UNIVERSIDADE DE SÃO PAULO

ESCOLA POLITÉCNICA

ANA CAROLINA MEIRELLES NEGRELLI

MARINA HIKARI DA SILVA

RODRIGO YUGI SAKAI

Previsão de demanda de chope através de algoritmos de Inteligência Artificial ligada à logística urbana

São Paulo

2019

ANA CAROLINA MEIRELLES NEGRELLI

MARINA HIKARI DA SILVA

RODRIGO YUGI SAKAI

Previsão de demanda de chope através de algoritmos de Inteligência Artificial ligada à logística urbana

Trabalho de Formatura apresentado à Escola Politécnica da Universidade de São Paulo para obtenção do bacharelado em Engenharia Civil

Orientador: Prof. Dr. Claudio Barbieri da Cunha

São Paulo

2019

Negrelli, Ana Carolina Meirelles

Da Silva, Marina HIkari

Sakai, Rodrigo Yugi

Previsão de demanda de chope através de algoritmos de Inteligência Artificial ligada à logística urbana

Trabalho de Formatura – Programa de Graduação em Engenharia Civil, Escola Politécnica, Universidade de São Paulo, São Paulo, 2019.

RESUMO

Este trabalho de formatura foi feito em associação com uma empresa produtora e fornecedora de bebidas e aborda o problema de previsões de demanda de um produto perecível - mais especificamente, o chope. Diferentemente de outras bebidas, o chope, tipicamente, não é pasteurizado, o que acaba por reduzir seu prazo de validade. Por não poder permanecer muito tempo estocado, ele cria um problema logístico, uma vez que tanto a falta quanto o excesso de produção acarretam prejuízos.

Utilizando dados históricos de comercialização dos produtos da empresa, propôs-se a construção de um modelo de Aprendizado Supervisionado que aprendesse a mapear uma função capaz de estimar, a partir das variáveis de entrada, o volume de consumo. Para comparar os efeitos de uma amostragem maior, foi feito um processo de aumento de dados, que resultou em um ganho de cinco vezes a quantidade de registros originais. Treinaram-se modelos separados com ambas bases de dados e seus resultados foram comparados com o modelo empregado pela empresa.

Foram testados um modelo paramétrico (Regressão Linear) e três não paramétricos (Árvore de Regressão, Random Forest e K-Nearest Neighbors), com base em suas performances na validação cruzada, em que o critério de avaliação utilizado foi o Erro Absoluto Médio. Usando esta mesma métrica, obteve-se uma comparação quantitativa entre o desempenho dos modelos de Aprendizado de Máquina e o empregado pela empresa.

Para complementar o estudo de demanda, foi feito um estudo de distribuição logística, a fim de trazer ainda mais eficiência na cadeia do chope. Foi comparado ao método tradicional de entrega, a instalação de centros de distribuição urbana que considera um centro adicional de armazenamento. Assim, foram feitos levantamentos de roteiros e custos em ambos casos, sendo um estudo preliminar, para determinar a necessidade de um estudo mais profundo sobre o tema.

Palavras-chave: Aprendizado de Máquinas. Previsão. Logística urbana. Alimentos perecíveis. Desperdício. Centro de Distribuição Urbana

ABSTRACT

This thesis was made in association with a beverage producer and supplier company. It approaches the problem of demand forecast of a perishable product: chopp. Unlike most beverages, chopp isn't pasteurized, which reduces its expiration date. Therefore, it cannot be stored for a long time, creating a logistic problem, once both the underproduction and the overproduction lead to a loss for the company.

Using historical data of the company's commercial operation, a Supervised Learning model was built in order to map a function capable of estimating the demand from input variables. In order to compare the effects of a larger sample, a data augmentation process was conducted, which resulted in a gain of five times the amount of data. Separated models were trained with each database, and the results were compared with the current strategy employed by the company.

One parametric (Linear Regression) and three non-parametric (Regression Tree, Random Forest and K-Nearest Neighbors) models were tested, and the ones with the best fit were chosen according to the least Mean Absolute Error in a cross-validation process. Using this metric, a quantitative comparison among the models developed and the one employed by the company was obtained.

To complete the demand study, a logistic distribution research was made, in order to improve the chain efficiency. The traditional delivery method was compared with the use of urban warehouses, a third tier to the delivery chain. To accomplish a preliminary study, travel routes and travel costs were simulated, aiming to verify the necessity of a further study.

Keywords: Machine Learning. Forecast. Urban Logistics. Perishable food. Food waste. Urban Warehouses

LISTA FIGURAS

SUMÁRIO

1. INTRODUÇÃO

Com a urbanização crescente no último século, a demanda logística aumentou muito, devido à concentração de pessoas em grandes centros, trazendo aumento do consumo de mercadorias. Isso trouxe um grande desafio logístico nas grandes cidades, aumentando o tráfego de veículos destinados às mercadorias, causando maior impacto nas vias urbanas. Dentre os produtos e serviços oferecidos para a população, o fornecimento de produtos perecíveis apresenta alguns desafios. Há dificuldade na previsão de sua demanda, no transporte e na gestão de estoques. Segundo Alcione Silva, integrante do comitê da ONU Save Food Brasil, "Enfrentamos, sobretudo, um desafio estrutural, na cadeia de distribuição, e o desafio comportamental, do consumidor" (Como O Desperdício De Alimentos Afeta O Brasil E O Seu Bolso | Huffpost Brasil, [S.D.])

Dado este contexto geral, esse trabalho terá um foco maior na logística de produtos perecíveis, em especial o chope, que possui dificuldades de previsão de demanda e, portanto, problema de desperdício.

Será estudado um caso específico de um grande fornecedor de produtos perecíveis a fim de analisar o possível impacto que modelos de previsão de demanda podem ter sobre a logística do negócio e sobre seu entorno: custos e viagens extras por conta de subdimensionamento, ociosidade por superdimensionamento e a possibilidade de se entregar o chope usando outros veículos.

Os centros de distribuição são o foco deste estudo. A rotina do produto começa nos pedidos feitos às cervejarias. Das cervejarias, os produtos acabados podem ser destinados a três categorias de distribuição. Os produtos podem ser enviados aos Centros de Distribuição, para os revendedores ou diretamente para os *key accounts*, que são os clientes de grande demanda, por exemplo supermercados e atacadistas.

Dada a breve contextualização, um problema da previsão de demanda de chope é que, por ser uma bebida não pasteurizada, ela tem prazo de validade curto, assim, a venda do chope é diferenciada. Os chopes têm, em média, um prazo de validade de 12 dias e é de responsabilidade da produtora entregar os barris com, pelo menos, três dias de prazo antes do vencimento. A solução para o problema atualmente é de se produzir chope sob demanda.

O vendedor, em seus turnos, pergunta para cada cliente o quanto ele vai precisar de chope na próxima semana, sem que o cliente precise comprar de fato essa quantidade informada no futuro. Essas previsões são computadas todos os dias e, na terça-feira de cada

semana, são processadas de forma diferenciada. Na quinta-feira, os barris chegam ao CDD e são distribuídos aos pontos de venda, por exemplo bares e restaurantes. Ou seja, os clientes estimam na semana anterior o quanto consumirão no final da semana seguinte.

Por isso, quando há um subdimensionamento da demanda, há uma grande dificuldade em produzir, envasar e entregar o chope dentro do intervalo de consumo e, por outro lado, quando há um superdimensionamento, o chope passa do seu prazo de validade e estraga. Diante disso, o chope apresenta-se como um produto difícil de se trabalhar no campo da logística.

A fim de criar um algoritmo de Inteligência Artificial capaz de prever a demanda de chope, serão estudadas, em colaboração com a empresa, as variáveis externas que influenciam o pedido inicial - como sazonalidade e indicadores econômicos - e suas correlações com a variação da demanda.

Em seguida, será modelado um algoritmo de Aprendizado de Máquinas com base nos fatores que mais se correlacionam com as variações do pedido inicial. Feito isso, o algoritmo será treinado com base nos dados analisados, prioritariamente utilizando o aprendizado supervisionado, e seu desempenho, como um algoritmo de regressão, será medido com base no Erro Quadrático Médio e no Erro Médio Absoluto.

Além da diminuição de perdas (desperdícios) na previsão de pedidos, a eficiência também deve ser estudada na entrega desses pedidos previstos ao ponto de venda. Neste trabalho, será estudado o uso de *mini-hubs*, ou Centros de Distribuição Urbana (CDU), para otimizar as entregas. Atualmente são feitas através de um caminhão de grande porte, que é carregado no CDD e depois faz um roteiro urbano para realizar entregas em cada ponto de venda. Isso gera muitos impactos negativos para a cidade, como poluição e barulho, além de uma ineficiência de entrega. O caminhão consome mais combustível e mais tempo para realizar entregas entre distâncias pequenas.

Uma alternativa logística para esse problema é a criação de centros de distribuição urbanos, que recebem o produto perecível e, deste, outros veículos de menor porte distribuem aos pontos de venda. Essa solução será estudada ao decorrer deste trabalho a partir de um estudo preliminar de custos com transporte, além do cálculo estimado de custos. Essa análise busca avaliar a viabilidade de se começarem estudos de implantação de um CDU de forma mais aprofundada.

2. REVISÃO BIBLIOGRÁFICA

2.1. INTELIGÊNCIA ARTIFICIAL

2.1.1. Modelagem Preditiva

Segundo Alpaydin (2010), para um conjunto de registros observados acerca de um fenômeno, principalmente quando este tange o comportamento humano, a probabilidade de que exista um conjunto de regras que explique seu funcionamento é maior do que a probabilidade de que este evento seja regido por completa aleatoriedade. Embora seja improvável obter e entender integralmente as normas que o governam, é possível elaborar uma aproximação capaz de computar parte da informação observada: um modelo. Ainda que não explique perfeitamente sua conduta, a estimativa pode identificar padrões que torne a interpretação humana mais intuitiva, ou auxiliar na predição de eventos futuros - assumindo que estes não divirjam muito dos já observados.

Um modelo é uma representação simplificada da realidade - as simplificações são feitas para descartar informações irrelevantes do problema, permitindo focar nos aspectos que se deseja entender. A modelagem preditiva é o processo de computar resultados conhecidos e desenvolver um modelo capaz de prever valores para novas observações - ela se apoia em dados históricos e em exemplos de treino para predizer eventos futuros. Há diversos tipos de técnicas de modelagem preditivas, como Regressão Linear, Regressão Logística, Árvores de Decisão, Redes Neurais, entre muitas outras. Seus mecanismos de funcionamento, no entanto, seguem o mesmo princípio: criar um conjunto de regras a partir do conjunto de treino para transformar variáveis de entrada, comumente chamadas de X, em variáveis de saída, também chamadas de Y (ALPAYDIN, 2010).

O teorema de que "Não existe almoço grátis" (WOLPERT e MACREADY, 1997), difundido no campo da Otimização, também se aplica à Modelagem Preditiva. Ele evidencia que, para problemas com iterações e funções de perda (funções que mapeiam valores de uma ou mais variáveis a um número real representando algum tipo de custo associado) finitos, não há um modelo com desempenho universal ótimo. Com isso, se um certo modelo apresenta resultados melhores do que os demais para determinado problema, então, para outros, ele terá, necessariamente, resultados piores.

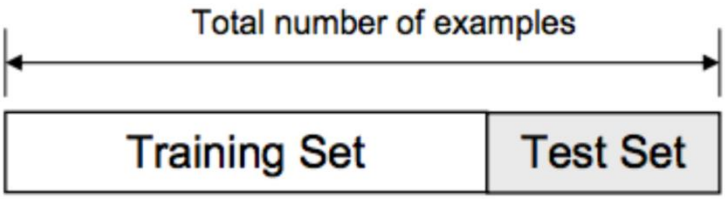
Dessa forma, uma das etapas principais da Modelagem Preditiva compreende a escolha do modelo a ser adotado. Dado que, globalmente, nenhuma técnica será superior às

demais, é necessário focar no problema em mãos, de forma a descobrir aquele cujas hipóteses mais se alinham ao problema e aos seus dados (ALPAYDIN, 2010).

2.1.2. Ajuste

Segundo Katsuri (2019), o desempenho de um modelo é medido através de sua capacidade de previsão e generalização do problema, *i.e.* reduzir o erro entre valor observado e valor previsto para qualquer observação. Para fazer essa avaliação, uma prática bastante difundida é dividir aleatoriamente o conjunto de dados observados em parcelas de treino e teste - em geral, a proporção adotada fica em torno de 75% e 25% respectivamente. Uma ilustração de seu funcionamento pode ser observada na Figura 1. O princípio deste método é criar o sistema de regras da metodologia preditiva através dos registros do primeiro conjunto e usá-los para estimar os resultados do segundo. Sabendo os resultados observados, é possível compará-los com os estimados e, assim, obter uma quantificação do desempenho para cada metodologia desejada. Dessa forma, escolhe-se uma técnica de modelagem com base na performance sobre os dados já coletados. Se os eventos futuros seguirem o mesmo padrão, a eficiência do modelo se manterá (ALPAYDIN, 2010).

Figura 1 - Ilustração da técnica de separação treino e teste



Fonte: Bronshtein (2017)

De acordo com Singh (2018), ao erro resultante entre o valor estimado e o observado, podem-se atribuir duas parcelas: irredutível e redutível. A primeira contabiliza uma fração que não será eliminada, independentemente da modelagem empregada e de procedimentos subsequentes. Frequentemente, tem sua origem atribuída a variáveis desconhecidas que influenciam a situação. A parcela redutível, por outro lado, corresponde à que pode ser limitada através de técnicas de modelagem. Para melhor entendê-la, a autora propõe dividi-la em dois componentes: viés e variância.

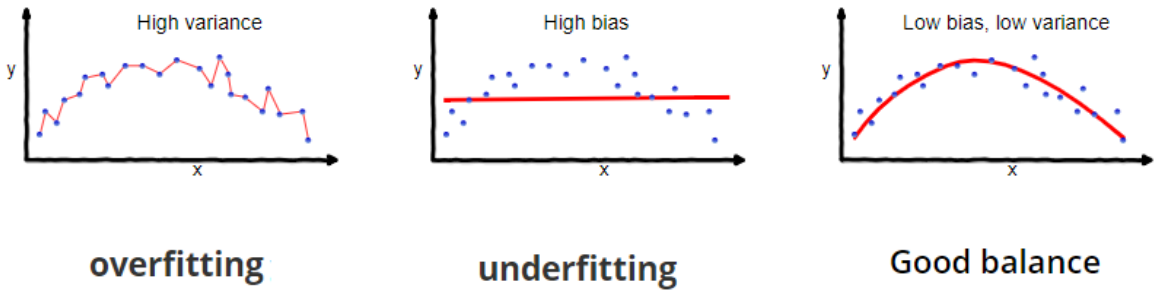
Viés indica o quão longe os valores estimados estão dos valores observados - quanto mais distantes, maior o viés. Alto viés aponta que a modelagem foi muito simples e incapaz de computar a complexidade da relação entre as variáveis de entrada e saída. Diz-se, então,

que houve um sobajuste (*underfitting*). Como resultado, um modelo subajustado produzirá, tanto para o conjunto de treino quanto para o de teste, estimativas consistentes, porém pouco fidedignas ao valor real (SINGH, 2018).

Variância indica o quão disperso os valores estimados estão dos valores observados - quanto mais dispersos, maior a variância. Alta variância aponta que o modelo foi muito complexo, ajustando-se exacerbadamente aos dados de treino. Diz-se, então, que houve um sobreajuste (*overfitting*). Como consequência, um modelo sobreajustado produzirá estimativas fidedignas com dados de treino, já conhecidos, e inexatas com dados de teste, jamais vistos (SINGH, 2018).

Um modelo balanceado é, portanto, definido por Singh (2018), como sendo aquele que não possui nem grande viés nem grande variância: ele não deve ser simples a ponto de não contabilizar as relações existentes entre as variáveis fornecidas, nem complexo a ponto de privilegiar o caso específico em detrimento do caso geral. Como consequência, um modelo balanceado deve produzir estimativas confiáveis tanto para o conjunto de treino quanto para o de teste. Tem-se, portanto, o dilema entre o balanceamento de viés e variância. Uma ilustração dos ajustes é fornecida na Figura 2.

Figura 2 - Ilustração de sobreajuste, sobajuste e ajuste balanceado



Fonte: Bronshtein (2017)

2.1.3. Validação Cruzada

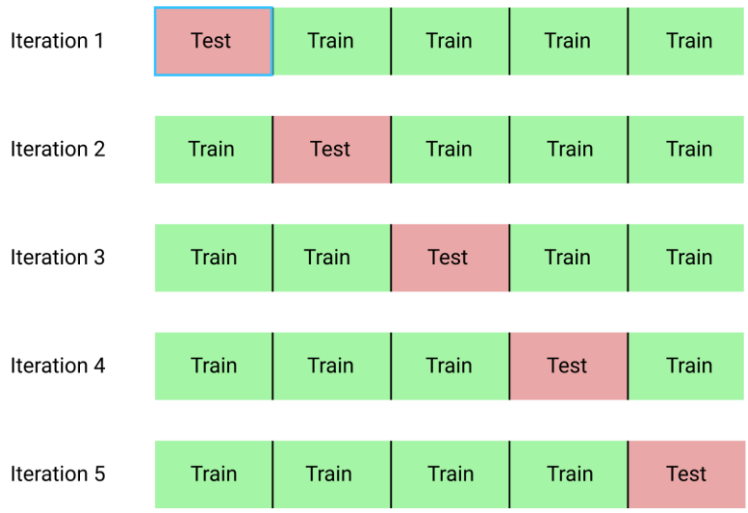
Sendo uma metodologia bastante simples, a separação entre treino e teste pode originar resultados enviesados, uma vez que a divisão nos dois subconjuntos também seguir um viés. Se a quantidade de dados for limitada, é bastante provável que um dos subconjuntos acumule uma parcela da distribuição em detrimento de outra. Quando este for o caso, o modelo acabará tendo um sobreajuste, adequando-se a uma situação específica do problema e não ao caso geral. Para minorar esta implicação e obter uma quantificação mais válida do

desempenho de modelos de previsão, recorre-se à validação cruzada (BRONSHTEIN, 2017). Essa técnica consiste em realizar uma série de análises nas quais um conjunto que antes foi usado para treino é usado para teste e vice-versa, de forma a produzir resultados complementares. Assim, cada registro é usado ao menos uma vez para treino e validação.

