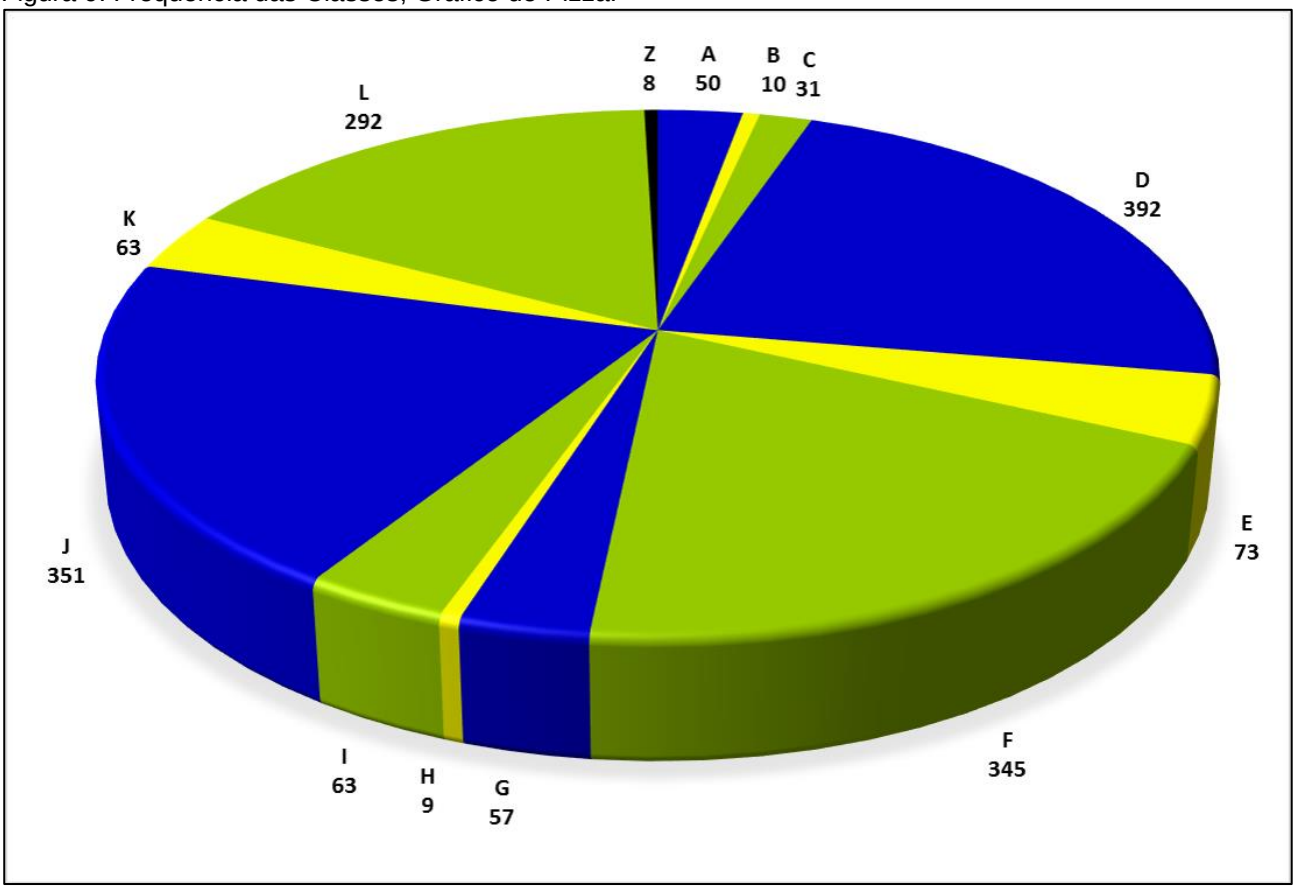
De posse dessas premissas o conjunto de dados é desbalanceado, podendo ser observado também na Figura 8 e 9.

Figura 8: Frequência das Classes, Gráfico de Barras.



Fonte: elaborado pelo autor.

