Relatório Desafio Geofusion

Gustavo Caltabiano Eichler

6 de dezembro de 2021

1 Introdução

Este documento retrata o processo de desenvolvimento da solução para o Desafio Técnico Cientista de Dados, proposto pela Geofusion. O desafio tem como objetivo montar uma análise dos bairros de São Paulo, estimando o faturamento e o potencial que uma empresa poderia atingir baseando-se em dados previamente conhecidos.

Foi fornecida uma base de dados da cidade do Rio de Janeiro com informações referente aos bairros, população, população por faixa etária, número de domicílios classificados por renda, a renda média de cada bairro, o faturamento de uma loja localizada no bairro e o potencial da loja.

De forma análoga, uma base de dados foi fornecida para a cidade de São Paulo. Como a empresa não possui os dados de faturamento e potencial para cada bairro de São Paulo, faz-se necessário o desenvolvimento de uma solução que possa estimar e classificar cada bairro presente nesta base de dados, possibilitando uma tomada de decisão, mais segura e confiável para a instalação de lojas.

O desenvolvimento da solução foi realizado utilizando o *Python* como linguagem de programação e as bibliotecas:

- Pandas;
- Pickle;
- Numpy;
- Seaborn;
- Matplotlib;
- Plotly;
- Os.

O restante do documento está organizado da seguinte maneira: A seção 2 apresenta como foi feita uma análise exploratório dos dados para melhor entender cada variável e buscar modelos que possuem melhor desempenho dada as distribuições e relações entre as variáveis. A seção 3 descreve os procedimentos realizados para o pré processamento das bases de dados. As seções 4 e 5 detalham os procedimentos para estimar o faturamento e classificar o potencial de cada loja, respectivamente. Os resultados são apresentados na seção 6 e a seção 7 conclui este relatório.

Os arquivos e códigos estão disponíveis no link: Github.

2 Análise Exploratória dos dados

A análise exploratória de dados é de suma importância para a concepção de soluções e modelos baseados em dados. A exploração dos dados permite uma visualização mais precisa das relações entre as variáveis de um problema e como elas estão distribuídas. De forma mais teórica, a análise exploratória permite chegar a questões de pesquisas mais bem articuladas, formalização de hipóteses mais robustas e descoberta de relações que a priori não poderiam ser identificadas [JPW17].

Como um dos problemas do desafio proposto é estimar o faturamento das lojas, a figura 1 demonstra como o faturamento das lojas do Rio de Janeiro é distribuído. É possível ver que o faturamento se aproxima de uma distribuição normal. Para uma melhor adequação dos modelos, todas as variáveis foram normalizadas para possuírem média 0 e desvio padrão igual a 1.