2.1.4. Método K-Folds

Seu princípio consiste na subdivisão do conjunto de treino em K partes iguais - tradicionalmente, por uma questão de custo computacional, K é igual a três ou cinco. A primeira parte será retida, enquanto as demais k-1 são usadas para treino. Uma vez treinado, o modelo tentará prever os resultados da primeira parte. Em seguida, a parte a ser retida para validação será a segunda, e o modelo será treinado com as k-1 partes restantes. Repete-se o processo, iterando o subconjunto que é usado para validação até que a última parte seja utilizada para teste - dessa forma, todo registro da base será incluído tanto no subconjunto de teste quanto no de validação. Para cada iteração, contabiliza-se o desempenho do modelo, de acordo com a medida definida. Ao fim, serão gerados K resultados, que fornecem um retrato mais confiável do modelo do que a divisão em conjuntos de treino e teste. Assim, é possível seguir com aquele que obteve melhor desempenho (SHULGA, 2018). Uma ilustração deste processo iterativo é fornecida na Figura 3.

Figura 3 - Ilustração da validação cruzada K-Fold



Fonte: Shaikh (2018)

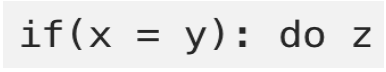
2.1.5. Aprendizado de Máquina

O Aprendizado de Máquina é um subconjunto da área de Inteligência Artificial e, segundo Nabi (2018), o termo surgiu pela primeira vez na década de 1950 e foi definido

como a área de estudo que fornece a computadores a habilidade de aprender sem terem sido explicitamente programados para tanto.

A grande inovação que o Aprendizado de Máquina proporcionou no campo da Inteligência Artificial foi viabilizar a criação de um tipo de arquitetura que extrapolasse a do agente tabela (EDWARDS, 2018). Como todo agente inteligente, o agente tabela deve possuir mecanismos de percepção do meio em que está inserido e mecanismos de atuação, de forma a agir com base na leitura do ambiente. No entanto, sua tomada de decisão é concebida com o auxílio de uma tabela, que deve conter as possíveis percepções do ambiente e as respectivas ações a serem realizadas. Na Figura 4, tem-se uma ilustração desta arquitetura. Sua ineficiência encontra-se na dificuldade ou inviabilidade de se compor uma tabela capaz de computar todas as percepções possíveis, que tornem o agente, de fato, racional (REALI COSTA, 2019).

Figura 4 - Ilustração de leitura e tomada de ação em uma arquitetura de agente tabela.



Fonte: Edwards (2018)

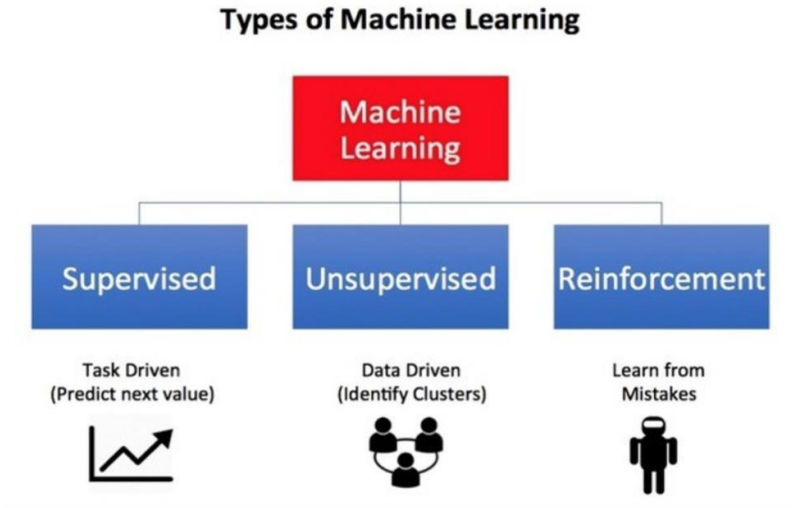
Com isso, se, antes, o computador precisava de uma série de instruções explicitamente fornecidas pelo programador a fim de transformar uma entrada em uma saída, *i.e.* um algoritmo, para resolver uma tarefa, agora, com o Aprendizado de Máquina, a máquina é capaz de extrair automaticamente um algoritmo a partir de dados que recebe e modelar o problema - uma vez fornecidos dados e exemplos o suficiente do problema proposto. Segundo Chollet (2017), tradicionalmente, a Engenharia de Software permitia combinar regras com informação para criar respostas a problemas. O Aprendizado de Máquina, entretanto, utiliza informação e respostas para descobrir as regras por trás de um problema.

Um modelo compreende parâmetros pré-definidos e o escopo do Aprendizado de Máquina, portanto, envolve a programação de computadores com o intuito de otimizá-los a partir de dados históricos. Para isso, utiliza-se a teoria estatística, uma vez que a tarefa envolve a criação de modelos matemáticos capazes de extrapolar uma amostragem ao caso geral (ALPAYDIN, 2010).

2.1.6. Tipos de Aprendizado

Segundo Heidenreich (2018), definem-se três tipos de aprendizado no campo do Aprendizado de Máquina: Supervisionado, Não Supervisionado e por Reforço. Eles se diferem tanto por metodologia quanto por objetivo. A Figura 5 demonstra essas vertentes.

Figura 5 - Ilustração dos tipos de aprendizado de máquina

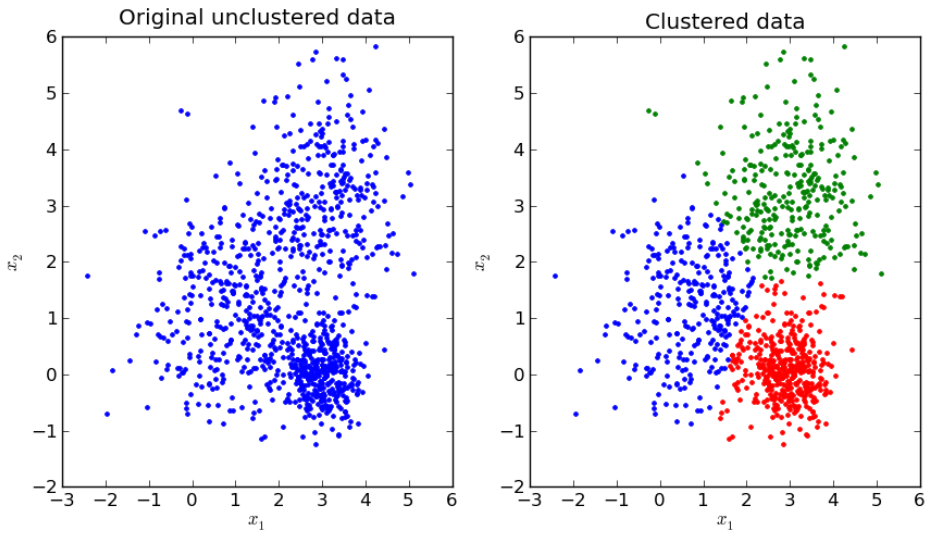


Fonte: Heidenreich (2019)

O Aprendizado Supervisionado é feito usando dados rotulados, ou seja, sabe-se, de antemão, os resultados que a amostra deve apresentar. Desta forma, o objetivo é obter uma função que retorne corretamente o rótulo de novas observações em a partir dos atributos conhecidos. Por isso, diz-se que é uma vertente orientada à tarefa de prever a próxima ocorrência. Quando o rótulo que se deseja estimar é uma variável categórica (ou discreta), então trata-se de um problema de classificação. Por outro lado, se este assume valores reais (ou contínuos), então tem-se um problema de regressão (HEIDENREICH, 2019).

O Aprendizado Não Supervisionado tem como objetivo agrupar conjuntos de dados não rotulados a partir da semelhança de seus atributos. Um exemplo de sua aplicação pode ser observado na Figura 6. Nela, a partir dos atributos x1 e x2, os dados observados são classificados em um dos três grupos: azul, verde ou vermelho, previamente desconhecidos - por isso, não rotulados.

Figura 6 - Ilustração do funcionamento de um aprendizado não supervisionado

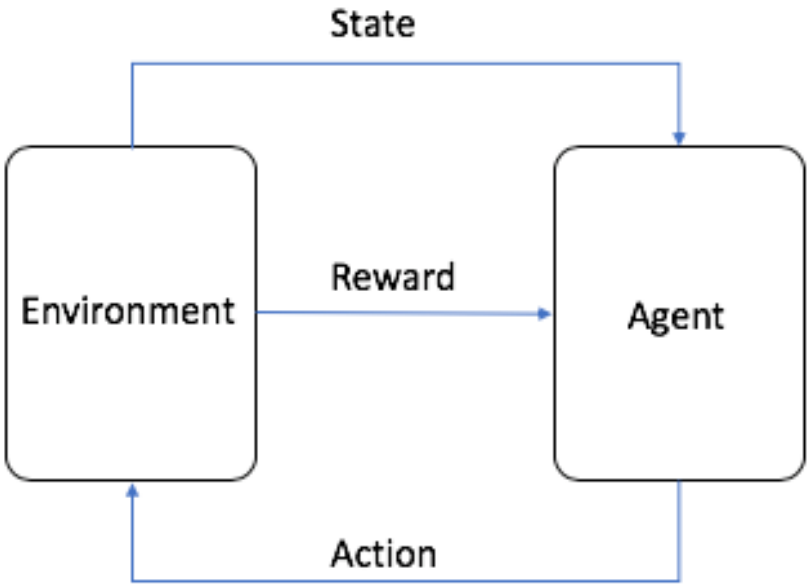


Fonte: Heidenreich (2019)

Este tipo de aprendizado baseia-se fortemente nos dados e em suas propriedades: as saídas dependem de como estes são estruturados antes de serem fornecidos ao algoritmo - o simples fato de alterar a escala de um atributo de metros para centímetros, por exemplo, pode provocar mudanças nos resultados. Por isso, pode-se dizer que é um tipo de aprendizado orientado à informação (HEIDENREICH, 2019).

Já o Aprendizado Por Reforço é orientado ao comportamento. Muito empregado em vídeo games e em robótica, esta variação consiste no fornecimento de avaliação das ações tomadas por um modelo de Inteligência Artificial inserido em um determinado ambiente. Espera-se que o agente cometa inúmeros erros no começo, porém, se sinalizado positiva ou negativamente após uma boa ou má ação tomada, com o tempo, o agente aprenderá a tomar decisões majoritariamente boas. Neste caso, há uma forte relação entre o agente inteligente e o ambiente em que está inserido, que é exemplificada na Figura 7. O agente precisa saber das ações possíveis que pode escolher e o meio precisa fornecer a avaliação, identificando se esta foi boa ou ruim. Por fim, o efeito da ação do agente deve promover mudanças no ambiente, de maneira que a nova leitura atualize o estado interno do algoritmo (HEIDENREICH, 2019).

Figura 7 - Ilustração da relação entre agente e ambiente no aprendizado por reforço



Fonte: Li, Meagher e Davis (2019)

2.1.7. Regressão

Regressão é um dos tipos de problemas inseridos no subconjunto do Aprendizado Supervisionado. Neste tipo de aprendizado, o objetivo é, a partir de conjuntos de dados conhecidos, obter uma função de mapeamento f que estime, para um conjunto de variáveis de entrada X, um valor contínuo y, ou seja aproximar ao máximo f(X) = y (GARBADE, 2018).

2.1.7.1. Regressão Linear

A Regressão Linear explora a relação entre variáveis de entrada e saída através de uma equação linear. Um modelo de Regressão Linear é dito simples, quando há apenas uma variável de entrada. Quando houver mais do que uma, ela é, então, dita múltipla. No caso simples, a equação linear é dada por:

$$Y = m*X + b$$

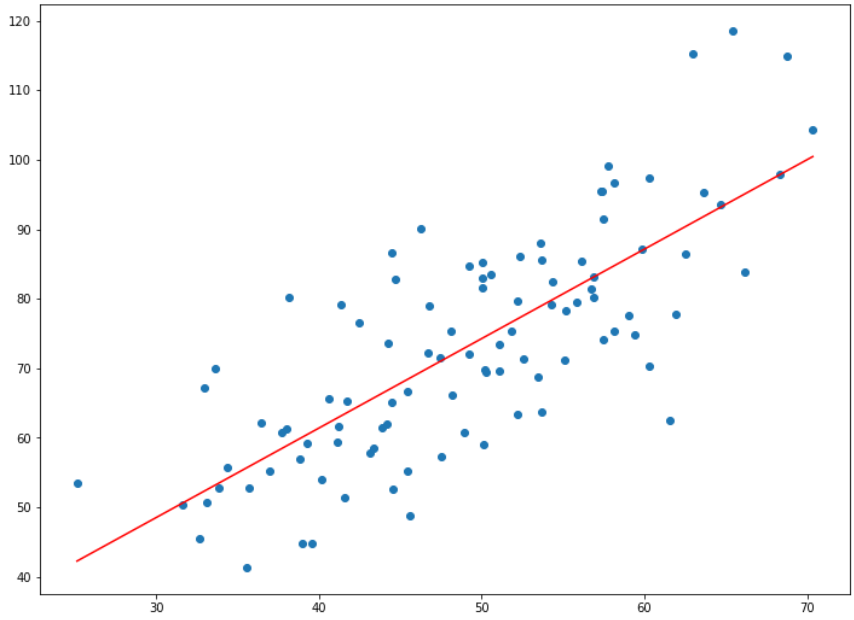
onde m é o gradiente (coeficiente angular da reta), X a variável em questão e b o intercepto do eixo Y. No caso de uma Regressão Linear Múltipla, a equação resultante é:

$$Y = m1*X1+m2*X2+...+mn*Xn + b$$

Onde Xi é cada variável de entrada e mi seu respectivo gradiente (BHATTACHARYYA, 2018).

Os gradientes e o intercepto são obtidos através do Método dos Mínimos Quadrados, que minimizam a soma das distâncias entre os pontos do conjunto de dados e o hiperplano resultante da Regressão Linear (MENON, 2018). A Figura 8 ilustra o resultado de uma Regressão Linear Simples.

Figura 8 - Gráfico ilustrando o resultado de uma Regressão Linear Simples



Fonte: Menon (2018)

Apesar de ser uma ferramenta capaz de resolver problemas complexos e computar as influências de múltiplas variáveis, seus resultados não são confiáveis se as hipóteses assumidas não forem verificadas. Segundo Soma (2018), elas são:

- Linearidade: um modelo de Regressão Linear é linear em seus parâmetros, com isso a mudança provocada em Y devido à variação por uma unidade de X deve ser constante;
- Multicolinearidade: duas variáveis são ditas correlatas, se a variação de uma implica em variação proporcional a outra. Variáveis independentes não devem ser correlatas entre si, do contrário trazem a mesma informação ao modelo;
- Resíduos devem ter esperança condicional zero: à diferença entre uma observação e uma estimativa atribui-se o nome de erro, ao conjunto de diferenças entre os dois é dado o nome de resíduos. Um modelo cujos resíduos possuem esperança condicional diferente de zero é enviesado;
- Endogeneidade (ausência de correlação entre resíduos e variáveis independentes): resíduos representam a parcela não explicada do modelo. Se

resíduos e variáveis independentes passam a ter correlação, então é possível usar os últimos para estimar os primeiros, o que impõe um erro fundamental;

- Homocedacidade (variância constante do resíduo): a variância do erro deve ser constante, independente do valor da observação;
- Ausência de autocorrelação nos resíduos: o erro de uma observação não deve influenciar o erro de observações subsequentes;
- Resíduos devem possuir distribuição normal.

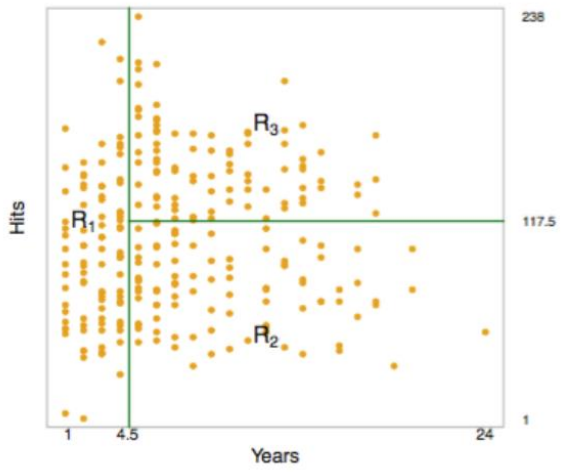
2.1.8. Regressão Não Paramétrica

A Regressão Não Paramétrica é uma outra metodologia utilizada para analisar o comportamento de uma variável dependente em função de outras independentes. Segundo Hazelton (2015), esta abordagem distingue-se da clássica, uma vez que não assume uma forma pré-determinada das variáveis nem hipóteses tão fortes sobre a relação entre elas. Ao invés disso, a função de estimação é derivada da proximidade entre os próprios registros. Por outro lado, modelos não paramétricos exigem amostras maiores para produzir boas estimativas. Dentre os vários métodos de Regressão Não Paramétrica, serão abordados, neste trabalho de formatura, algoritmos baseados em árvores (Árvores de Regressão e *Random Forests*) e *K-Nearest Neighbors* (K-Vizinhos Mais Próximos).