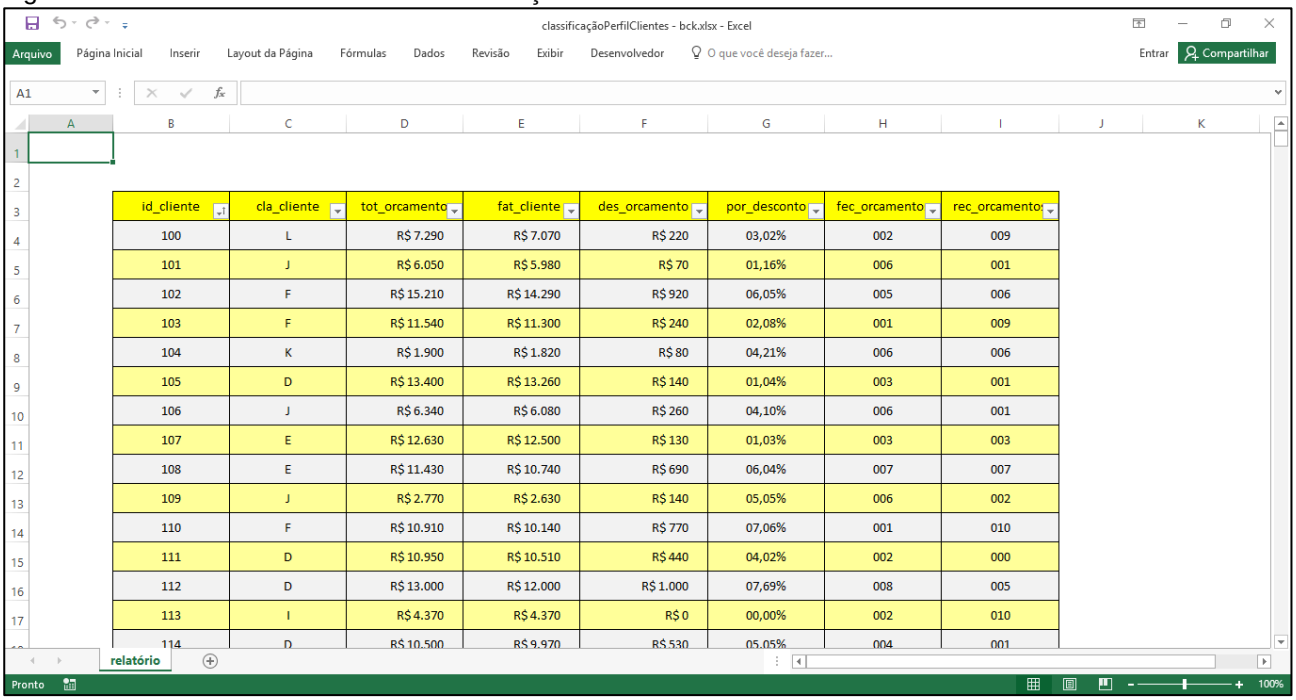
Figura 9: Frequência das Classes, Gráfico de Pizza.



Fonte: elaborado pelo autor.

Anexo I - *PrintScreen* Planilha de Classificação de Clientes

Figura 10: *PrintScreen* Planilha de Classificação de Clientes.