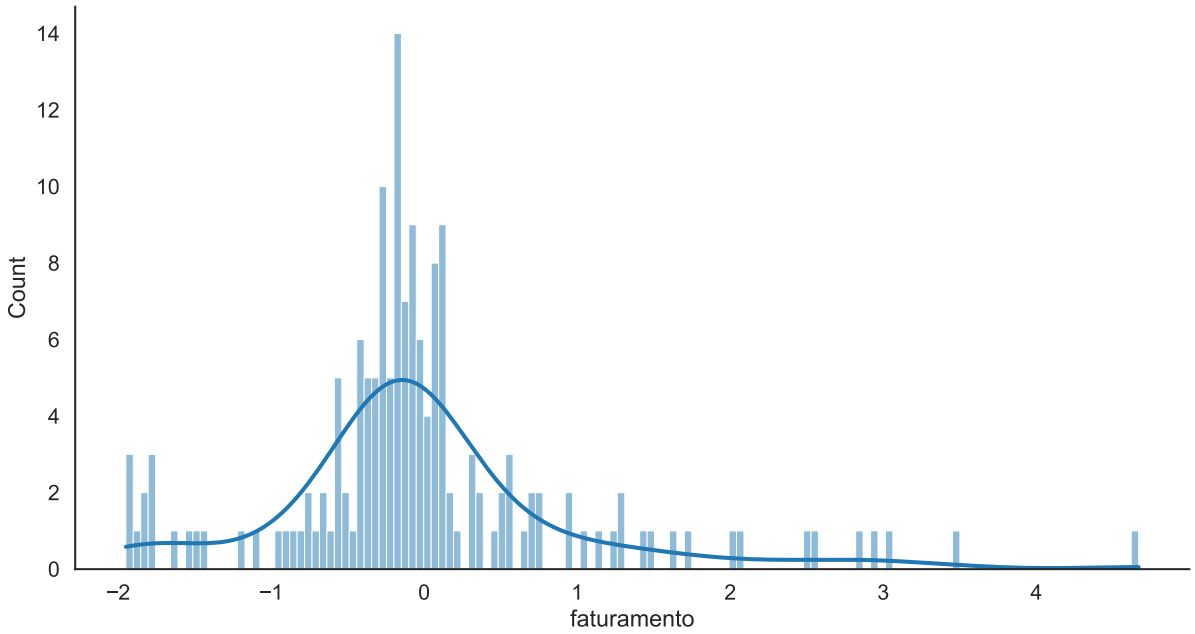


Figura 1: Distribuição do faturamento

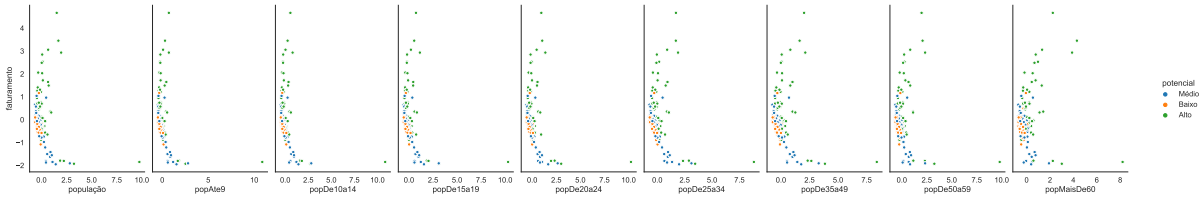


Figura 2: Relação entre o faturamento e as variáveis de população

Analisando as relações entre as variáveis, a figura 2 apresenta a relação entre os potenciais de cada loja, representado pelas cores verde, azul e laranja para potencial Alto, Médio e Baixo, respectivamente, com as informações de faturamento, no eixo Y e as informações sociodemográficas de população e população por faixa etária. É possível inferir que não há uma separação explicita entre os potenciais de acordo com a faixa etária, porém é possível ver que existe uma tendência entre o potencial e o faturamento. Os pontos que representam o potencial Alto costumam se isolar dos demais potenciais conforme o faturamento aumenta.

Buscando relações com as informações sobre as classes de domicílios, a figura 3 demonstra a relação da quantidade de domicílios com o faturamento e o potencial de cada localização. Diferentemente do que foi mostrado anteriormente, na figura 2, existe uma relação clara entre a quantidade de domicílios A1 e A2, o faturamento, e o potencial de cada loja. O faturamento aumenta conforme a quantidade de domicílios A1, A2 e B1 presentes nos bairros, assim como o potencial passa de baixo para médio e posteriormente para alto. As outras classes de domicílios B2 a E não possuem uma relação tão forte quanto a apresentada pelas classes A1 a B1, podendo ser comprovada pela figura 4 que representa a correlação entre todas as variáveis presentes.