2.1.8.1. Árvore de Regressão

Uma Árvore de Regressão é definida como sendo uma árvore de decisão que aproxima funções contínuas em suas folhas. Seu princípio de funcionamento é segmentar o espaço preditor em diversas regiões simples. Quando uma nova observação é inserida dentro de um destes espaços, então o rótulo que lhe é atribuído equivale à média dos valores do conjunto de treino utilizados para compor esta região (STEORTS, 2017). A Figura 9 exemplifica a segmentação de um espaço preditor.

Figura 9 - Ilustração da segmentação de um espaço preditor em três subregiões



Fonte: Steorts (2017)

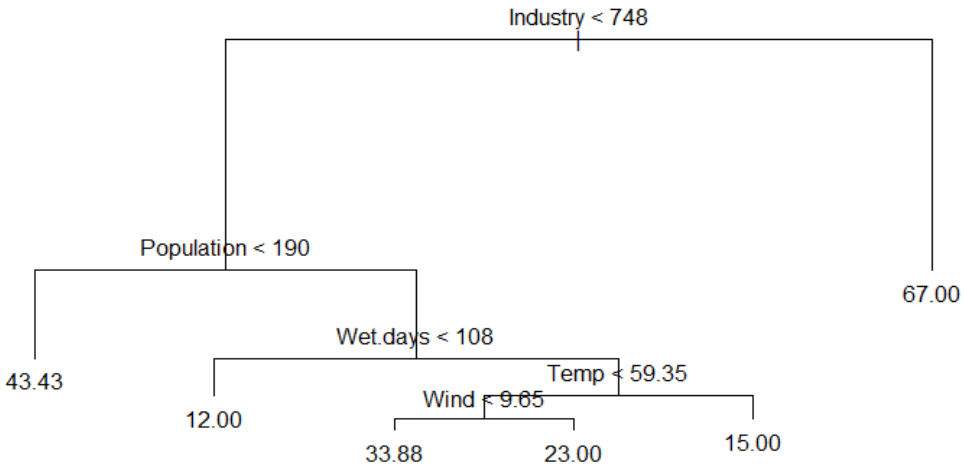
O algoritmo utiliza-se de uma árvore binária para particionar os exemplos baseados em seus atributos, de forma a ter estimativas da variável independente em suas folhas - nós que não levam a nenhum outro nó. A árvore é percorrida de cima para baixo, aplicando-se um teste lógico a cada nó interno que é atravessado. De acordo com o resultado do teste, continua-se o percurso seguindo ou à direita ou à esquerda. O procedimento é repetido até que se atinja um nó folha, cujo valor representa a média de todos os dados de treino que percorreram o mesmo caminho (STEORTS, 2017).

A construção da árvore também é feita de cima para baixo, particionando de maneira recursiva os dados de treino com a inserção de um teste lógico juntamente com o novo nó. O teste a ser realizado pode levar em conta uma variável contínua, categórica ou binário e dividirá a amostra de maneira a reduzir o Erro Quadrático Médio das subregiões, dado pela fórmula:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

onde Ŷi é a média do valor da variável dependente dentro de uma subregião e Yi os valores individuais observados. Caso a amostra passe o teste lógico, então o subcaminho seguirá para o nó à direita, do contrário seguirá para o nó da esquerda. O procedimento se repete até que os atributos tenham sido esgotados e as médias dos valores acumulados estejam nos nós folhas (HRUSCHKA, 2019). A Figura 10 demonstra o resultado deste processo no formato clássico de fluxograma.

Figura 10 - Fluxograma resultado de uma árvore de regressão



Fonte: Vala (2019)

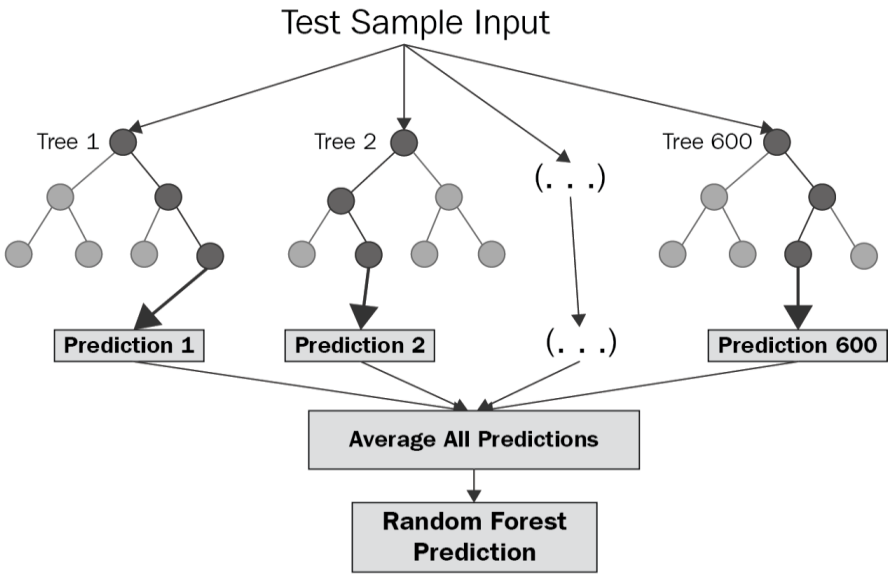
A grande vantagem de algoritmos de árvore é a sua interpretabilidade, uma vez que constroem regras intuitivas ao entendimento humano. No entanto, quando empregadas em problemas complexos, elas podem atingir alturas que a tornam indecifráveis. Além disso, ainda que sejam pouco sensíveis a valores atípicos (*outliers*), são um modelo altamente sujeito ao sobreajuste, pois sua construção é feita de maneira gulosa: a partição em um nó intermediário é feita levando-se em conta a minimização do Erro Quadrático Médio unicamente no nó em questão - e não na árvore como um todo. Consequentemente, se um novo conjunto de dados apresentar flutuações amostrais, a árvore produzirá estimativas com base nos dados anteriores, resultado em predições inexatas devido à variância (HRUSCHKA, 2019).

2.1.8.2.Random Forest

Segundo Chakure (2019), para mitigar esta variância, pode-se recorrer a um outro algoritmo baseado em árvore: *Random Forest* (Floresta Aleatória). Este algoritmo consiste em um aglomerado de árvores de decisão, cujos resultados individuais são agregados para contabilizar um resultado final do modelo através de uma média. A *Random Forest* reduz o erro por variância, uma vez que incluem apenas algumas das variáveis do conjunto de treino na modelagem de cada árvore. Com isso, um indivíduo é montado com apenas parte da informação em mãos. Se o desempenho individual diminui, por não terem sido concebidos com todo o conhecimento disponível, ele é compensado pela melhora do desempenho do conjunto. A estimativa com alta variância produzida por algumas das árvores pode ser mitigada pela estimativa daquelas com menor variância, resultando em previsões mais

consistentes e menos sujeitas a flutuações. É importante notar que a redução da variância não implica em aumento do viés. A Figura 11 demonstra como o resultado final do algoritmo é obtido.

Figura 11 - Ilustração de uma *Random Forest*

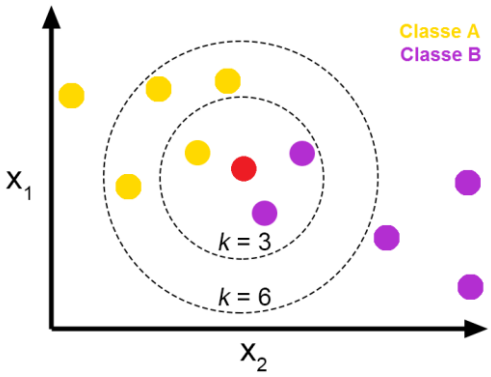


Fonte: Chakure (2019)

2.1.8.3.K-Nearest Neighbors

K-Nearest Neighbors (K-Vizinhos mais Próximos) é um algoritmo baseado em distância e semelhança. A ideia por trás é bastante intuitiva: atribuir ao valor de uma nova observação o valor da observação conhecida mais semelhante. Essa semelhança é calculada a partir da distância entre os atributos dos dois registros (MILLER, 2019).

Figura 12 - Ilustração do funcionamento do KNN em um problema de classificação.



Fonte: José (2018)

KNN é dito um algoritmo "preguiçoso", pois, em sua fase de treino, apenas assimila os registros sem efetuar nenhum tipo de cálculo. Ao tentar estimar o rótulo de uma nova

observação, no entanto, o modelo calcula a distância entre ela e todas aquelas armazenadas durante o treinamento. Como deseja-se obter a menor, é preciso conhecer todas as distâncias - o que pode tornar seu emprego computacionalmente custoso. Calculadas as distâncias, o rótulo da nova observação será igual à média entre os K registros mais próximos. No caso em que se define K=1, atribui-se à nova observação o mesmo rótulo do registro mais semelhante. (MILLER, 2019). A Figura 12 ilustra o método de funcionamento deste modelo para um problema de classificação, em que a decisão é feita com base em uma votação.

Segundo Miller (2019), as propriedades que mais influenciam no desempenho do modelo são:

- O número de vizinhos a serem levados em conta (K): não há um K ótimo para todos os problemas. Números pequenos tendem a provocar um sobreajuste, aumentando a variância para pequenas flutuações. Por outro lado, números elevados podem incluir na estimativa registros que não condizem com a observação, produzindo resultados discrepantes.;
- Distância empregada: há diversos tipos de distância que podem ser empregadas e as diferenças entre elas se realçam quando empregadas em espaços com alto número de dimensões. A distância de Minkowski incorpora boa parte das distâncias convencionalmente empregadas. Sua fórmula é dada por:

$$d(X,Y) = \sum_{i=1}^{n} |x_i - y_i|^p \quad ^{\frac{1}{p}}$$

- Quando P = 1, então tem-se a distância denominada Manhattan, que contabiliza a soma das diferenças em cada dimensão em valor absoluto. Para P = 2, a fórmula assume a clássica distância Euclidiana, que traduz distância em soma de quadrados;
- Escala/normalização: Por ser um algoritmo baseado em distância, o KNN é altamente sensível às escalas das variáveis. Dessa forma, o simples fato de alterar a unidade de uma das variáveis, de metros para centímetros por exemplo, acarretará novos resultados. Além disso, incluir variáveis de ordens de grandeza discrepantes - como 10 e 1.000.000, produzirá saídas enviesadas, em que a última variável terá uma relevância preponderante. Uma solução

para remediar este tipo de efeito é aplicar uma normalização para as variáveis, de forma que fiquem limitadas a um intervalo.

A grande vantagem do algoritmo é não possuir a etapa de treino propriamente dita. Nesta fase, ao invés de se construir um modelo, os registros são simplesmente armazenados. Por isso, ele pode ser evoluído constantemente, à medida em que se coletam novos dados. Em contrapartida, uma vez armazenada uma grande quantidade de registros de treino, o modelo pode tornar-se lento para produzir estimativas, uma vez que será necessário calcular as distâncias entre todos os casos assimilados e as novas observações. Além disso, o modelo não possui interpretabilidade. Enquanto é possível obter estimativas da relevância das variáveis em outras técnicas de modelagem, como Regressão Linear (através dos coeficientes) e Árvores de Decisão (critério de partição dos dados), o KNN assume que todos os atributos são igualmente importantes (MILLER, 2019).

2.1.9. AVALIAÇÃO DE MODELOS DE REGRESSÃO

Segundo KARBHARI (2018), há três métricas principais de se avaliar um modelo de regressão: Erro Absoluto Médio, Erro Quadrático Médio e Coeficiente de Determinação (r2).

O Erro Absoluto Médio é resultado da média entre as diferenças, em valor absoluto, entre valor real e valor previsto. É uma métrica bastante intuitiva que obtém o erro na mesma unidade da variável (KARBHARI, 2018). Ele é dado pela fórmula:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

O Erro Quadrático Médio, de forma análoga, calcula a média entre as diferenças entre valor real e estimado, no entanto ele computa o erro elevado ao quadrado, de maneira que grandes diferenças resultam em maiores penalizações (KARBHARI, 2018). Ele é obtido pela fórmula:

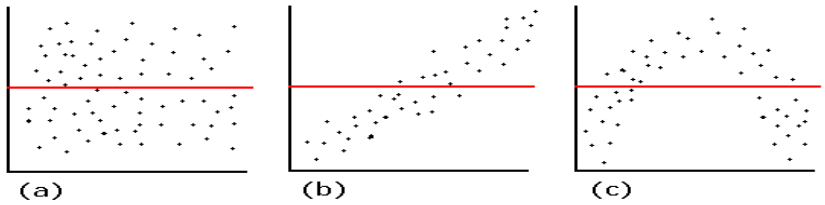
$$MSE = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

O Coeficiente de Determinação, por outro lado, não mede explicitamente a precisão do modelo em estimar uma variável. De forma simplificada, ele determina a porcentagem da variância de uma variável que é explicada por outra (MAKLIN, 2019).

Uma forma gráfica de avaliar o desempenho do modelo de regressão é através dos resíduos. O gráfico a ser avaliado consiste na estimativa do modelo no eixo X e os resíduos no eixo Y. Um bom gráfico não deve possuir *outliers* nem padrões de distribuição e/ou um formato (parabólico, linear, pilha). Além disso, ele deve ser igualmente distribuído, no eixo Y, em torno de zero (HAGERMAN, 2017). Em uma Regressão Linear, essas condições são exigidas pelas hipóteses de normalidade e homocedacidade. Nos modelos não paramétricos, que não possuem hipóteses, estes resultados são desejados.

A Figura 13 ilustra três exemplos de gráficos de resíduos: (a) segue uma distribuição aleatória - o que garante que os erros do modelo são independentes um do outro, enquanto (b) e (c) possuem formatos linear e parabólico - o que indica que o valor do erro é influenciado pelo valor da estimativa (DEO, 2016).

Figura 13 - Exemplos de gráficos de resíduo com e sem formato



Fonte: Deo (2016)

2.2. CENTRO DE DISTRIBUIÇÃO URBANA

Um centro de distribuição urbano, CDU, é uma solução logística a fim de trazer melhorias à região urbana em relação à entrega de mercadorias. A organização tradicional da distribuição de uma cadeia de produção é feita através de roteiros com caminhões, passando por vias urbanas, podendo elas ser congestionadas, estreitas, contribuindo para o aumento do tráfego urbano. Existem também diversas restrições de circulação nessas zonas, tanto de acesso, janela de tempo em que os veículos podem ficar parados, zonas ambientais. Tudo isso diminui a eficiência da logística de entregas, além de prejudicar o trânsito, poluição ambiental e sonora da cidade (OLIVEIRA; CORREIA, 2014).

Segundo Dablanc (2007), independente das características da cidade, gera demandas logísticas de acordo com as atividades comerciais locais e o importante é a existência dessas atividades em si. Isto é, existe uma “neutralização” do território urbano. Portanto, a distribuição de mercadorias sempre vai existir, de maneiras similares, não existe logística planejada para uma cidade ou região específica. Dablanc continua, diz que o transporte

urbano é caracterizado por veículos velhos, poluentes, eles devem se adaptar ao ambiente urbano, com restrições físicas como ruas estreitas, obstáculos físicos, trânsito, etc.

De acordo com Quak (2008), os CDU's são uma solução logística que visam melhorar a sustentabilidade das cidades através de mudanças estruturais físicas usadas pelo transporte de cargas na cidade. É de separar as atividades provenientes de fora da cidade das atividades internas. O estudo de CDU's não é recente, tendo sido iniciado no Reino Unido, depois na Holanda, Mônaco, sendo parte da política nacional em diversos países europeus. Esse tópico começou a se popularizar da década de 1990, tendo suas primeiras aplicações práticas na Itália, sendo o primeiro país a adotar uma estratégia nacional de CDU, seguindo-se pela Alemanha dois anos depois, e em 1993 pela França (TRANS et al., 2007).

Mesmo com diversos estudos e políticas na década de 1990, não houve muitos modelos práticos, e os existentes tiveram de ser encerrados por causa do baixo volume de mercadorias e baixo nível de serviço apresentado. (Browne et al., 2005). Atualmente, os CDU's são vistos como alternativa interessante devido ao crescimento da cultura sustentável ambiental, trazendo necessidade de soluções e mitigações que não prejudiquem o meio ambiente, ou que ao menos diminuam os danos em relação às soluções atuais.

A solução em si se caracteriza através de pequenos centros de consolidação de mercadorias, que recebem as mercadorias através dos caminhões vindos de outro centro maior ou diretamente das fábricas. Assim, os grandes veículos circulam pouco dentro da cidade, e veículos de pequeno porte fazem as entregas aos locais finais, sendo mais conveniente para o acesso físico e menos poluente.

3. METODOLOGIA

3.1. INTELIGÊNCIA ARTIFICIAL

3.1.1. Base de Dados

Um modelo de previsão de demanda apoia-se sobre dados históricos para produzir estimativas de eventos futuros. Por isso, a empresa forneceu dados de produção, venda, falta e vencimento de sete produtos diferentes referentes aos anos de 2017 e 2018 do centro de distribuição (CDD) da Praia Grande. As bases de dados contemplavam as seguintes informações de cada produto:

- Volume de chope vendido pelo CDD aos pontos de venda - denominado volume vendido. Equivale ao volume solicitado pelo ponto de venda e que se encontrava disponível no CDD;
- Volume de chope faltante na hora da venda - denominado volume de falta. Equivale ao volume que seria comprado pelo ponto de venda, porém que não pode ser vendido por não estar disponível no CDD;
- Volume de chope vencido no CDD - denominado volume vencido.
- Volume de chope recebido pelo CDD da cervejaria - denominado volume pedido. Equivale à soma dos pedidos dos pontos de venda, encaminhado à cervejaria para produção. Depois de entregues no CDD, será vendido aos pontos de venda. É o único valor conhecido anteriormente à produção e à venda e passível de ser incluído como variável independente no modelo de previsão.