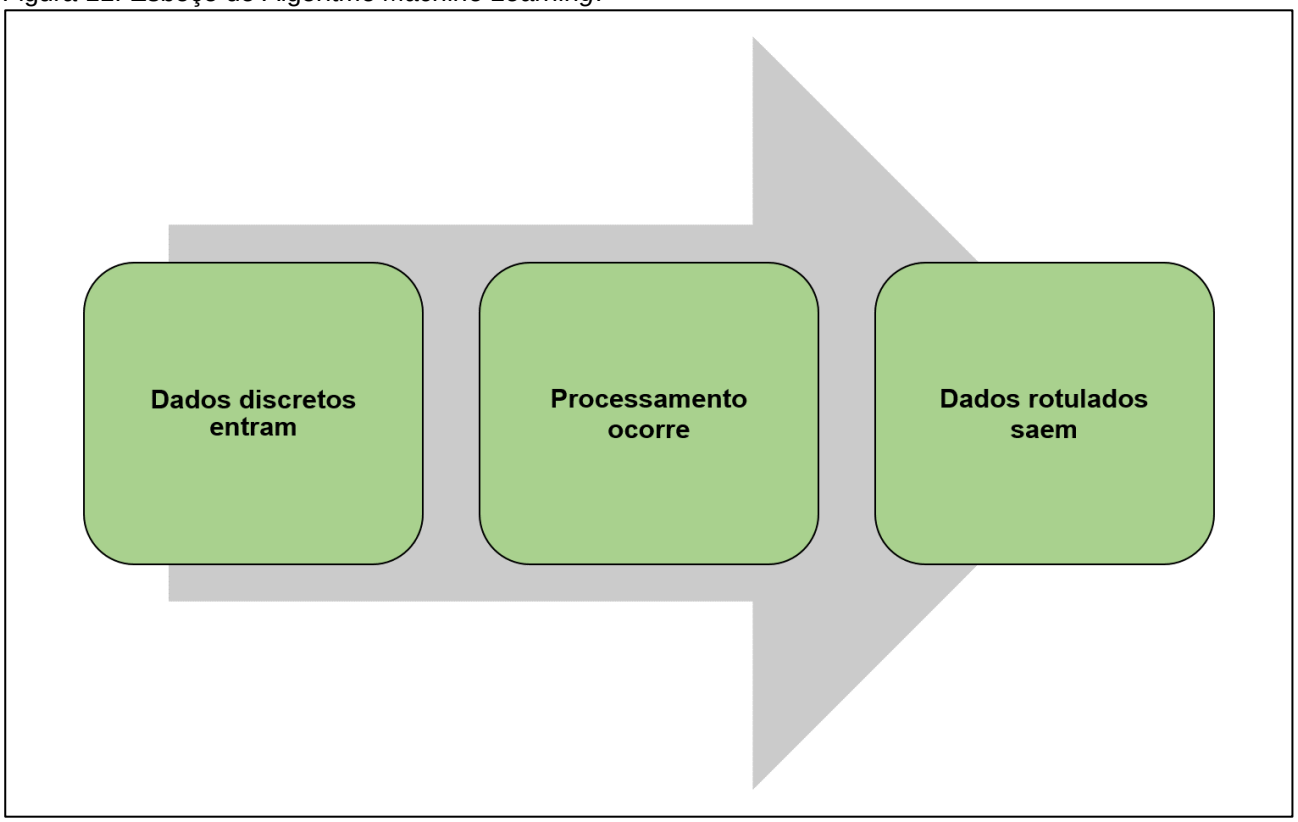
classificaçãoPerfilClientes - bck.xlsx - Excel

id_cliente	cla_cliente	tot_orcamento	fat_cliente	des_orcamento	por_desconto	fec_orcamento	rec_orcamento
100	L	R$ 7.290	R$ 7.070	R$ 220	03,02%	002	009
101	J	R$ 6.050	R$ 5.980	R$ 70	01,16%	006	001
102	F	R$ 15.210	R$ 14.290	R$ 920	06,05%	005	006
103	F	R$ 11.540	R$ 11.300	R$ 240	02,08%	001	009
104	K	R$ 1.900	R$ 1.820	R$ 80	04,21%	006	006
105	D	R$ 13.400	R$ 13.260	R$ 140	01,04%	003	001
106	J	R$ 6.340	R$ 6.080	R$ 260	04,10%	006	001
107	E	R$ 12.630	R$ 12.500	R$ 130	01,03%	003	003
108	E	R$ 11.430	R$ 10.740	R$ 690	06,04%	007	007
109	J	R$ 2.770	R$ 2.630	R$ 140	05,05%	006	002
110	F	R$ 10.910	R$ 10.140	R$ 770	07,06%	001	010
111	D	R$ 10.950	R$ 10.510	R$ 440	04,02%	002	000
112	D	R$ 13.000	R$ 12.000	R$ 1.000	07,69%	008	005
113	I	R$ 4.370	R$ 4.370	R$ 0	00,00%	002	010
114	D	R$ 10.500	R$ 9.970	R$ 530	05,05%	004	001

Fonte: elaborado pelo autor.