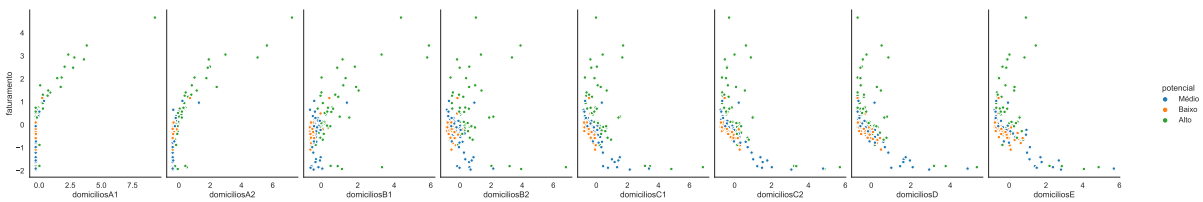


Figura 3: Relação entre o faturamento e as variáveis de domicílio

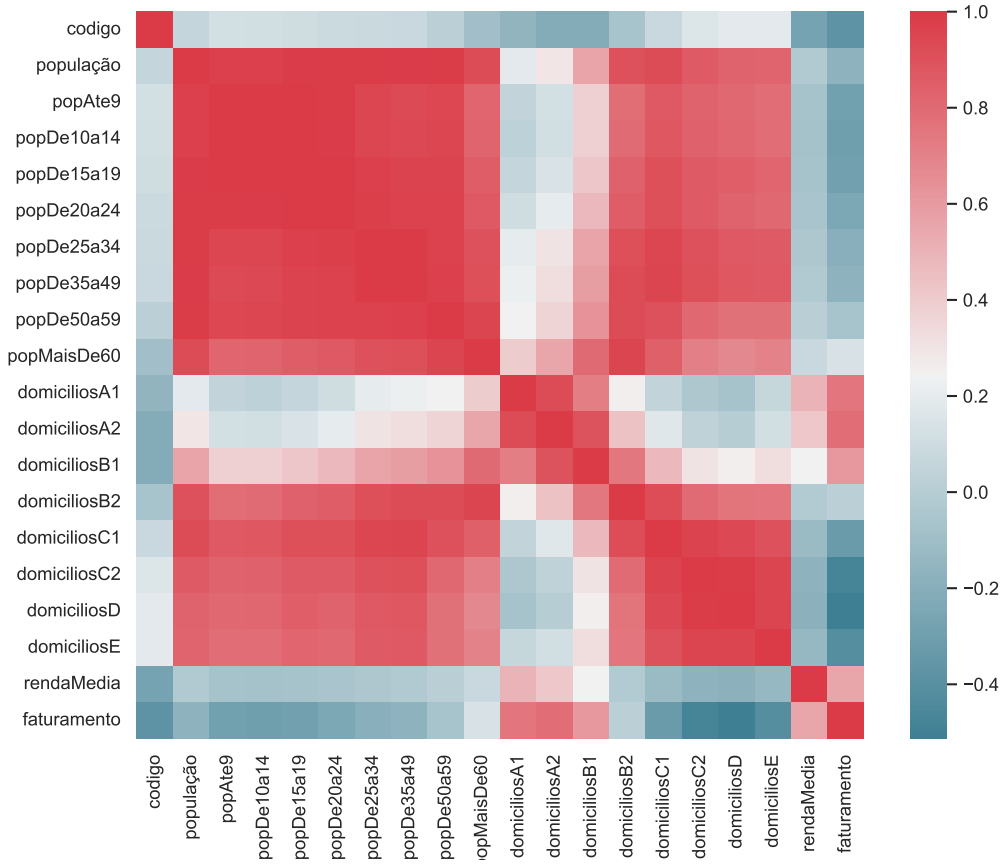


Figura 4: Correlação entre as variáveis

Analisando a correlação entre as variáveis, principalmente a correlação com a variável *faturamento*, é possível inferir que as variáveis com a maior correlação com o faturamento são as variáveis domiciliosA1, domiciliosA2, domiciliosB1 e rendaMedia.

Por se tratar de uma base de dados pequena não é necessário fazer escolhas entre as variáveis para o desenvolvimento de um modelo de classificação e predição. Porém é possível perceber que as variáveis com maior influência sobre o faturamento e potencial, são aquelas que apresentam uma maior correlação, indicadas pela figura 4. A seguir, o pré processamento dos dados será detalhado.

3 Pré Processamento dos Dados

O pré processamento dos dados é uma das etapas com maior influência no desenvolvimento de soluções. Dados não balanceados e não padronizados podem levar a inferências equivocadas, arruinando o desempenho de qualquer modelo.