Obtidos os dados, foi possível observar que os volumes vendidos e de falta não eram vinculados a uma unidade armazenada no centro de distribuição. Por exemplo, se foram recebidos dois barris iguais em dois dias diferentes e houve uma venda de barril, não se sabe se o que saiu do CDD foi o barril recebido no primeiro dia ou no segundo. Ou seja, a partir das informações das bases fornecidas, não é possível vincular a venda à unidade específica do produto que foi vendido. Como um modelo precisa ser alimentado com instâncias cíclicas e não foi possível estabelecer um controle a partir dos dados fornecidos, definiu-se, em conjunto com o representante da empresa, o agrupamento dos dados em intervalos semanais começando na segunda-feira e encerrando no domingo. Desta forma, toda venda e toda falta que acontecem até domingo de uma semana passaram a ser referenciadas ao volume recebido desde a última segunda-feira.

Feito este agrupamento, notou-se que o registro do volume vencido, no entanto, não corresponde à real data de vencimento dos produtos. Ao invés, ele é feito de maneira periódica, no qual um lançamento acumula os vencimentos de semanas, como pode ser observado na Tabela 1- as unidades de volume foram removidas e os valores multiplicados por uma constante para proteger o sigilo dos dados da empresa.

Tabela 1 - Trecho da base de dados consolidada

	Data	Produto	Vol Pedido	Vol Vendido	Vol Falta	Vol Vencido
286	2017-10-30	2	202.5	152.0	0.00	4.5
287	2017-11-06	2	117.0	101.0	0.00	705.5
288	2017-11-13	2	122.0	154.5	9.00	0.0
289	2017-11-20	2	92.0	102.5	0.00	27.5
290	2017-11-27	2	140.0	139.5	0.00	0.0
291	2017-12-04	2	150.0	157.5	0.00	0.0
292	2017-12-11	2	193.0	184.0	23.07	0.0
293	2017-12-18	2	273.5	271.0	0.00	0.0
294	2017-12-25	2	306.0	294.0	0.00	0.0
295	2018-01-01	2	220.0	209.5	30.50	0.0
296	2018-01-08	2	105.0	118.5	0.00	0.0
297	2018-01-15	2	131.5	146.0	29.21	0.0
298	2018-01-22	2	193.5	157.5	0.00	0.0
299	2018-01-29	2	162.5	106.5	0.00	0.0
300	2018-02-05	2	119.5	168.0	12.00	402.0
301	2018-02-12	2	120.0	113.5	0.00	0.0

Fonte: Autores

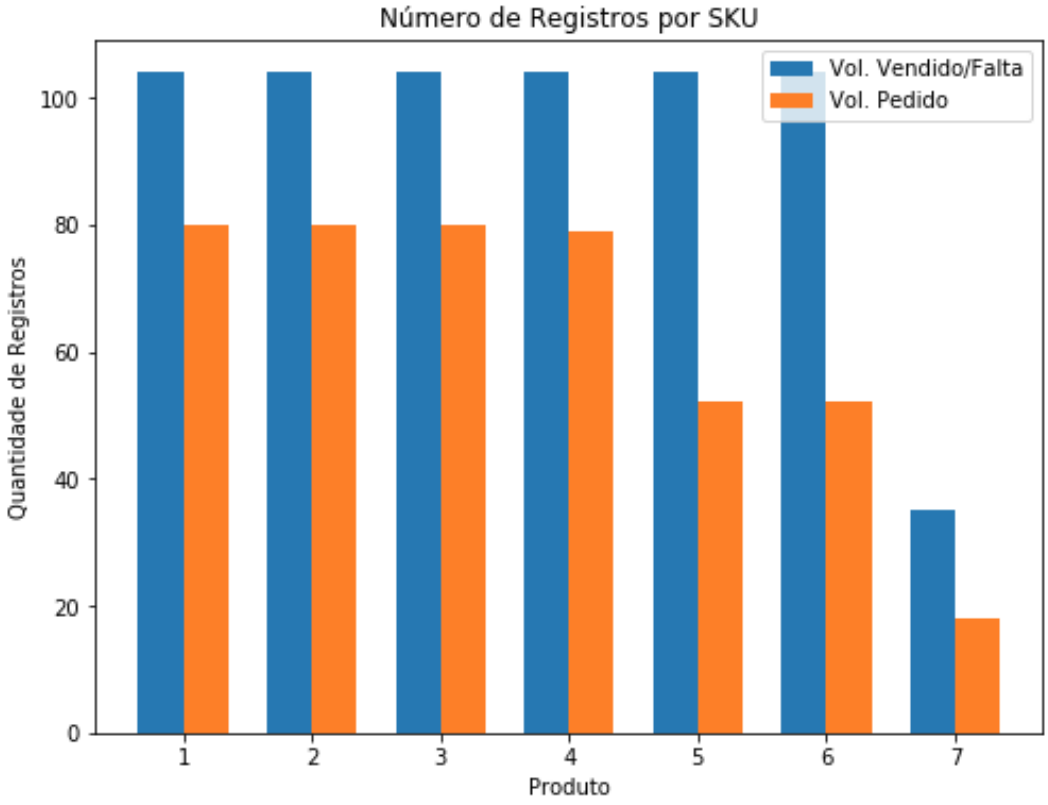
Desta forma, o grupo e o representante da empresa chegaram a um consenso de que esta variável não deveria ser incluída no modelo, uma vez que consiste em uma informação inverossímil. Por fim, foi acordado com o representante da empresa que a variável objetivo, que deveria ser estimada pelo algoritmo, deveria ser igual à soma entre os volumes entregues e faltantes, *i.e.* a soma entre o que foi, de fato, vendido e o que poderia ter sido, mas que não

foi devido à ausência do produto no centro de distribuição. A esta variável de saída, foi atribuído o nome de volume ideal.

3.1.2. Pré-Processamento

Tendo dados de dois anos agrupados em intervalos semanais e sabendo que um ano possui, em média, 52 semanas, cada produto deveria ter por volta de 104 registros na base consolidada. Porém, como pode ser verificado na Figura 14, o produto de número 7 possui quantidades significativamente menores de registros que os demais. Isso torna difícil a distinção entre representação e ruído da informação e, por isso, o produto foi excluído da modelagem. Os demais possuem a quantidade de observações esperadas de volumes vendidos e de falta. Observam-se cerca de 80 registros de volume pedido para três produtos e 50 para outros dois, que correspondem a, aproximadamente, um ano e meio e um ano de dados respectivamente.

Figura 14 - Gráfico da quantidade de registros por produto



Fonte: Autores

Como a quantidade de entradas do algoritmo deve ser igual à quantidade de saídas, têm-se duas possibilidades para prosseguir com a modelagem: remover os registros sem equivalências, ou preencher a informação faltante - da forma mais verossímil possível.

Para não descartar essa diferença de cerca de 20 e 50 observações, que corresponderia a uma perda de, respectivamente, 25% e 50% da amostragem, optou-se por preenchê-los. Para tanto, foi utilizado o algoritmo *K-Nearest Neighbors*, de forma a estimar os valores faltantes a partir de observações semelhantes (pedidos do mesmo produto em épocas próximas do ano). A combinação de parâmetros que gerou os melhores resultados, isto é, valores que acompanhassem as curvas sem, no entanto, apenas replicá-las foi a seguinte:

- K = 3, ou seja, estimativa igual à média entre os três registros mais semelhantes;
- Distância: Manhattan, dada pela fórmula (MUNOZ et al., 2009):

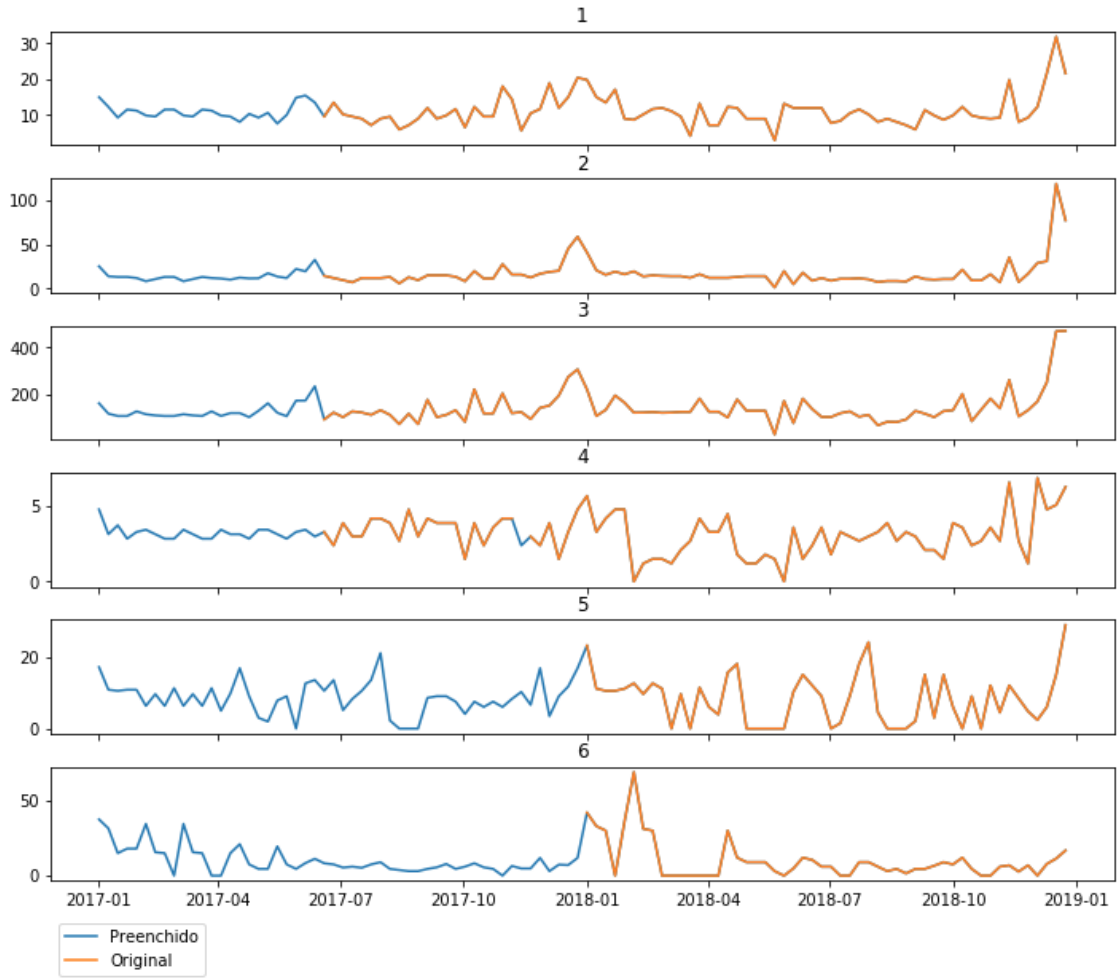
$$d(A,B) = |X_B - X_A| + |Y_B - Y_A|$$

- Normalização: Min-Max, dada pela fórmula (KEEN, 2017):

$$x' = \frac{x - min\,(x)}{(x)\; - min\,(x)}$$

As variáveis explicativas utilizadas foram o Índice de Consumo (ICON), o SKU e as variáveis temporais “dia do mês”, “mês”, “semana do ano” e “ano”. Na Figura 15, é possível observar os resultados do preenchimento dos dados faltantes em comparação com os observados. Novamente, a unidade do volume foi omitida e os dados multiplicados por um fator de correção para preservar o sigilo da empresa.

Figura 15 - Gráfico do volume pedido: valores observados e valores preenchidos



Fonte: Autores

Tendo a quantidade de dados de entrada igual à quantidade de dados de saída, prosseguiu-se, então, para a remoção de *outliers* através do Teste Z (Escore Padronizado). A ideia por trás do teste é descrever os registros através de sua relação com a média e o desvio padrão do conjunto. A variável é normalizada para ter média zero e desvio padrão igual a um, através da fórmula:

$$z = \frac{x - \mu}{\sigma}$$

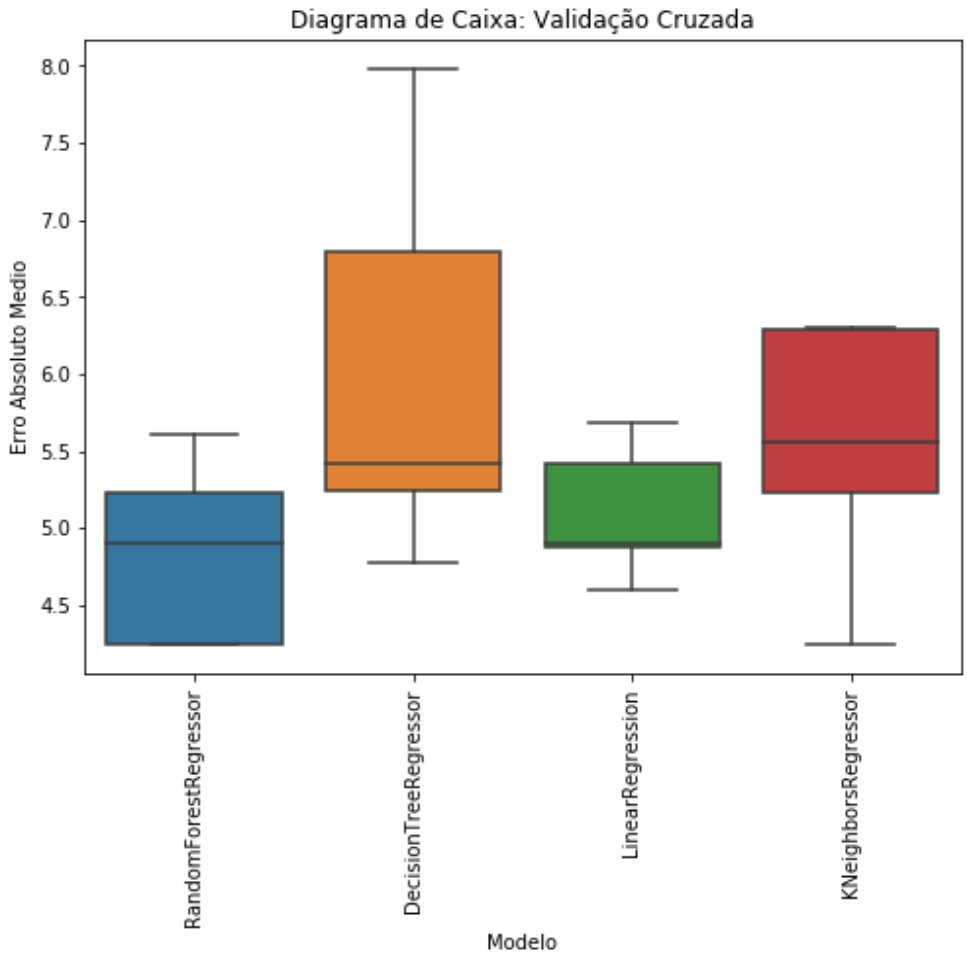
onde x é o valor observado, μ a média das observações e σ o desvio padrão (Sharma, 2018). Assim, valores que obtiveram um escore igual ou maior, em módulo a três, *i.e*, que são maiores ou menores do que a média por pelo menos três vezes o valor do desvio padrão, foram retirados da amostragem por serem valores anormais.

3.1.3. Validação Cruzada

Com os *outliers* removidos, parte-se para a validação cruzada, a fim de obterem-se os melhores algoritmos para o problema em mãos. Testaram-se os modelos de Regressão Linear, Árvore de Regressão, *Random Forest* e *K-Nearest Neighbors*, e a métrica de avaliação foi o Erro Absoluto Médio, uma vez que retorna o erro na mesma unidade da variável objetivo, o que facilita a interpretação dos resultados.

As variáveis de entrada incluem o volume pedido, o produto e variáveis temporais: dia do mês, número da semana no ano, mês e ano. A variável de saída constitui o volume ideal de recebimento. O subconjunto de treino acumulou 75% dos registros e o de teste 25%. Sobre o primeiro, foi aplicada a validação cruzada usando o método *K-Fold* com K = 5 e utilizando a normalização MinMax. Obtiveram-se cinco valores que, para melhor interpretação, foram exibidos em um diagrama de caixa, na Figura 16.

Figura 16 - Diagrama de Caixa com resultados da validação cruzada



Fonte: Autores

Observa-se que a Regressão Linear e a *Random Forest* obtiveram os melhores desempenhos, tendo produzido estimativas consistentes e mais próximas aos valores observados e serão incluídos na etapa de modelagem. Ainda que tenha produzido resultados piores, o *K-Nearest Neighbors*, por ter um funcionamento bastante diferente destes dois algoritmos, será incluído na etapa de modelagem para efeitos de comparação. A Árvore de Regressão, por ter tido desempenho pior do que outro algoritmo baseado em árvore, não será testada.

3.1.4. Aumento de Dados

Observando os gráficos da Figura 15, nota-se um pico de consumo entre os meses de Dezembro e Janeiro. Por ser uma área litorânea próxima à cidade de São Paulo, pode-se supor que uma alta nesta época do ano não represente apenas uma flutuação, mas que o consumo de chope esteja sob um efeito de sazonalidade.