Anexo J - Esboço do Algoritmo *Machine Learning*

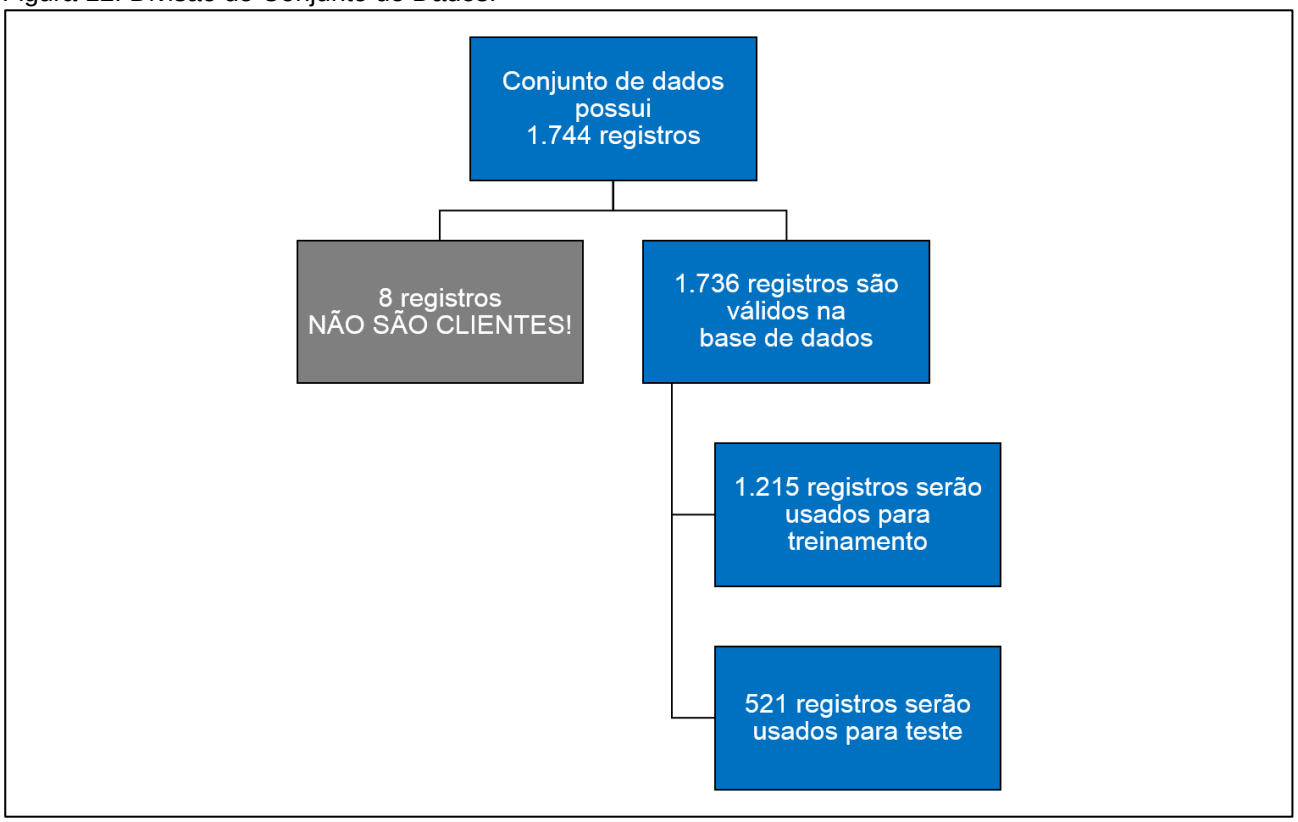
Figura 11: Esboço do Algoritmo *Machine Learning*.



Fonte: elaborado pelo autor.