Os dados, tanto de São Paulo quanto do Rio de Janeiro foram enviados em uma única base de dados e portanto foram separados em dois arquivos. O primeiro contém os dados referentes ao estado do Rio de Janeiro e o segundo referente ao estado de São Paulo. Dessa forma, a base do estado do Rio de Janeiro será utilizada como uma base de aprendizagem e validação dos modelos. Após a separação dos estados, 6 registros da base de dados do Rio de Janeiro na coluna de renda média estavam vazios, por se tratar de uma base de dados com apenas 160 entradas, para não perder os 6 registros que representam quase 4% dos registros, os valores em branco foram preenchidos com a média de todos os outros registros, dessa forma a média geral da tabela não sofre nenhuma alteração, não prejudicando o desenvolvimento da solução com informações não condizentes com a realidade.

O próximo passo tomado foi a separação das colunas que não devem fazer parte das bases de dados de treino e validação dos modelos, que são as colunas de código, nome, cidade e estado. Após retirar as

colunas indesejadas, o o pré processamento segue para a montagem das bases de treino e validação.

Para que um modelo possa aprender é necessário que a base seja separada nas entradas do modelo e nos valores que o modelo deve aprender a classificar ou estimar. A base então deve ter as colunas de faturamento e potencial separadas e então a base como um todo pode ser dividida entre treinamento e validação, também chamado de teste. O processo está resumido de forma gráfica na figura **??**.

A base de dados foi divida de forma que 85% dos registros foram utilizados para o aprendizado do modelo os 15% restantes foram separados para os testes e validação do modelo. Como um dos objetivos do desafio se trata de uma classificação de variáveis descritivas como o potencial Alto, Médio e Baixo, para que um modelo possa aprender a classificar é necessário realizar uma codificação destes valores de texto para números. Utilizando a ferramenta *LabelEncoder* da biblioteca *Sklearn*, os valores Alto, Médio e Baixo foram substituídos respectivamente por 0, 2 e 1.

A seguir os procedimentos para o desenvolvimento do modelo para estimar o faturamento das lojas será apresentado.

4 Estimar o faturamento de cada loja

Por se tratar de um problema para estimar valores contínuos, ou seja, o faturamento de cada loja, diferentes abordagens pode ser utilizadas, como a regressão linear e redes neurais. Para obter o melhor desempenho as duas abordagens foram comparadas e a abordagem com o melhor desempenho foi escolhida.

4.1 Regressão Linear

Utilizando a biblioteca *Sklearn* o modelo de regressão linear foi utilizado para estimar o faturamento de cada loja da base de dados de treinamento. Como métricas para avaliar o modelo gerado pelo algoritmo de regressão linear, foram utilizados o erro médio absoluto, o erro quadrático e a média do erro quadrático. Também foi utilizada a função *explained_variance_score* que serve como uma métrica de acurácia. Os valores obtidos para cada métrica está apresentado na tabela 1.

Erro médio absoluto	129238,72
Erro quadrático	83451573271,49
Média do erro quadrático	288879,85
Acurácia	78,75%

Tabela 1: Tabela de métrica Regressão Linear

4.2 Rede Neural

Também utilizando a biblioteca *Sklearn* uma abordagem de rede neural foi utilizada para predizer os valores de faturamento. Por se tratar de uma abordagem menos complexa para redes neurais os resultados obtidos com essa abordagem não foram tão satisfatórios como mostram as métricas na tabela 2

Erro médio absoluto	173481,87
Erro quadrático	101811089857,87
Média do erro quadrático	319078,50
Acurácia	71,20%

Tabela 2: Tabela de métrica Rede Neural com *Sklearn*

Buscando uma solução que atingisse uma melhor acurácia, uma nova arquitetura de rede neural, um pouco mais complexa foi desenvolvida com o *framework* do *TensorFlow*. Com 5 camadas ocultas e uma de ativação, conforme mostra a figura 6, uma outra abordagem de rede neural foi criada buscando atingir melhores métricas. A tabela 3 mostra que a acurácia utilizando este modelo de rede neural atingiu um valor muito superior as outras abordagens apresentadas.

Após comparar as abordagens apresentadas, o modelo da rede neural produzida com o *TensorFlow* foi adotado e a base de dados de São Paulo foi passada para o modelo onde os valores de faturamento para cada bairro de São Paulo foi estimado. Com a base de São Paulo atualizada com as estimativas de faturamento para cada bairro, a próxima seção apresenta a classificação de potencial para cada bairro.