Como foram fornecidos dados de apenas dois anos, este efeito não é tão facilmente percebido e pode não ser contabilizado pelos modelos durante os processos de construção de relações entre as variáveis - sobretudo para os modelos não paramétricos, que, segundo Hazelton (2015), precisam de amostragens maiores para produzirem boas estimativas. Para remediar este efeito, que levaria a um alto erro de previsão devido à variância, gerou-se um aumento dos dados. Para verificar se o desempenho melhora com o aumento da amostragem, serão testados modelos de previsão com a base original e com a base aumentada.

Este aumento de dados será feito com base nos registros observados e no Índice de Consumo (ICON), contabilizado pela bolsa de valores do Brasil (B3). O indicador, segundo a instituição, corresponde ao "desempenho médio das cotações dos ativos de maior negociabilidade e representatividade dos setores de consumo cíclico, consumo não cíclico e saúde". Incluiu-se esta variável externa, para que os dados aumentados não resultassem em uma sequência de cópias dos fornecidos - do contrário o problema deixaria de ser sobre previsão e passaria a ser sobre memorização. Como a B3 disponibiliza os dados desde 2007, foi possível gerar dados de dez anos.

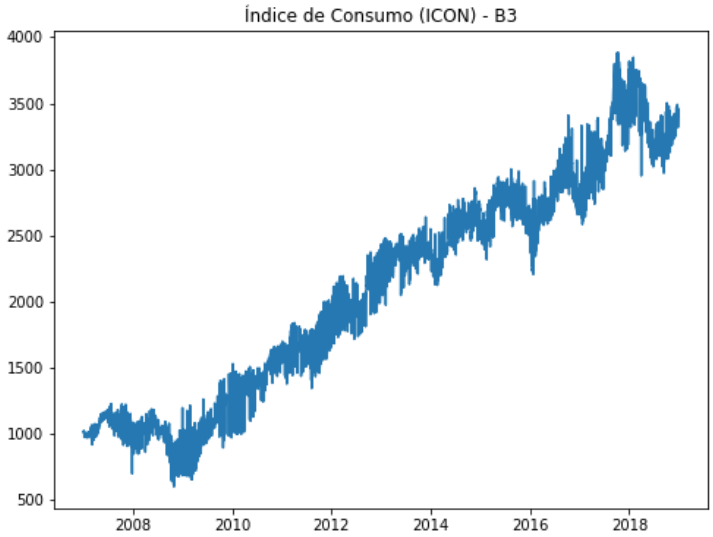
O aumento de dados foi realizado usando *K-Nearest Neighbors*, e a combinação de parâmetros que gerou os resultados mais compatíveis com as curvas foi:

- K = 2, ou seja: estimativa igual à média entre as duas observações mais semelhantes;
- Distância Euclidiana;

- Normalização Min-Max;

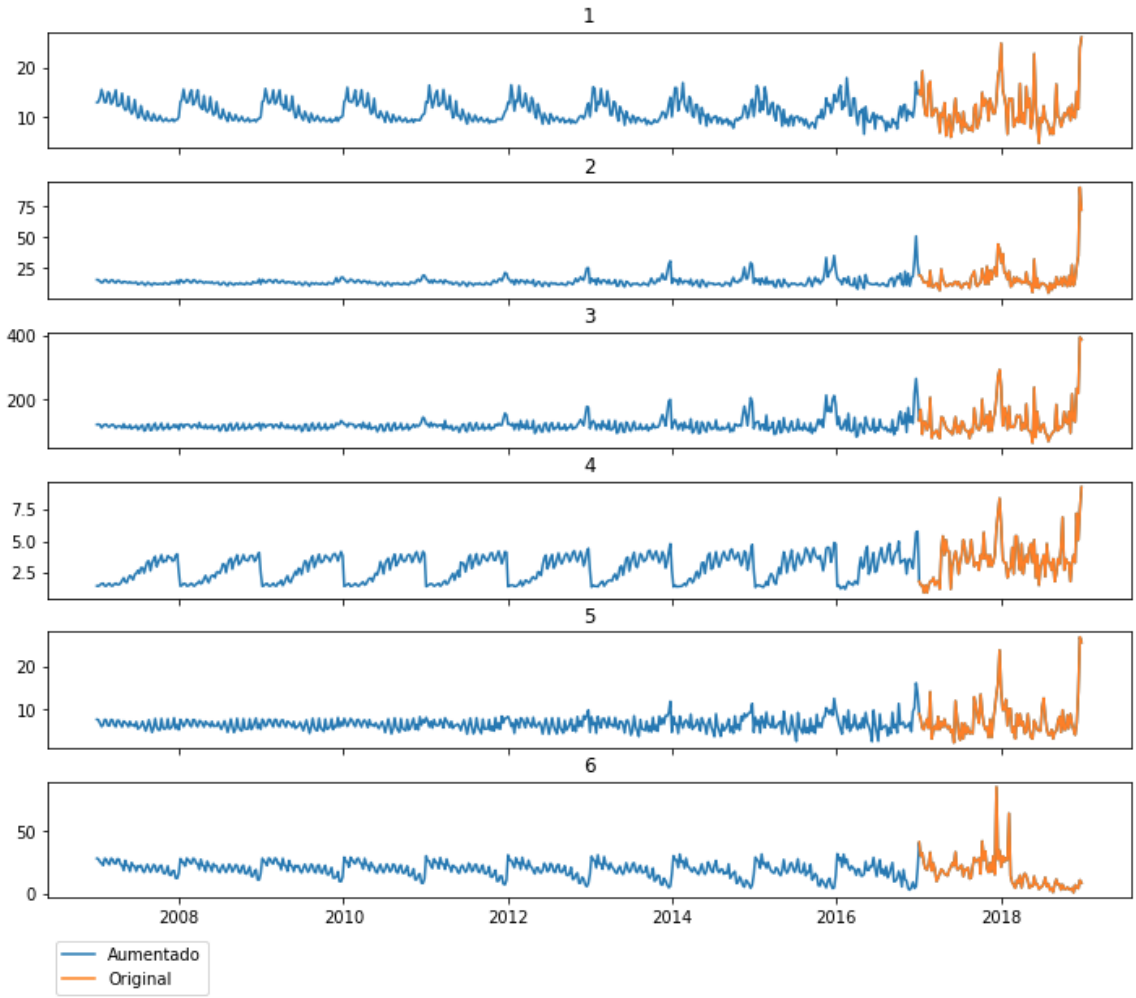
As Figuras 17, 18 e 19 a seguir ilustram respectivamente, a evolução do ICON e os resultados do aumento de dados para o volume ideal e o volume pedido.

Figura 17 - Gráfico Evolução do ICON (B3)



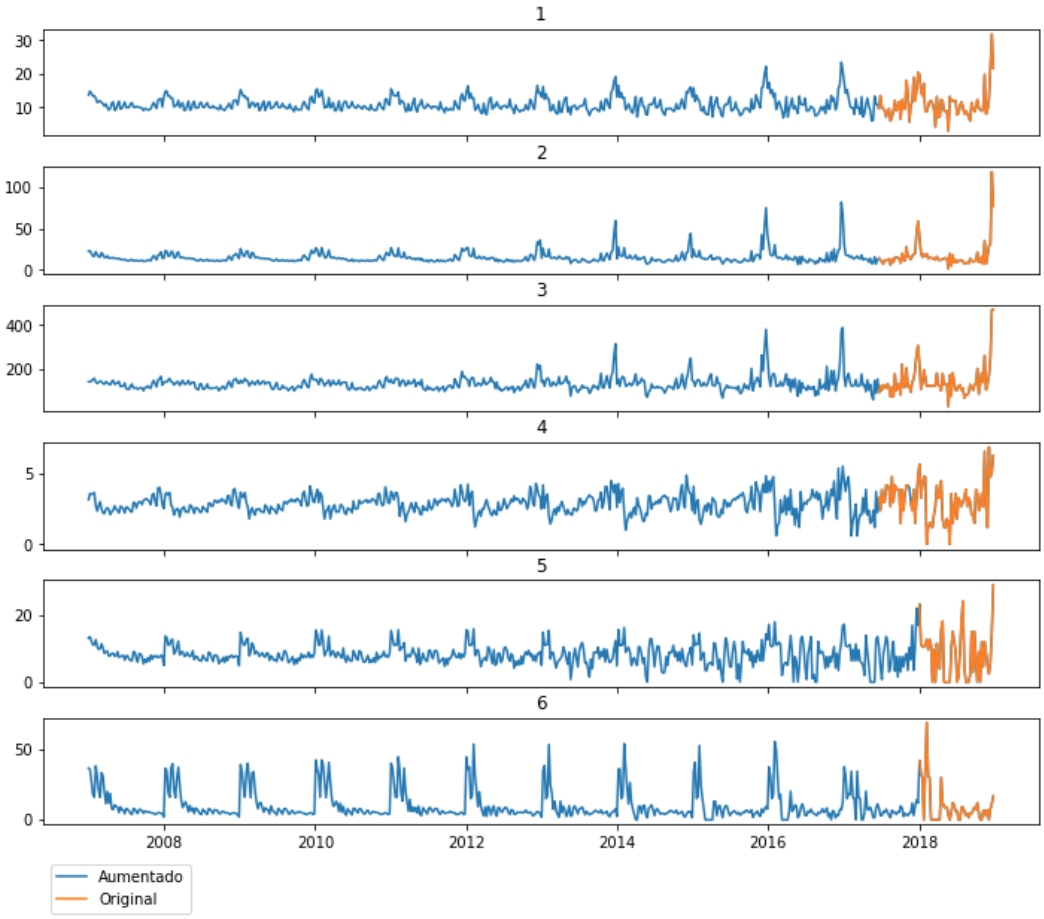
Fonte: Autores

Figura 18 - Gráficos ilustrando o resultado do aumento do volume ideal (por produto)



Fonte: Autores

Figura 19 - Gráficos ilustrando o resultado do aumento do volume pedido (por produto)

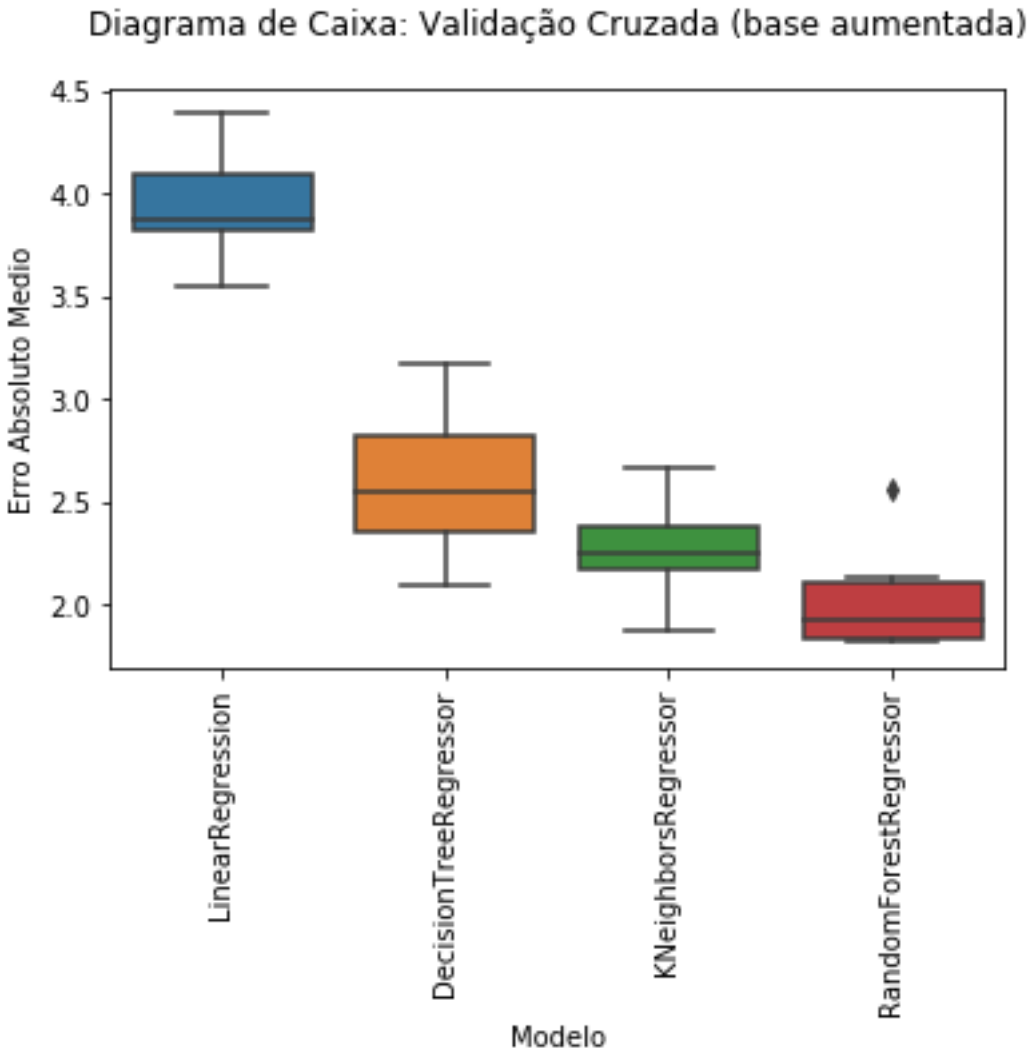


Fonte: Autores

3.1.5. Validação Cruzada: Base Aumentada

Para a base aumentada, foi feito o mesmo procedimento na validação cruzada: o subconjunto de treino acumulou 75% dos dados e o de teste os 25% restantes. O método *K-Fold* foi aplicado ao primeiro sob a normalização MinMax, tendo K=5. Para cada iteração, calculou-se o Erro Absoluto Médio entre valor real e valor estimado, que foram utilizados para compor o diagrama de caixa a seguir, ilustrada na Figura 20:

Figura 20 - Diagrama de caixa com resultados da validação cruzada para a base aumentada



Fonte: Autores

Com a base aumentada, nota-se uma melhoria considerável do desempenho dos modelos, com exceção da Regressão Linear, que não será testada nesta modelagem. Novamente, a Árvore de Regressão obteve desempenho pior do que o outro modelo baseado em árvore, por isso também não será utilizada na modelagem. *Random Fores*t e *K-Nearest Neighbors*, que obtiveram os melhores resultados, serão os algoritmos a serem testados.

3.2. CENTRO DE DISTRIBUIÇÃO URBANA

Ao se analisar os dados de entrega e os dados de falta da base de dados do CDD da Praia Grande, nota-se que a empresa já é bastante eficiente na previsão de demanda de chope. Com um algoritmo de Inteligência Artificial essa eficiência aumentaria ainda mais. Contudo, essa não é a única medida possível para redução de custos da empresa e impactos gerados,

podendo ser aliada ao estudo de custo de entrega desses produtos aos pontos de venda. Ou seja, o estudo do custo com pessoal e por quilometragem rodada.

Uma opção estudada pela empresa para grandes centros urbanos é a instalação de Centros de Distribuição Urbanos (CDUs) que estariam entre o CDD e o ponto de venda no ciclo de entrega. Os CDUs apresentam diversas vantagens já mencionadas na revisão bibliográfica, como redução de gastos com combustível e diminuição do tempo de entrega.

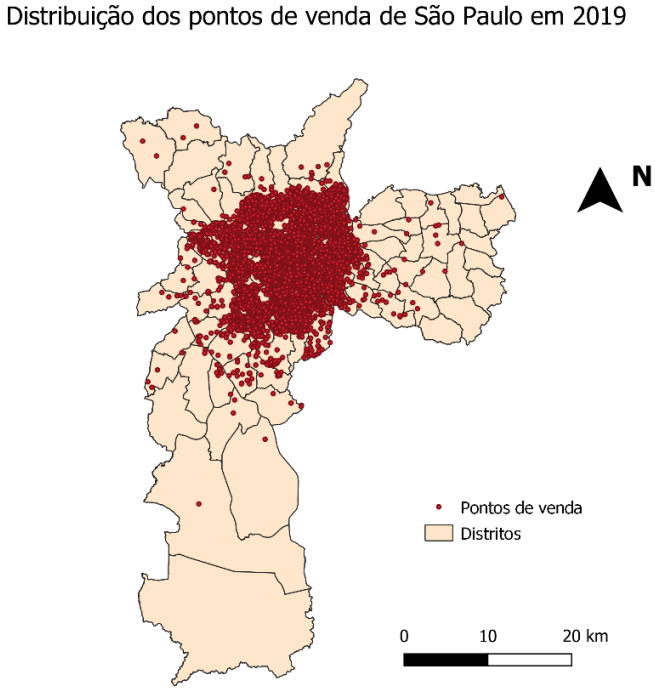
Esses CDUs são vantajosos, no entanto, quando são feitas entregas em grandes cidades, que possuem problemas de tráfego intenso, dificuldade de estacionamento e restrição de circulação de veículos. Entende-se, portanto, que um centro de distribuição urbano traria melhores resultados para a cidade de São Paulo que para a região da Baixada Santista. Desta forma, o estudo de aplicação do CDU será feito para as entregas de chope a partir do Centro de Distribuição da Mooca, em São Paulo. O objetivo, nesta fase, é a comparação dos custos de entrega com roteiros elaborados com e sem a existência de CDU.

3.2.1. Definição da região de interesse

Por definição, os centros de distribuição urbanos possuem uma área de abrangência menor que um CDD. Assim, é necessário que se defina a região de estudo.

Partindo do número total de pontos de venda em São Paulo georreferenciados nos anos de 2015 a 2019, dado fornecido pela empresa, pode-se encontrar o número de pontos de venda em cada distrito da cidade. Para isso, utilizou-se o *software Qgis* e o mapa de distritos da cidade obtido pelo portal *Geosampa* para plotar a distribuição de pontos em cada distrito, como demostrado na Figura 21 para o ano de 2019.