Anexo K - Divisão do Conjunto de Dados

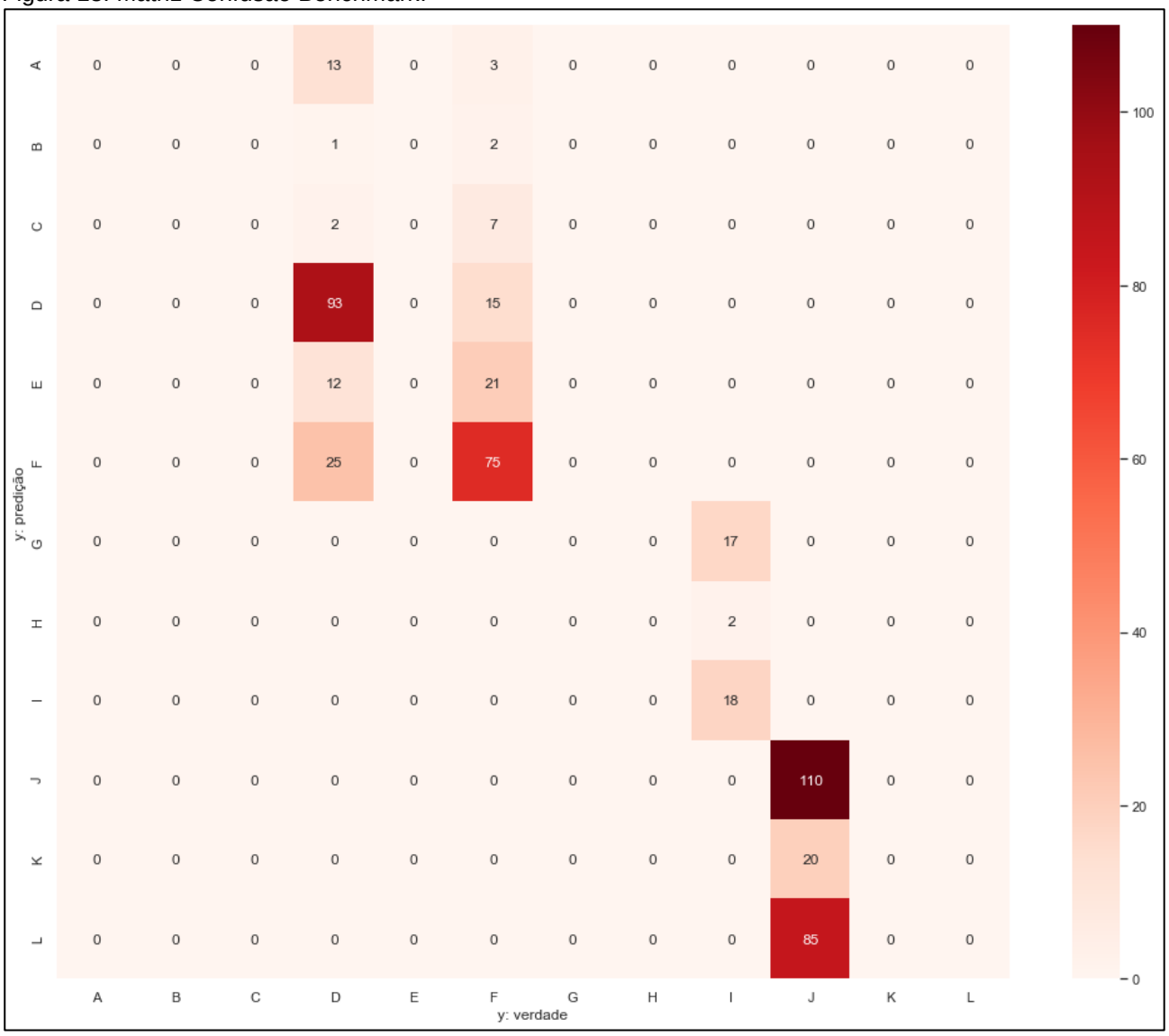
Figura 12: Divisão do Conjunto de Dados.



Fonte: elaborado pelo autor.

Anexo L - Matriz Confusão Benchmark

Figura 13: Matriz Confusão Benchmark.