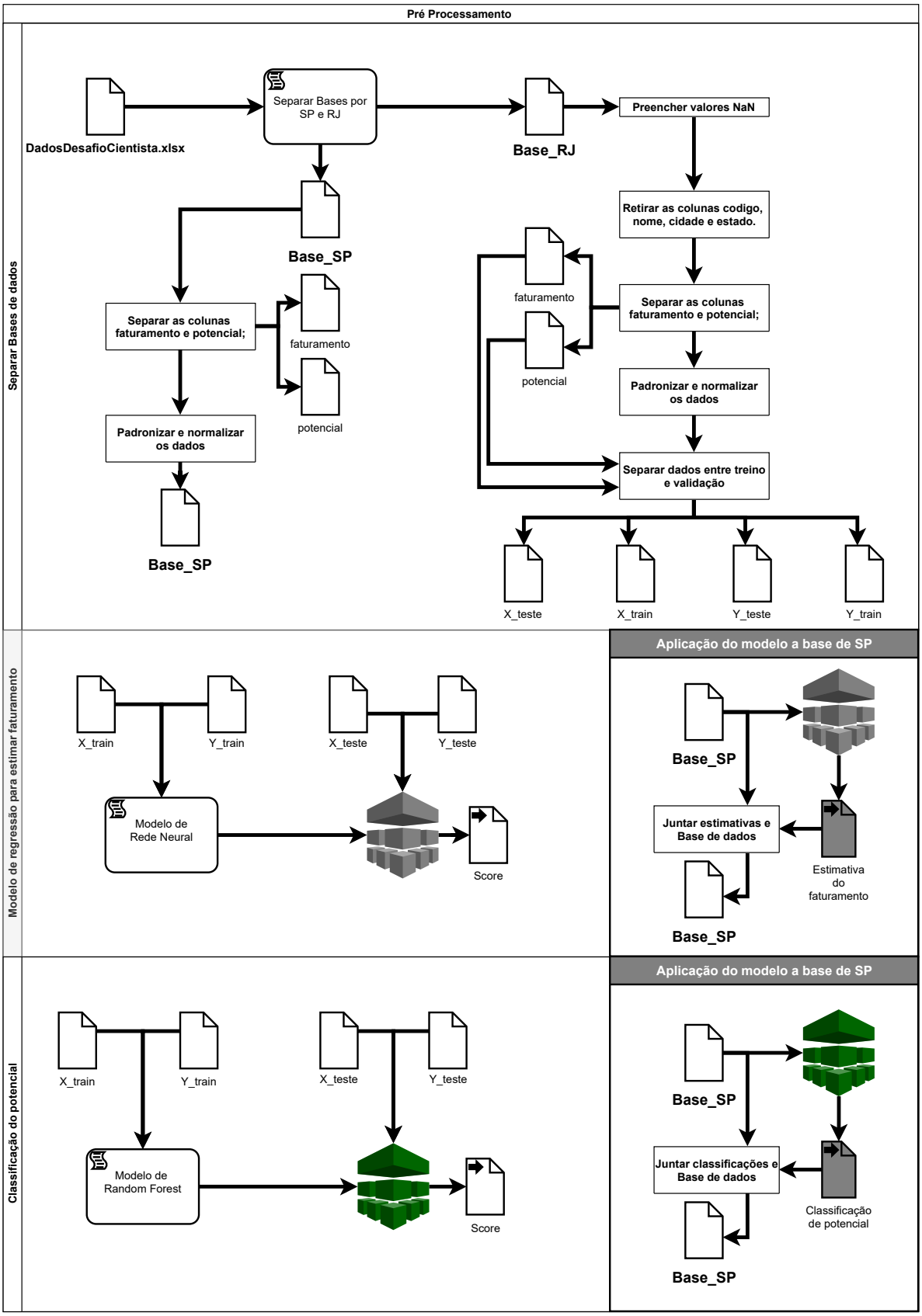
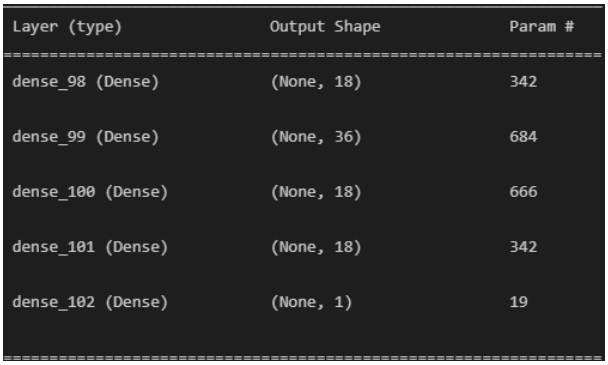