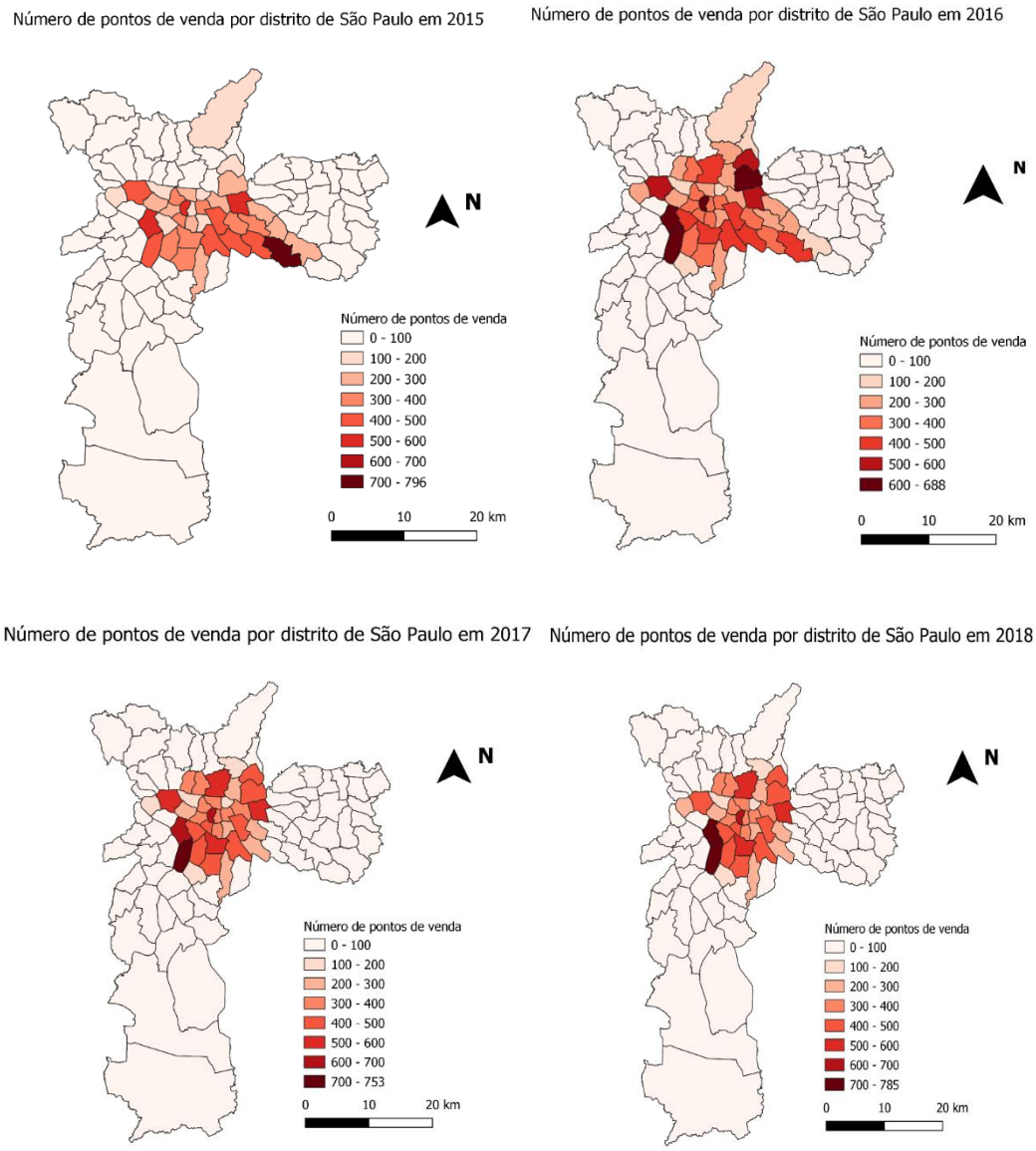
Figura 21 - Distribuição dos pontos de venda de São Paulo em 2019



Fonte: Autores

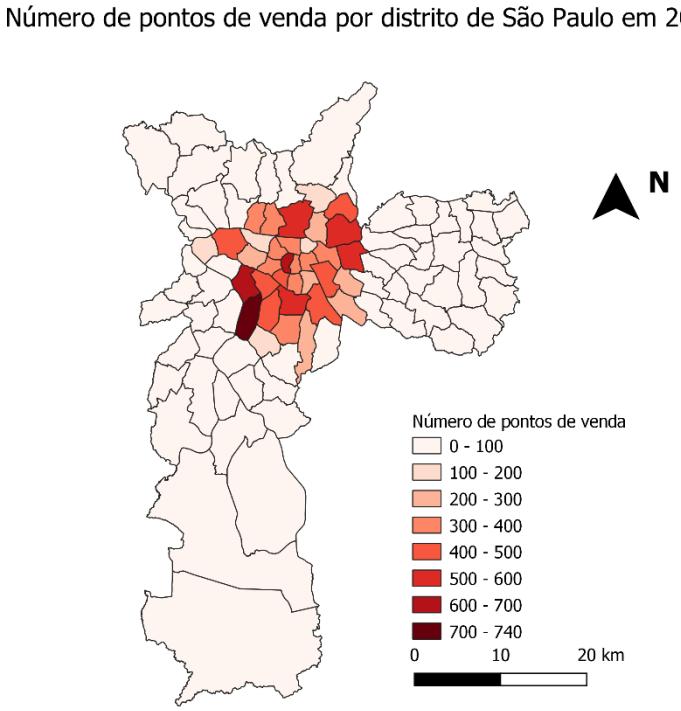
Utilizando a ferramenta de interseção das camadas vetoriais, geraram-se mapas de classificação por número de pontos de venda em cada distrito. Essa classificação foi feita por *pretty breaks* (intervalos iguais de valor arredondado). Os mapas gerados com o histórico das classificações são apresentados na Figura 22 e a classificação para 2019 Figura 23.

Figura 22 - Histórico de classificações por número de pontos de venda de São Paulo.



Fonte: Autores

Figura 23 - Classificação por número de pontos de venda de São Paulo em 2019

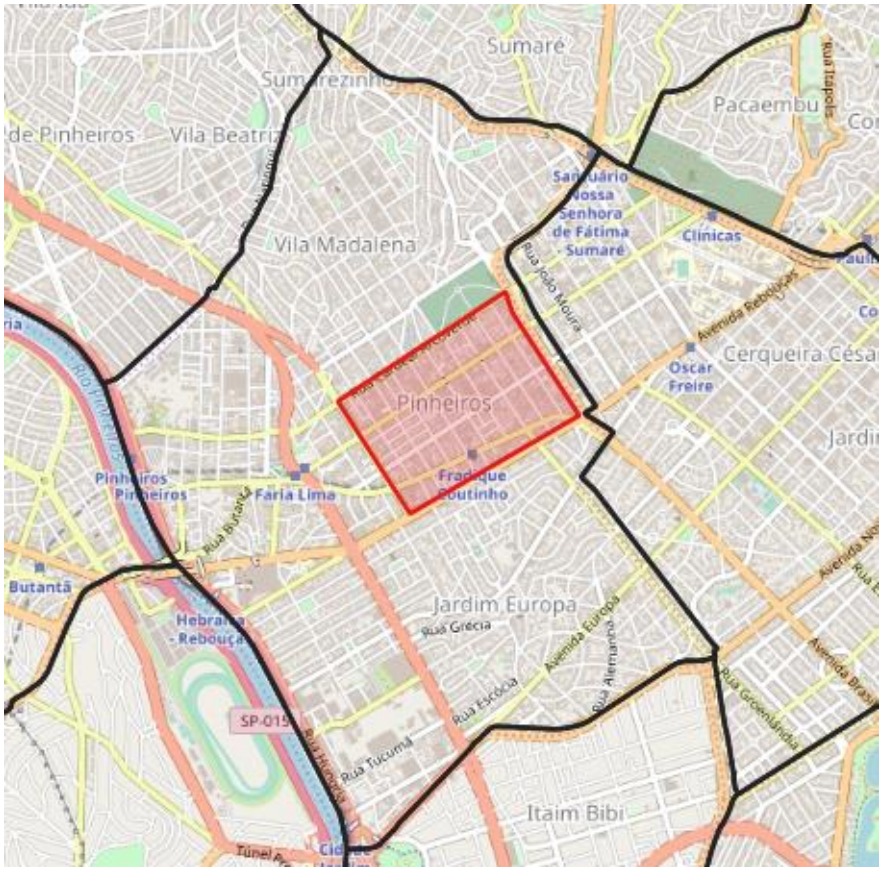


Fonte: Autores

Percebe-se que há um padrão de distribuição dos pontos atendidos pela empresa nestes últimos 5 anos. A maioria dos pontos de venda se localiza nas regiões mais centrais da cidade, com destaque para a região da Subprefeitura da Sé e entornos, o distrito do Itaim Bibi e o distrito de Pinheiros. Escolhe-se para a análise, então, o distrito de Pinheiros, pela frequência em que aparece em categorias maiores no histórico e sua classificação em 2019. O distrito de pinheiros apresenta duas áreas de grande presença de bares e restaurantes: a região da Vila Madalena e a da Fradique Coutinho.

Ainda assim, deve-se delimitar ainda mais a região de interesse. Escolhe-se a região da Fradique Coutinho, por fatores como a dificuldade de estacionamento e a concentração de bares. A região foi delimitada pela Avenida Rebouças, Rua Cardeal Arcoverde, Avenida Pedroso de Morais e a Rua Henrique Schaumann, como na Figura 24, que apresenta o distrito de Pinheiros em preto e a área escolhida em vermelho (mapa gerado com base fornecida pelo *OpenStreetMap*)

Figura 24 – Região de estudo escolhida.



Fonte: Autores

A região demarcada em vermelho abrange 132 pontos de venda, sendo que não se sabe quais vendem chope e quais pedem outros produtos da empresa. A Figura 25 mostra o recorte do mapa gerado com os pontos de venda.

Figura 25– Pontos de venda na região da Fradique Coutinho

Fonte: Autores

3.2.2. Definição do volume a ser entregue

Como dados de entrada, têm-se a base de dados de contagem do volume de chope entregue na Praia Grande em 2017 e 2018 (resumo da base apresentado na Figura 26 com unidades omitidas e fator multiplicador) e a base de Aderência da Praia Grande, que apresenta os pontos de venda georreferenciados.

Figura 26– Recorte da base de contagem da Praia Grande

Data Puxada	mês	Semana	CMN	Grupo Em	Cod. Prod	Volume
02/01/2018	jan-18	1	5780023	BARRIL	8284	17,4
02/01/2018	jan-18	1	5780023	BARRIL	8383	90
02/01/2018	jan-18	1	5780023	BARRIL	8276	3,3
02/01/2018	jan-18	1	5780023	BARRIL	87767	7,5
02/01/2018	jan-18	1	5780024	BARRIL	80374	11,1
02/01/2018	jan-18	1	5780024	BIB 12L	146605	14,4
02/01/2018	jan-18	1	5780024	BARRIL	133322	33
03/01/2018	jan-18	1	5780058	BARRIL	8284	8,7
03/01/2018	jan-18	1	5780058	BARRIL	8383	45
04/01/2018	jan-18	1	5780005	BARRIL	8284	8,7
04/01/2018	jan-18	1	5780005	BARRIL	8383	45
04/01/2018	jan-18	1	5780005	BARRIL	87767	7,5
08/01/2018	jan-18	2	5780022	BARRIL	8284	13,8
08/01/2018	jan-18	2	5780022	BARRIL	8383	70
08/01/2018	jan-18	2	5780022	BARRIL	8276	3,3
08/01/2018	jan-18	2	5780022	BARRIL	87767	7,5
09/01/2018	jan-18	2	5780024	BARRIL	8284	6,9
09/01/2018	jan-18	2	5780024	BARRIL	8383	35
09/01/2018	jan-18	2	5780025	BARRIL	80374	11,1

Fonte: Autores (adaptado da base de dados da empresa)

Com essas duas informações é possível localizar qual o volume pedido por cada cliente na Baixada Santista por semana. Do CDD da Mooca se tem a base histórica de vendas do CDD em volume total por mês para os anos de 2016 a 2019 (Figura 27 com unidades omitidas e fator multiplicador) e os pontos georreferenciados.

Figura 27– Parte da base de volume de chope de São Paulo

Realizado 2016	Total Operação	Total Tipo Cliente Volume									
Total Unidade Destino											
Produto	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
A	98,1	95,4	96,3	98,7	85,2	125,4	117,9	112,8	111,3	124,2	104,7	147,3
B	548,1	614,4	655,2	679,5	582,3	545,7	646,2	615,3	551,7	609	569,4	736,5
C	4565	4774,5	4598	5223,5	3862,5	3823,5	4686,5	4429,5	4646,5	4734	4385,5	6582,5
D	4,6	9,6	16,6	8	18	19,2	19,9	15,6	37,2	29,1	26,2	12,8
E	480,3	492,3	473,7	462,9	523,2	503,4	528,6	490,2	527,1	495,9	522,6	608,4
F	334,5	350,4	425,4	417,3	370,5	374,4	432,6	332,4	391,2	382,5	410,7	522,9
G	0,9	0,3	0	0	0	0	0	0	0	0	0	0
H	1,8	1,2	0,9	0	0	0	0	0	0	0	0	0
I	2,4	0,9	0,3	0	0	0	0	0	0	0	0	0
J	253,8	271,5	233,1	232,5	139,2	120	78,9	78,9	58,2	55,5	44,4	44,7
K	0	0	0	0	0	0	3,3	1,5	2,1	4,8	9,9	7,2
L	0	0	0	0	0	3,9	22,2	25,2	26,7	25,5	31,8	40,8
M	0	0	0	0	0	6	24,3	27,3	26,7	36,6	35,4	38,7
N	0	0	0	0	0	0	0	0	0	0	0	14,82
O	0	0	0	0	0	0	0	0	0	0	0	12,48
P	0	0	0	0	0	0	0	0	0	0	0	10,14
Total	6289,5	6610,5	6499,5	7122,4	5580,9	5521,5	6560,4	6128,7	6378,7	6497,1	6140,6	8779,24

Fonte: Autores (adaptado da base de dados da empresa)

Não é possível, no entanto, saber o quanto foi vendido para cada ponto de venda por semana especificamente em São Paulo. Assim, realizam-se estimativas.

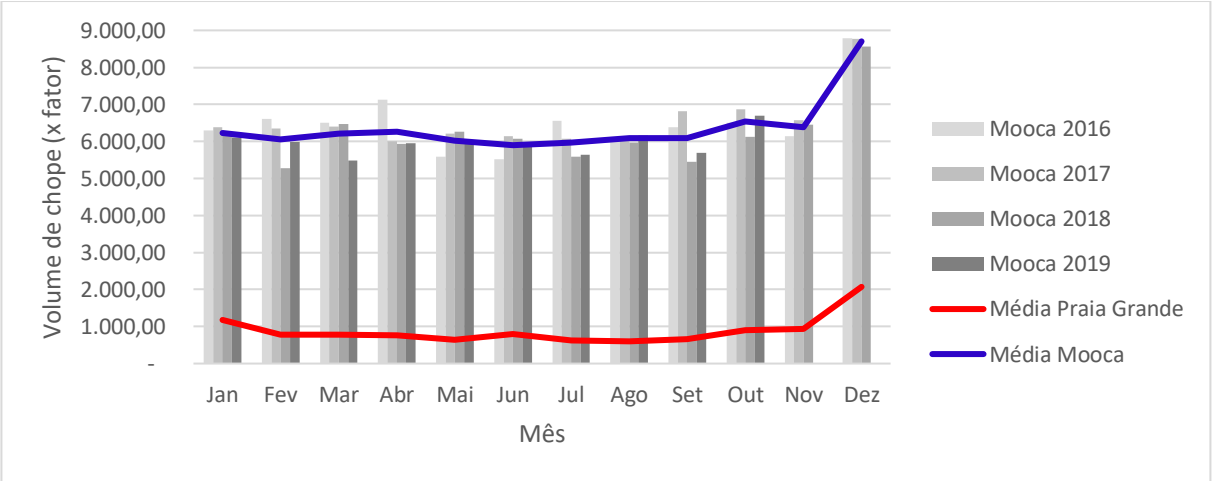
Para isso, deseja-se encontrar um padrão de distribuição de volumes entre pontos de venda da Praia Grande e atribuí-lo a São Paulo, respeitando as devidas proporções. Primeiramente, estuda-se o histórico de vendas agrupando todas as marcas de chope da empresa por mês para a Mooca e Praia Grande. A Tabela 2 apresenta as médias mensais obtidas pelo histórico e as médias por ponto de venda, multiplicados por um fator aleatório para preservação dos dados da companhia. E a Figura 28 apresenta esses mesmos valores por ano.

Tabela 2 - Pontos de venda na região Fradique Coutinho.

Mês	Média mensal		Média mensal por ponto de venda	
	Mooca	Praia Grande	Mooca	Praia Grande
Jan	6.235,94	1.172,60	0,44	0,37
Fev	6.051,87	782,00	0,43	0,25
Mar	6.217,35	773,80	0,44	0,24
Abr	6.255,68	762,40	0,44	0,24
Mai	6.019,79	645,70	0,42	0,20
Jun	5.898,73	804,30	0,41	0,25
Jul	5.965,69	631,30	0,42	0,20
Ago	6.095,70	600,50	0,43	0,19
Set	6.084,02	663,60	0,43	0,21
Out	6.544,18	903,30	0,46	0,28
Nov	6.387,37	937,40	0,45	0,30
Dez	8.702,61	2.072,70	0,61	0,65

Fonte: Autores

Figura 28– Comparativo do volume de chope entregue por mês.