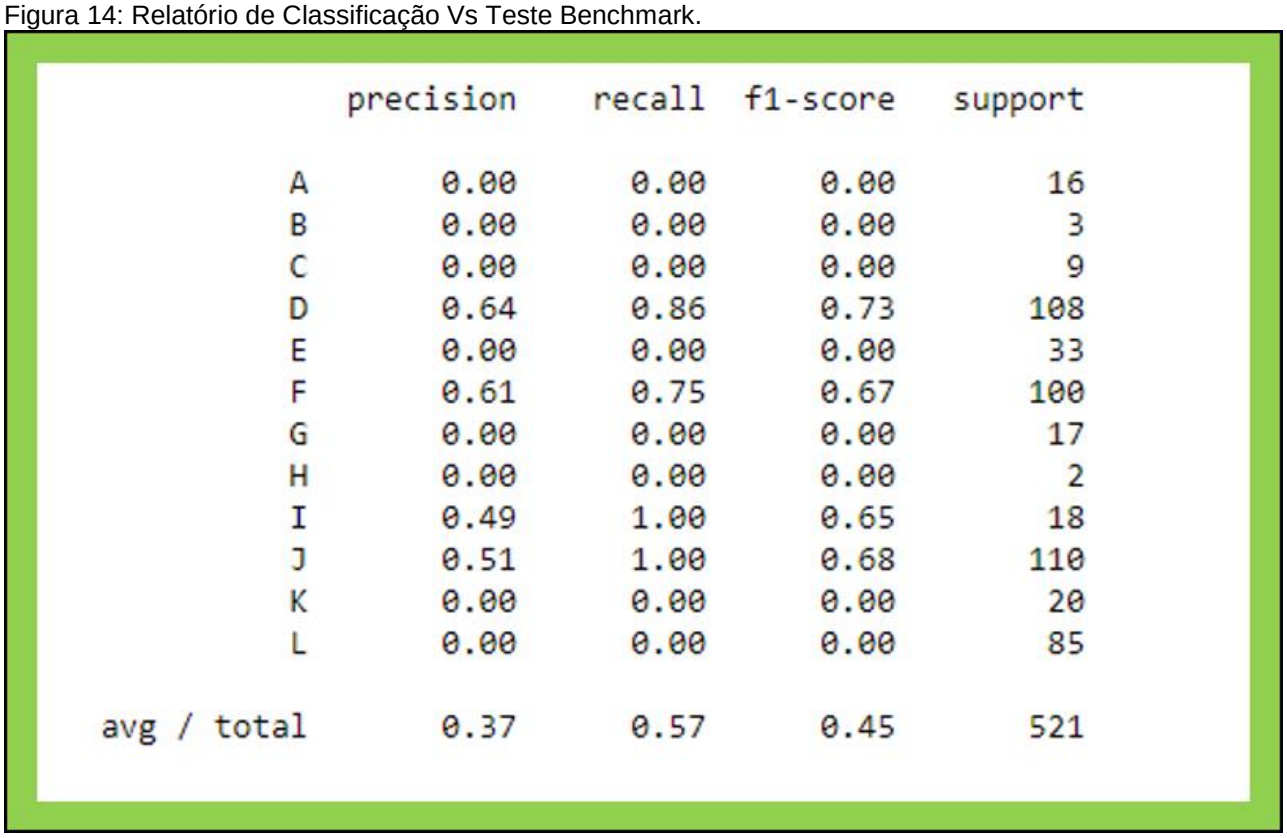


Fonte: elaborado pelo autor.

Conforme apresentado na Matriz Confusão, o modelo Benchmark apresenta grande parte dos registros classificados errados, fora da diagonal principal da esquerda para a direita. Sendo 296 registros classificados corretamente e 225 registros incorretamente, de um total de 521 registros utilizados, ou seja 43,19% de erro na classificação.

Anexo M - Relatório de Classificação Vs Teste Benchmark