Figura 5: Diagrama de processamento da base de dados



Layer (type)	Output Shape	Param #
dense_98 (Dense)	(None, 18)	342
dense_99 (Dense)	(None, 36)	684
dense_100 (Dense)	(None, 18)	666
dense_101 (Dense)	(None, 18)	342
dense_102 (Dense)	(None, 1)	19

Figura 6: Configuração da rede neural com *TensorFlow*

Erro médio absoluto	50412,29
Erro quadrático	17175720144,35
Média do erro quadrático	131056,17
Acurácia	95,23%

Tabela 3: Tabela de métrica Rede Neural com *TensorFlow*

5 Classificar o potencial de cada loja

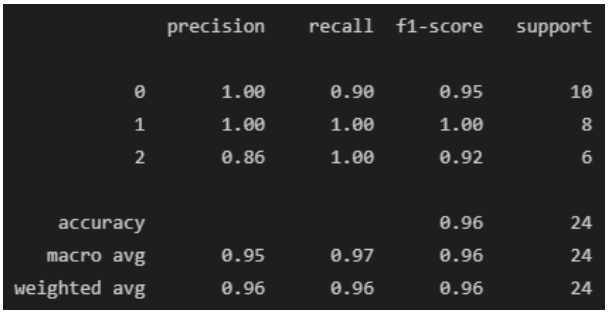
Por se tratar de um problema de classificação, outras abordagens devem ser utilizadas por se encaixarem melhor em problemas de classificação em comparação a estimação de valores contínuos, uma vez que a classificação se torna uma estimativa de valores discretos. Abordagens como *Naive Bayes*, Arvores de Decisão, *K-nearest Neighbors*, *Support Verctor Machines* e *Random Forests*, costumam ser mais utilizadas para problemas de classificação.

Com o objetivo de obter o melhor desempenho para a classificação dos potenciais das lojas todas as abordagens citadas anteriormente foram comparadas e a acurácia obtida para cada uma das abordagens está classificada na tabela 4

Abordagem	**Acurácia (%)**
Naive Bayes	50%
Arvore de Decisão	83%
Random Forest	96%
KNN	83%
SVM	79%

Tabela 4: Comparação entre as abordagens

A abordagem com maior acurácia foram as *Random Forests*. A figura 7 apresenta as métricas de precisão, que de forma simples significa: Quantos itens classificados são classificados corretamente. *Recall*, que pode ser explicado como: Quantos itens classificados para uma classe são corretamente classificados. *F1-score* representa a média harmônica entre as métricas de precisão e *recall*.



	precision	recall	f1-score	support
0	1.00	0.90	0.95	10
1	1.00	1.00	1.00	8
2	0.86	1.00	0.92	6
accuracy			0.96	24
macro avg	0.95	0.97	0.96	24
weighted avg	0.96	0.96	0.96	24

Figura 7: Relatório de Métricas

6 Resultados

De posse da Base de dados de São Paulo com a estimativa de faturamento e a classificação dos bairros em três diferentes potenciais, alguns resultados podem ser gerados para que decisões possam ser tomadas com mais segurança e confiabilidade. Como os domicílios de classes A1, A2, B1 e B2 são de interesse da empresa, assim como a população com faixa etária entre 25 e 50 anos, foram geradas duas novas informações para focar os resultados no público alvo e domicílios de interesse.

Os domicílios alvo, mostrados na figura 8, são representados pela soma dos domicílios A1, A2, B1 e B2. É possível perceber que os bairros de São Paulo, cada um representado por um ponto na figura 8, podem ser segmentados. O raio de cada bairro na figura é definido pelo tamanho da população alvo, ou seja, a população entre 25 e 50 anos. Pontos maiores possuem uma maior quantidade de pessoas da população alvo vivendo no bairro.