Fonte: Autores.

Nota-se que, pode-se estimar o volume pedido pelos pontos de venda de São Paulo, a partir dos volumes pedidos na Baixada Santista multiplicando o valor encontrado por aproximadamente 1,72. Assim, dividindo-se o volume semanal pelo número de pontos de venda na Praia Grande, multiplicando-se pelo fator e pelo número de pontos de venda da Fradique Coutinho, tem-se o volume semanal de entrega na Fradique Coutinho, mostrado na Figura 29 (tabela sem unidades por confidencialidade).

Figura 29– Volume semanal entregue na Fradique Coutinho.

Região de estudo: Fradique Coutinho	
Número de pontos de venda	132
Volume na semana	627,29

Fonte: Autores.

O volume semanal, no entanto, não pode ser distribuído uniformemente entre os pontos, isso porque nem todos os pontos da Fradique Coutinho pedem chope e, dentre os que pedem, nem todos pedem toda semana. Uma atribuição deste tipo não retrataria as cargas reais a serem transportadas.

Analisando a base de dados semanal da Praia Grande, nota-se que diversos clientes não realizam pedidos em diversas semanas. A Figura 30 apresenta um recorte da base de dados semanal por cliente. Clientes que não pediram chope em determinada semana aparecem zerados na figura.

Figura 30– Recorte da base de dados de volume entregue por semana por cliente.

	SEMANA																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
5780023	118,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	57,4	0,0	0,0
5780024	58,5	41,9	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
5780058	53,7	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
5780005	61,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	63,3
5780022	0,0	94,6	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	41,8	0,0	0,0	0,0
5780025	0,0	72,9	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
5780032	0,0	7,5	0,0	58,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
5780049	0,0	0,0	55,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	56,7	0,0	0,0	0,0	0,0	0,0	0,0
5780042	0,0	0,0	40,5	0,0	0,0	0,0	0,0	0,0	59,6	0,0	0,0	44,3	0,0	0,0	0,0	22,8	0,0	0,0
5780053	0,0	0,0	27,6	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	45,4	0,0	0,0	0,0	0,0
5780051	0,0	0,0	26,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
5780056	0,0	0,0	57,6	0,0	39,3	94,3	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
5780067	0,0	0,0	0,0	120,6	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	14,4	0,0	0,0	0,0	0,0	0,0
5780054	0,0	0,0	0,0	53,7	0,0	51,6	0,0	0,0	0,0	0,0	0,0	0,0	0,0	47,8	0,0	0,0	0,0	0,0
5780013	0,0	0,0	0,0	57,4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
5780050	0,0	0,0	0,0	0,0	80,9	0,0	0,0	49,6	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
5780043	0,0	0,0	0,0	0,0	47,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
5780037	0,0	0,0	0,0	0,0	72,3	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	59,1	0,0	0,0	0,0	0,0
5780048	0,0	0,0	0,0	0,0	0,0	49,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
5780061	0,0	0,0	0,0	0,0	0,0	6,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
5780062	0,0	0,0	0,0	0,0	0,0	30,0	72,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
5780091	0,0	0,0	0,0	0,0	0,0	0,0	73,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
5780063	0,0	0,0	0,0	0,0	0,0	0,0	9,6	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
5780064	0,0	0,0	0,0	0,0	0,0	0,0	31,2	0,0	0,0	0,0	0,0	9,6	0,0	0,0	0,0	0,0	0,0	0,0
5780065	0,0	0,0	0,0	0,0	0,0	0,0	0,0	48,1	0,0	0,0	0,0	0,0	46,3	0,0	0,0	0,0	0,0	0,0
5780055	0,0	0,0	0,0	0,0	0,0	0,0	0,0	42,6	0,0	45,2	41,9	0,0	0,0	0,0	0,0	0,0	0,0	0,0

Fonte: Autores

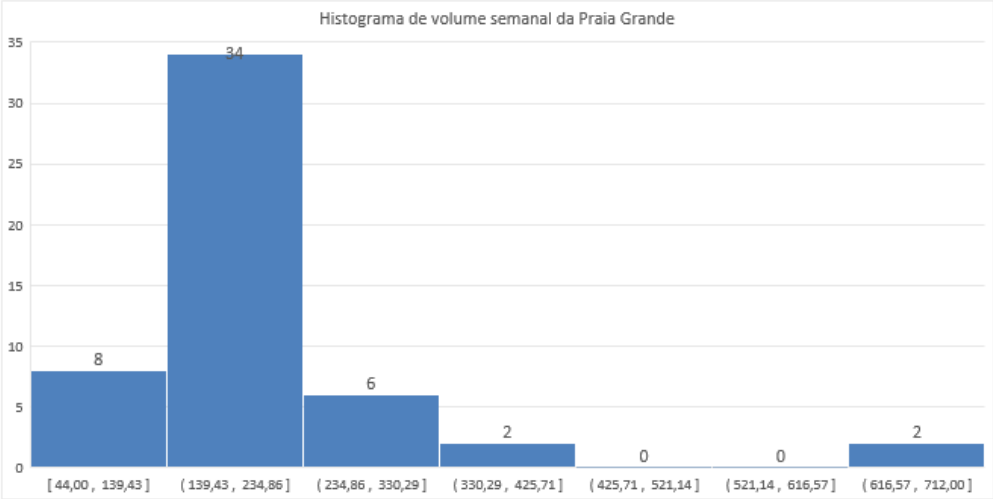
Para elaborar uma distribuição mais condizente com a realidade, estuda-se o padrão entre os clientes que, de fato, realizam pedidos em uma “semana padrão” de 2018. A Figura 31 apresenta o histograma de volume semanal entregue na Praia Grande com 7 classes

(intervalos). O número de classes foi definido a partir da regra de Sturges, a partir da seguinte equação:

$$k = 1 + log_2 n$$

Sendo k o número de classes e n o número de observações (HYNDMAN, 1995).

Figura 31– Histograma de volume semanal de entregas.



Fonte: Autores

Nota-se que existem semanas *outliers* que causariam distorções no estudo da "semana padrão". Assim, primeiramente, retiraram-se os dados *outliers* utilizando o método *z-score* excluindo dados que estivessem fora do intervalo $-3 < z < +3$.

Pelo estudo de volumes semanais, foram retiradas as últimas duas semanas de 2018. Pela soma do total pedido pelos clientes, retira-se 1 cliente da análise. Se ao invés do volume pedido por cada cliente, fosse feita uma análise da frequência de pedidos ao invés do volume, o mesmo cliente seria eliminado.

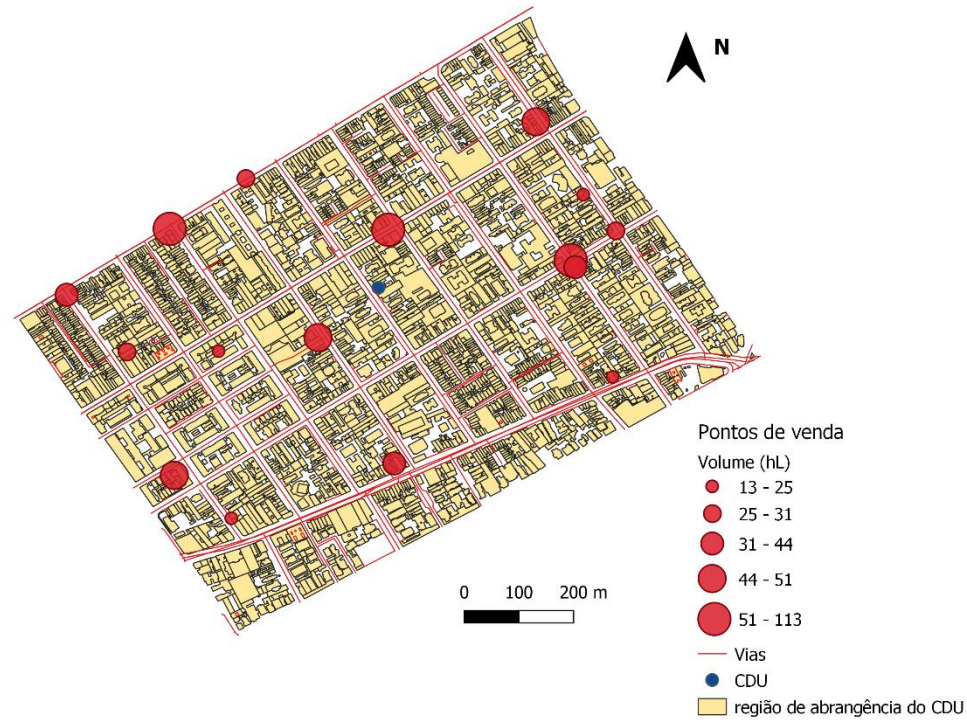
Calcula-se a média e desvio padrão semanal entre os clientes que, de fato, realizaram pedidos. Esses valores são 38,87 e 27,11 respectivamente (unidades omitidas). Assim, gerou-se uma distribuição normal aleatória com média e desvio padrão calculados para Praia Grande e aplicou-se essa distribuição para os pontos da Fradique Coutinho de forma que a soma dos pedidos fosse o mais próximo do valor total da semana apresentado na Figura 29.

Para isso, dos 132 pontos de venda na região, 116 tiveram seus pedidos zerados. Esses 116 valores corresponderiam aos estabelecimentos que não pedem chope especificamente e aqueles que pedem chope, mas não pediram na "semana padrão" analisada. Os outros 16 pontos de análise seguem a distribuição normal estipulada a partir dos pontos da Baixada

Santista. Com a distribuição aleatória gerada, o volume semanal calculado final foi de 688,78 para a região da Fradique Coutinho. Considerou-se que um “barril padrão” a ser transportado equivale a aproximadamente 1,98 vezes o volume.

Realiza-se também o cálculo do posicionamento do CDU. Alocou-se o CDU no centro de gravidade do sistema, ou seja, multiplicou-se o volume pedido pela latitude e longitude de cada ponto e dividiu-se pelo volume total, encontrando a latitude e longitude do CDU. Esses pontos estão representados na Figura 32. Quanto maior o volume pedido, maior o diâmetro do marcador (valores multiplicados por fator aleatório), sendo o ponto em azul o CDU.

Figura 32– Mapa de volume de pedidos por cada ponto de venda.



Fonte: Autores

3.2.3. Roteirização

Para a roteirização dos veículos, os volumes entregues são transformados em “barris padrão” (Tabela 3).

Tabela 3– Número de barris entregues por semana por cliente.

cod. cliente	latitude atual	longitude atual	volume entregue	número aproximado de barris
19457888	-23,5639	-46,6861	50,86	102
19476202	-23,561338	-46,681347	24,53	49
19484001	-23,5619875	-46,6807616	29,35	59
19439182	-23,5600343	-46,6821951	44,74	89
19497800	-23,5619668	-46,6848399	77,67	155
19448847	-23,5625114	-46,681571	113,14	226
19404655	-23,566367	-46,688674	50,59	101
19404050	-23,56105	-46,68739	31,47	63
19499599	-23,5631329	-46,6906028	42,27	85
19412152	-23,56414	-46,687881	13,48	27
19455875	-23,5671355	-46,6876513	15,72	31
19498661	-23,5646062	-46,6808207	20,46	41
19481389	-23,5619515	-46,6887564	64,11	128
19439918	-23,5626374	-46,6814896	44,15	88
19464756	-23,566153	-46,684733	38,02	76
19408268	-23,564156	-46,68952	28,22	56

Fonte: Autores

A roteirização foi feita com a utilização da ferramenta *online SimpliRoute.* O critério adotado é a diminuição do número de dias de entrega. Os cenários com maior ou menor número de veículos foram simulados separadamente pelo grupo.

Para rotas diretamente do CDD, foi considerada a entrega dos barris em caminhões toco de capacidade igual a 92 barris de chope. De acordo com a montadora Volvo (2019), um caminhão toco, ou semi-pesado, possui um eixo frontal simples e um eixo traseiro simples ou duplo e possui comprimento máximo de 14 metros. Sua velocidade e tempo de parada foram adotados com base no relatório de aderência da Praia Grande, sendo a velocidade média do veículo igual a 10,19 km/h e o tempo total médio parado no ponto de venda de 27 minutos.

Para rotas a partir do CDU, foi considerada a entrega com motos. A velocidade média das motos que fazem esse tipo de serviço é de 22km/h, de acordo com a pesquisa “Entregadores - Questões Socioeconômicas nos serviços por aplicativos” desenvolvida pela Fundação Instituto de Administração (FIA). Já o tempo de espera também foi obtido pela base de aderência da Praia Grande, considerando apenas o tempo de espera no ponto de venda ao invés do tempo total no ponto de venda, já que foi considerado que o tempo de descarga ou estacionamento, por exemplo, não eram significativos para a moto. Obtém-se, assim, um valor de 12 minutos de espera no ponto de venda para as motos.(JORNAL CORREIO ,

[s.d.]). A capacidade da moto foi adotada como 7 barris de chope, sendo que haverá um sidecar ou uma “carretinha” acoplada à moto como na Figura 33.

Figura 33 – “Carretinha” para entrega de gás.

Fonte: Empresa Motoprático

A ferramenta *SimpliRoute* não considera velocidades dos veículos, somente fatores de tráfego. Para aproximar o modelo da realidade, adotaram-se os fatores de tráfego 1 e 2 para toco e moto, respectivamente, numa escala de 0 a 3. Os tempos calculados pelo programa, no entanto, não foram considerados, somente a quilometragem percorrida. Os cálculos dos tempos foram feitos com a utilização das velocidades médias mencionadas.

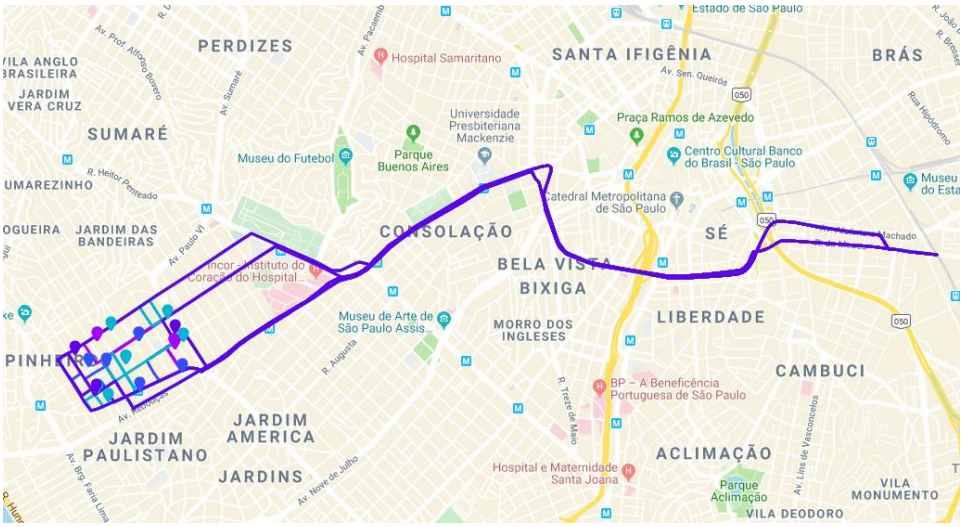
Nota-se que, existem alguns clientes da lista com pedidos que excedem a capacidade de carga dos veículos. Nesse caso, existem duas opções pela ferramenta: ou o entregador volta ao CDU para buscar o restante do pedido, sendo somados ao percurso “2 tempos de entrega”, por exemplo 24 minutos para o mesmo cliente, no caso da entrega com moto, ou realiza-se a entrega com 2 veículos simultaneamente e considerando apenas um “tempo de entrega’ no roteiro por cliente. Sendo assim, realizaram-se 7 simulações com esses dois casos:

- “1 TOCO” e “2 TOCOS em conjunto”: entrega dos pedidos agrupados por semana para todos os pontos de venda utilizando um ou dois caminhões toco, respectivamente, a partir do CDD da Mooca.
- “2 motos em conjunto” e “3 motos em conjunto”: entrega dos pedidos agrupados por semana para todos os pontos de venda utilizando duas ou três motos, respectivamente, fazendo as entregas juntas a partir do CDU.

- “1 moto – entrega diária”, “2 motos – entrega diária” e “3 motos – entrega diária”: entrega dos pedidos diariamente para cada ponto de venda. Dividiu-se o número de barris na semana por 5 e se considerou que o motoqueiro fará 1 visita por dia a cada cliente. Utilizam-se 1, 2 e 3 motos, respectivamente, a partir do CDU.

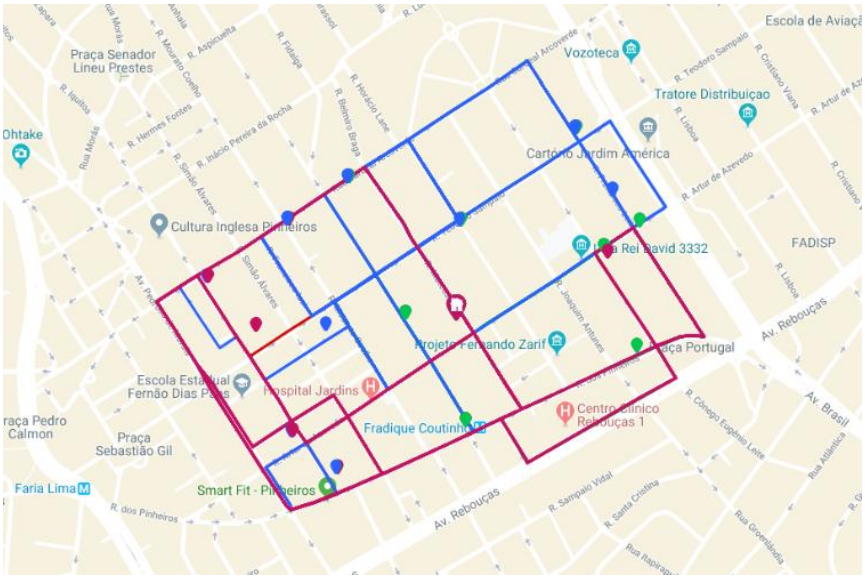
Exemplos de percursos criados a partir das roteirizações podem ser vistos nas Figuras 34, 35 e 36:

Figura 34 – Simulação com “1 TOCO”



Fonte: Autores

Figura 35– Simulação com “2 motos em conjunto”.