Figura 14: Relatório de Classificação Vs Teste Benchmark.



	precision	recall	f1-score	support
A	0.00	0.00	0.00	16
B	0.00	0.00	0.00	3
C	0.00	0.00	0.00	9
D	0.64	0.86	0.73	108
E	0.00	0.00	0.00	33
F	0.61	0.75	0.67	100
G	0.00	0.00	0.00	17
H	0.00	0.00	0.00	2
I	0.49	1.00	0.65	18
J	0.51	1.00	0.68	110
K	0.00	0.00	0.00	20
L	0.00	0.00	0.00	85
avg / total	0.37	0.57	0.45	521

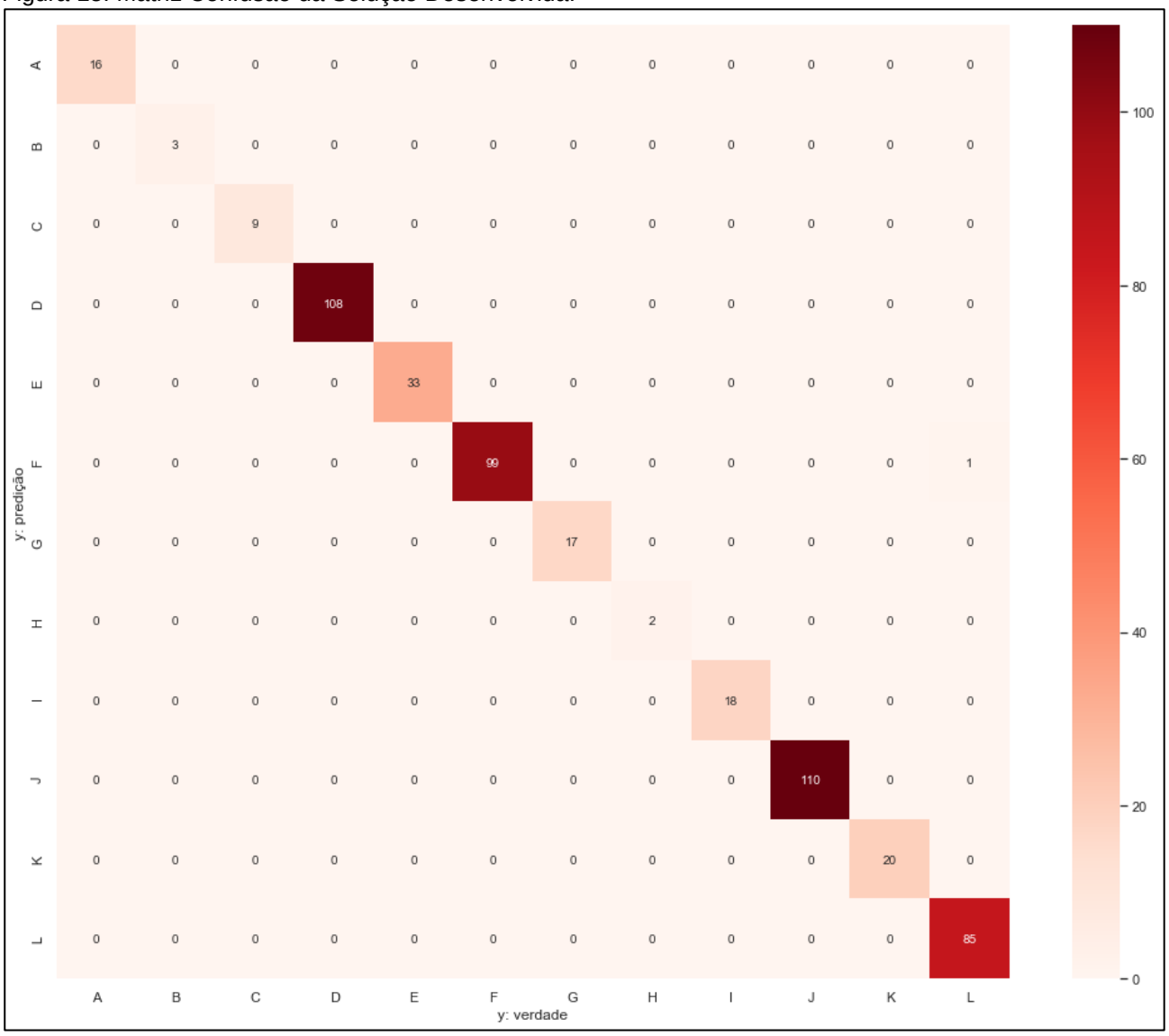
Fonte: elaborado pelo autor.