Segmentando os bairros que possuem mais de 5 mil domicílios alvo e bairros com estimativa de faturamento acima de 1 Milhão de reais é possível capturar quase todas os bairros que possuem um potencial elevado. Além disso alguns bairros dentro desse segmento e que estão classificados como um potencial Médio podem ser estudados com uma maior profundidade. Como os modelos fazem estimativas e classificações, alguns erros podem ser esperados e nem todos os bairros necessitam ser desconsiderados por estar com outras classificações.

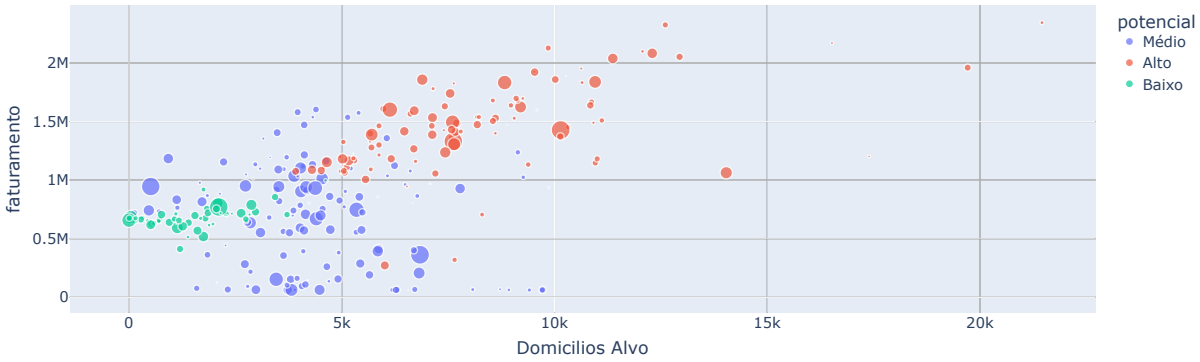


Figura 8: Faturamento por Quantidade de domicílios alvo

Utilizando a população alvo como forma de análise também é possível obter diferentes respostas para os bairros. A quantidade de domicílios alvo é demonstrada pelo raio de cada amostra presente na figura 9. Podemos ver que bairros com uma população alvo abaixo de 5 mil e acima de 50 mil pessoas tem uma grande chance de não possui um potencial Alto. Se segmentarmos ainda mais os bairros podemos ver que a maioria dos bairros de alto potencial possuem uma estimativa de faturamento superior a 1 Milhão de Reais e uma população alvo entre 5 mil e 30 mil pessoas.

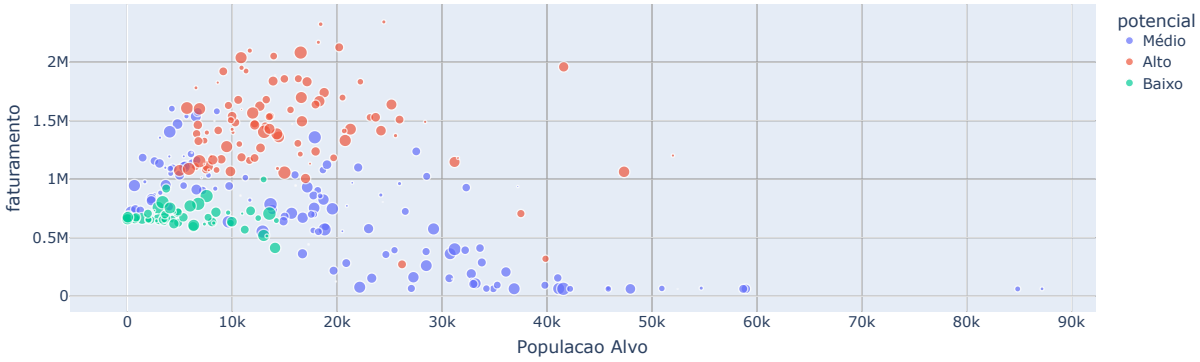


Figura 9: Faturamento por Quantidade de população alvo

Outra ferramenta visual que ajuda a segmentar os bairros é o gráfico de *box plot* apresentado na figura 10. A maioria dos bairros de São Paulo classificados como potencial Alto possuem um faturamento

superior a 1 Milhão de Reais. Comparando com a classe de bairros com potencial Médio, 75% dos bairros categorizados com esse potencial apresentam um faturamento estimado inferior a R$ 932515, 00. Quando comparado com bairros de potencial Baixo, apenas dois bairros apresentam uma estimativa de faturamento superior a 1 Milhão de Reais.