O relatório de classificação apresenta as 12 classes de perfis de clientes, presentes nos 521 registros destinados a teste (30% da base de dados). Algo extremamente válido, uma vez que a base de dados se encontra desbalanceada. Do relatório pode-se observar:

- baixa precisão (em inglês precision) média na classificação;
- média sensibilidade (em inglês recall) na classificação;
- baixo F1 Score (média ponderada entre a precisão e a sensibilidade) na classificação;

Anexo N - Matriz Confusão da Solução Desenvolvida

Figura 15: Matriz Confusão da Solução Desenvolvida.

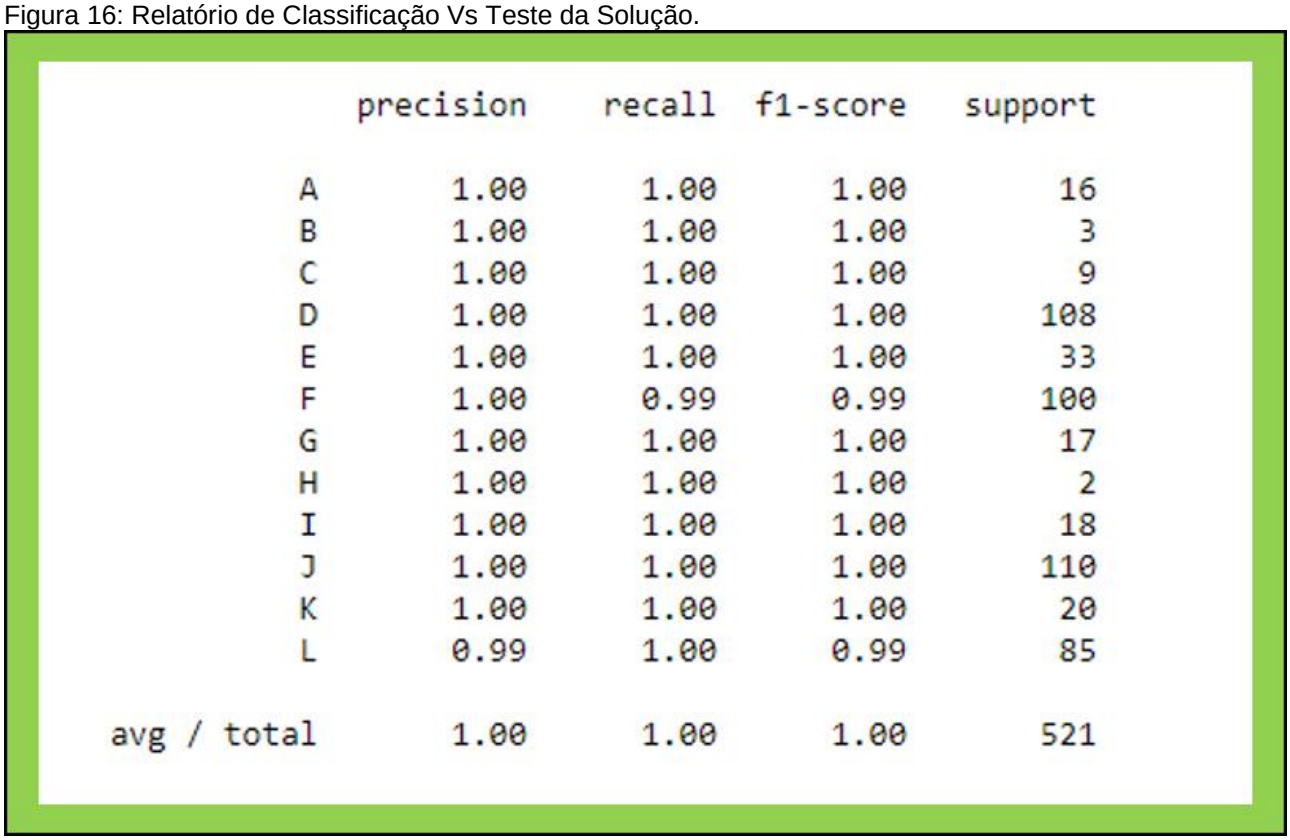


Fonte: elaborado pelo autor.

Conforme apresentado na Matriz Confusão, a solução desenvolvida apresenta a maior parte dos registros classificados corretamente, na diagonal principal da esquerda para a direita. Sendo 520 registros classificados corretamente e somente 1 registro incorretamente, de um total de 521 registros utilizados, ou seja 0,19% de erro na classificação.

Anexo O - Relatório de Classificação Vs Teste da Solução

Figura 16: Relatório de Classificação Vs Teste da Solução.



	precision	recall	f1-score	support
A	1.00	1.00	1.00	16
B	1.00	1.00	1.00	3
C	1.00	1.00	1.00	9
D	1.00	1.00	1.00	108
E	1.00	1.00	1.00	33
F	1.00	0.99	0.99	100
G	1.00	1.00	1.00	17
H	1.00	1.00	1.00	2
I	1.00	1.00	1.00	18
J	1.00	1.00	1.00	110
K	1.00	1.00	1.00	20
L	0.99	1.00	0.99	85
avg / total	1.00	1.00	1.00	521

Fonte: elaborado pelo autor.

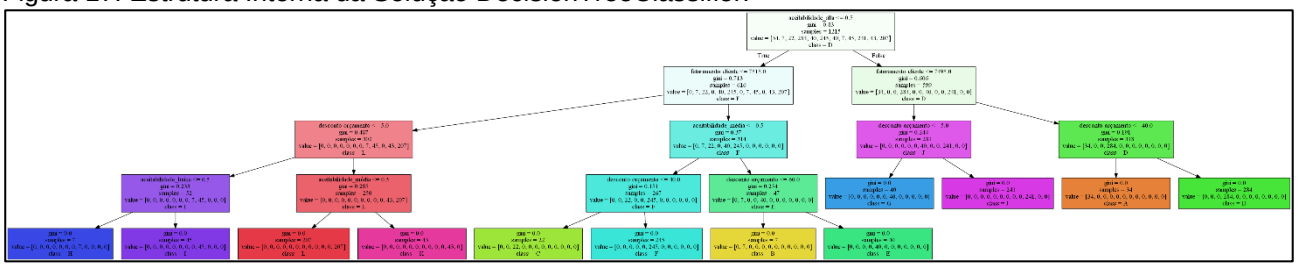
O relatório de classificação apresenta as 12 classes de perfis de clientes, presentes nos 521 registros destinados a teste (30% da base de dados). Algo extremamente válido, uma vez que a base de dados se encontra desbalanceada. Do relatório pode-se observar:

- alta precisão (em inglês precision) média na classificação;
- alta sensibilidade (em inglês recall) na classificação;
- alto F1 Score (média ponderada entre a precisão e a sensibilidade) na classificação;

Lembrando que somente 1 registro foi classificado errado, representando 0,19% de margem de erro, por isso as métricas médias resultaram em 100% aproximadamente.

Anexo P - Estrutura Interna da Solução DecisionTreeClassifier

Figura 17: Estrutura Interna da Solução DecisionTreeClassifier.



Fonte: elaborado pelo autor.