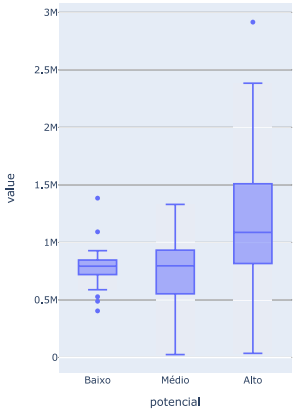


Figura 10: Box plot para o faturamento de cada potencial

Os resultados apresentados na figura 10 ajudam a formular a hipótese de que bairros com uma previsão de faturamento acima de 1 Milhão de reais podem ser muito atrativos para receber unidades da empresa. Dessa forma uma lista de bairros que se encaixam nas características citadas anteriormente pode ser gerada, apresentando as lojas com os maiores potenciais no estado de São Paulo.

Pesquisando bairros com uma população alvo entre 5 mil e 30 mil pessoas com mais de 5 mil domicílios alvo e com uma estimativa de faturamento superior a 1 Milhão de Reais, foi possível chegar a uma lista com 97 bairros dos 296 bairros presentes na base de dados. Destes 97 bairros selecionados, 9 bairros foram classificados com o potencial Médio e nenhum bairro classificado com potencial Baixo. Restando 88 bairros com potencial alto, uma busca de informações adicionais referentes a estes bairros poderia ser realizada.

Informações como custo de aluguel médio, quantidade de lojas concorrentes presentes nos bairros, presença de facilidades como estacionamentos, paradas de ônibus, metrôs, entre outros, poderiam ser utilizadas para reduzir ainda mais o número de bairros. Com informações adicionais, a decisão entre os bairros poderia ser realizada com uma precisão ainda maior, reduzindo o risco de investimento da empresa para abrir novas unidades em locais que devem retornar os melhores resultados.

Os cinco bairros de São Paulo com maior faturamento dentro da segmentação realizada são apresentados na tabela abaixo:

Nome	**Faturamento (R$)**	**Potencial**	**População Alvo**	**Domicílios Alvo**
Moema	2344377.25	Alto	24444.0	21456.0
Tatuapé	2324855.75	Alto	18425.0	12604.0
Saúde	2302159.75	Alto	21984.0	15445.0
Perdizes	2170066.00	Alto	18203.0	16520.0
Vila Andrade	2127303.25	Alto	20187.0	9852.0

Tabela 5: 5 bairros com maior faturamento

7 Conclusão

Tomadas de decisões baseadas em dados possuem uma grande relevância atualmente para empresas e afetando diversas áreas como produtividade, faturamento, pessoas, entre outras mais [PF13]. Neste documento, foi demonstrado o processo do desenvolvimento de uma solução para tomadas de decisão baseadas em dados. O caso estudado envolve dados sociodemográficos dos estados do Rio de Janeiro e

São Paulo, além do faturamento e potencial de uma empresa alimentícia. Com o objetivo de expandir suas lojas do estado do Rio de Janeiro para São Paulo, uma segmentação dos bairros de São Paulo que podem receber futuras lojas foi realizada, mostrando as relações existentes entre a população alvo, a quantidade de domicílios alvo e o faturamento que a empresa pretende atingir para cada bairro.

Referências

[JPW17] Andrew T Jebb, Scott Parrigon, and Sang Eun Woo. Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*, 27(2):265–276, 2017.

[PF13] Foster Provost and Tom Fawcett. Data science and its relationship to big data and data-driven decision making. *Big data*, 1(1):51–59, 2013.