Fonte: Autores

Figura 36 – Simulação com “1 moto – entrega diária”.

Fonte: Autores

3.2.4. Análise Econômica

Ressalta-se, primeiramente, que o objetivo da análise de implantação de um CDU deste trabalho é uma avaliação preliminar de possíveis vantagens ou não desse sistema. Não se trata de um estudo completo, de forma que não foram considerados custos de implantação como construção do CDU, compra do terreno e compra de veículos. Os valores de entrega de barris também não são os reais da região, tendo, esses, sido estimados a partir de um padrão encontrado na Praia Grande. Assim, esta análise busca somente responder sobre a viabilidade de se investir em estudos mais aprofundados sobre a aplicação do modelo de CDU.

Para o cálculo dos custos de cada roteiro, foram considerados alguns fatores para ser possível comparar as diferentes soluções encontradas nas simulações.

- Distância – retirada diretamente das simulações

$$D_{total} = D_{trajeto} + D_{abastecer}$$

- Tempo de cada roteiro – retirada das simulações

$$T_{total} = \frac{D_{total}}{V_m} + T_{espera}$$

- Salário do motorista – tempo trabalhado vezes trabalho por hora

$$Salário = T_{total} \times \frac{Salário}{h}$$

- Combustível consumido – consumo médio do veículo vezes a distância vezes o preço do combustível

$$C_{combustível} = D_{total} \times \frac{Consumo\ Veículo}{km} \times Preço\ Combustível$$

O custo total se dará por:

$$C_{total} = (Salário + C_{combustível}) \times N_{veículos}$$

Além disso, de modo a contabilizar os impactos gerados para a cidade, foram levantadas as emissões de CO2 equivalente, que contabiliza CO2, CH4, N2O, HFC-125, HFC-134a, HFC143a, HFC-152a, CF4, C2F6, SF6. (MMA – Ministério Do Meio Ambiente, [s.d.]) bem como o ruído total gerado pelos veículos em meio urbano.

O cálculo de emissão de CO2 equivalente foi feito pela seguinte equação:

$$Emissão_{total} = \frac{Taxa\ CO_{2\ eq}}{km} \times D_{total}$$

O ruído em dB foi calculado através do ruído máximo permitido pelas normas brasileiras e multiplicada pela quantidade de veículos utilizados no roteiro. (Resolução CONAMA Nº 252, 1999)

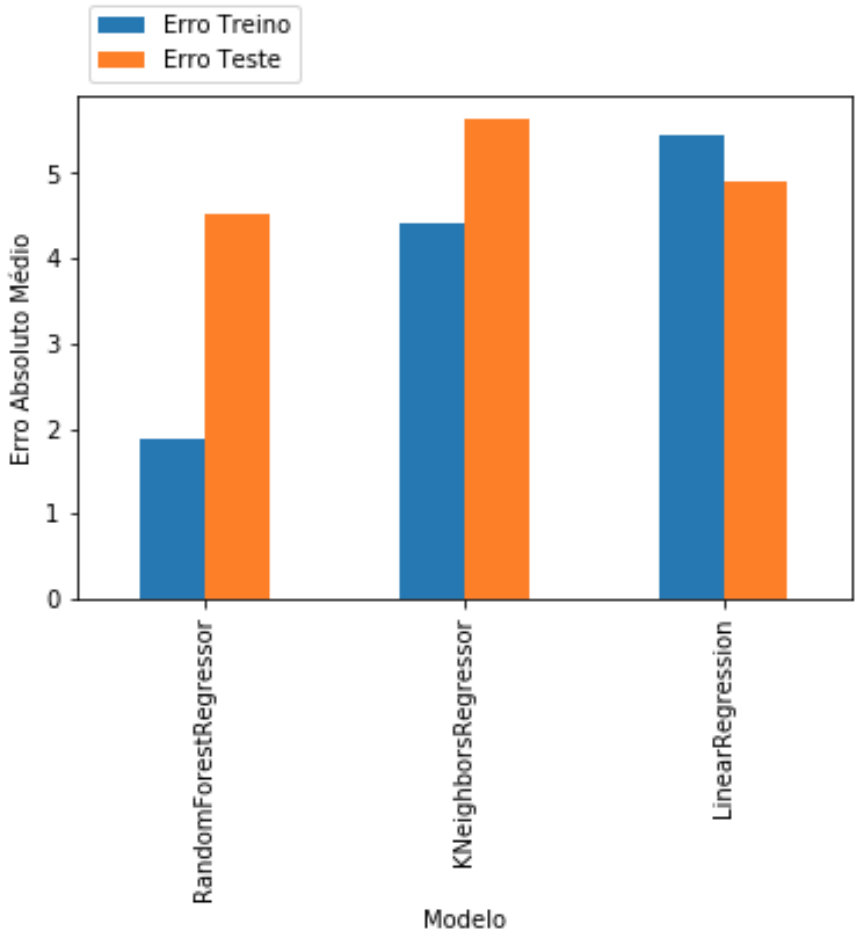
4. RESULTADOS

4.1. INTELIGÊNCIA ARTIFICIAL

4.1.1. Dados Originais

De acordo com o desempenho medido na validação cruzada, os modelos que melhor se adaptaram ao problema em questão foram Regressão Linear, *Random Forest* e *K-Nearest Neighbors*. Os algoritmos foram treinados com o subconjunto de treino e, em seguida, postos para estimar os valores do conjunto de treino. Na Figura 37, mostram-se suas performances na previsão de valores de treino e de teste segundo o Erro Absoluto Médio. Como visto na seção de Ajuste, um modelo bem balanceado produz estimativas precisas tanto para o subconjunto de treino quanto para o de teste; um modelo superajustado produz estimativas precisas apenas para o conjunto de treino e um modelo sobajustado produz estimativas imprecisas para ambos (SINGH, 2018).

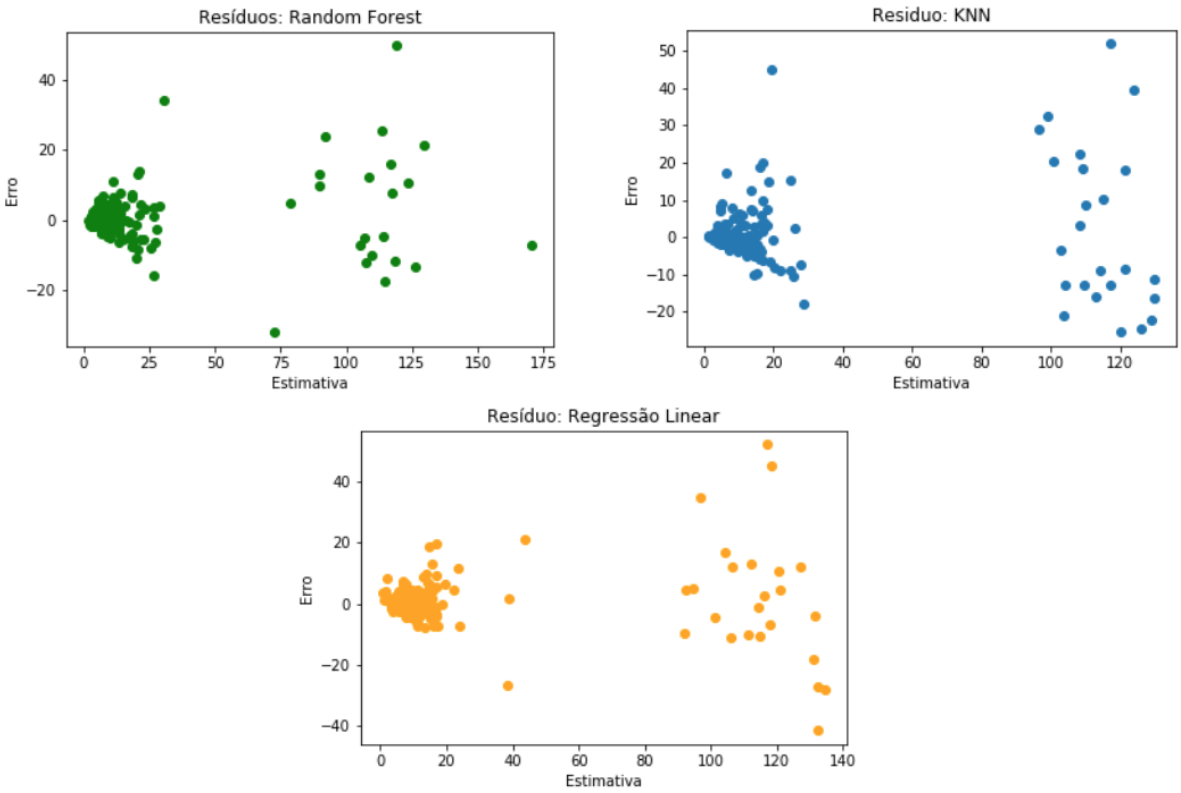
Figura 37- Verificação do ajuste



Fonte: Autores

Notam-se desempenhos melhores nos dados de treino do que nos dados de teste - a maior discrepância da *Random Forest* pode indicar um sobreajuste do modelo. Para ter um melhor entendimento dos resultados, analisam-se os resíduos, que podem ser visualizados na Figura 38.

Figura 38- Resíduos das regressões



Fonte: Autores

Nos três gráficos, observam-se dois padrões de distribuições dos pontos: uma mais numerosa e mais concentrada à esquerda e outra menos expressiva e mais dispersa à direita. O formato cônico que as distribuições apresentam indicam ausência de homocedacidade, *i.e* a variância do erro não é uniforme e ele aumenta, em valor absoluto, conforme aumenta o valor da estimativa. Isto é suficiente para invalidar, estatisticamente, o modelo de Regressão Linear e afirmar que os outros dois não produzem resultados confiáveis quando a estimativa ultrapassa 30.

No entanto, pode-se contrastar as previsões dos algoritmos com o modelo empregado pela empresa e verificar se houve melhora no desempenho. Essa comparação será feita através do Erro Absoluto Médio entre estimativas e valor observado (volume ideal) e, para o modelo da empresa, entre volume ideal e o volume pedido.

Tabela 4- Comparação entre erros absolutos médios

	Erro Absoluto Médio (hL)
Modelo Atual	6,61
Random Forest	4,80
KNN	5,90
Regressão Linear	5,80

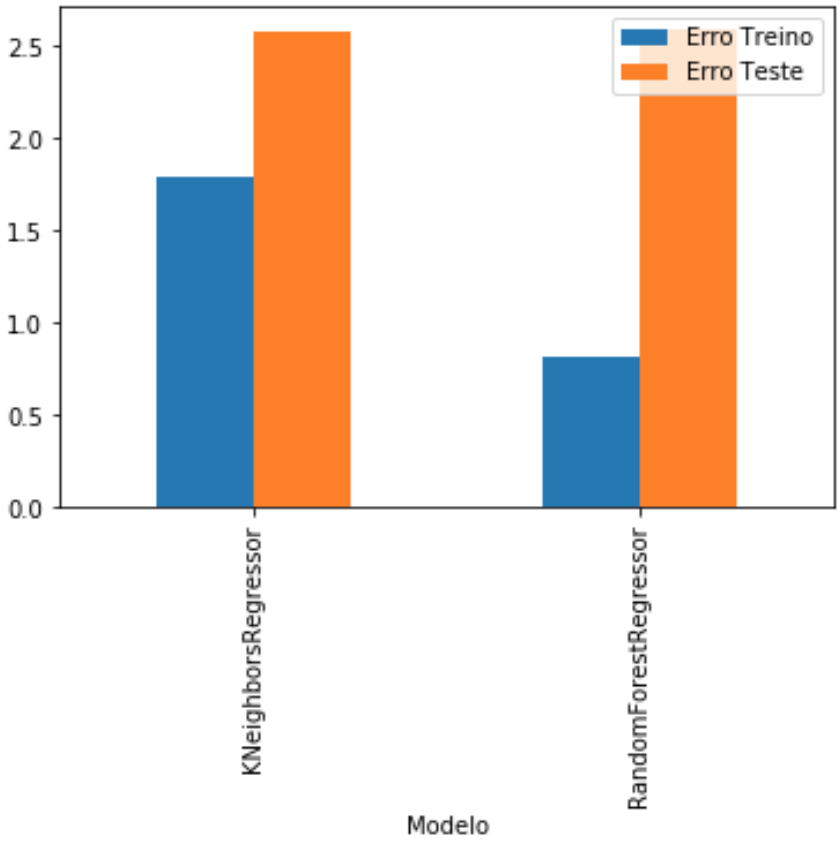
Fonte: Autores

Comparando o modelo empregado pela empresa com o algoritmo de menor erro, tem-se uma diferença de 1,81 hL. Ou seja, a estimativa do modelo de Aprendizado de Máquina ficou, na média, 1,81 hL mais próxima do volume ideal do que a previsão tradicionalmente feita pela empresa. Como há um pedido de produto por semana e seis produtos analisados, acertam-se 10,86 hL ou 1.086L a mais o volume de consumo a cada semana.

4.1.2. Dados Aumentados

Para os dados aumentados, os algoritmos que se mostraram mais aderentes, de acordo com a validação cruzada, foram *K-Nearest Neighbors* e *Random Forest*. Assim, dividiu-se a base aumentada em subconjuntos de treino e teste e realizou-se a etapa de treinamento. Em seguida, verificaram-se seus ajustes medindo o Erro Absoluto Médio entre as estimativas do modelo e os dados de treino e teste, ilustrados na Figura 39.

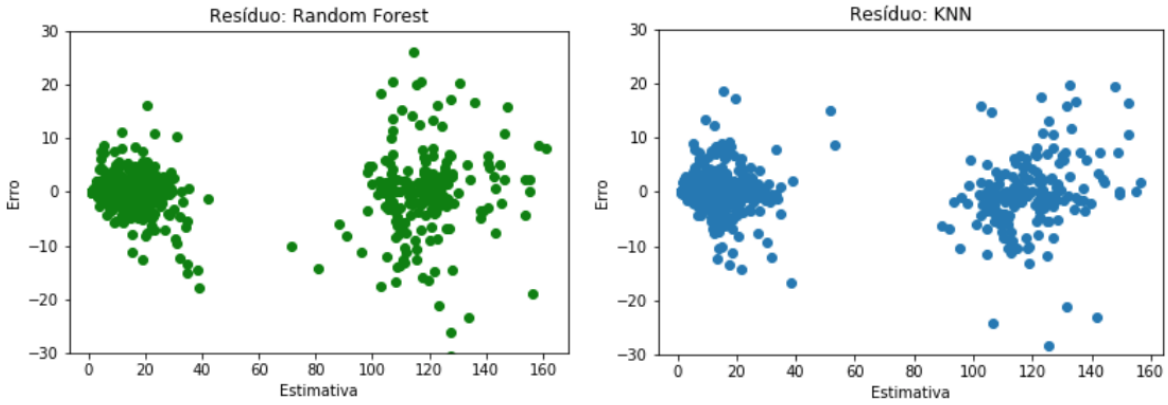
Figura 39- Verificação do ajuste após processo de aumento de dados



Fonte: Autores

Nota-se uma redução do erro, o que implica em melhora da performance dos modelos. A discrepância entre erro de treino e de teste da *Random Forest* ainda sugere, no entanto, um sobreajuste. Para ter um melhor entendimento dos resultados, analisam-se seus resíduos, vistos na Figura 40.

Figura 40- Resíduos das regressões após processo de aumento de dados



Fonte: Autores

Diferentemente dos gráficos da Figura 38, observa-se que estes não possuem o aspecto divergente à medida em que as estimativas aumentam e a quase totalidade dos pontos encontram-se no intervalo [-20, 20] do eixo Y. Isso indica que as estimativas possuem homocedacidade e os resultados são confiáveis, uma vez que o erro entre os valores real e estimado não aumenta conforme a estimativa aumenta. Com uma amostragem maior, os modelos puderam obter um melhor entendimento das relações entre as variáveis no cenário em que a variável objetivo assume um valor maior.

Comparando seus desempenhos com o modelo atual da empresa através do Erro Absoluto Médio, tem-se:

Tabela 5- Comparação entre erros absolutos médios após processo de aumento de dados

	Erro Absoluto Médio (hL)
Modelo Atual	6,61
Random Forest	1,88
KNN	1,76

Fonte: Autores

Nota-se uma diferença de 4,85 hL por pedido de produto nas estimativas entre o modelo da empresa e o modelo de Aprendizado de Máquina com melhor desempenho. Sabendo que há seis produtos pedidos por semana, então o último produziu estimativas que ficaram, em média, 29,1 hL ou 2.910 L mais próximas do volume ideal de consumo.

4.2. CENTRO DISTRIBUIÇÃO URBANA

A partir da roteirização feita, foram feitas análises de custo relativos a cada roteiro simulado. A Tabela 6 mostra os resultados, cada linha representa uma simulação.

A tabela mostra em destaque os principais resultados, em que o vermelho representa maior custo e o verde menor. Nota-se que em todas as dimensões analisadas, de custo financeiro e de emissão de poluentes e sonora, os caminhões toco tem uma performance inferior. 2 motos em conjunto e uma moto diária tiveram destaque nos custos, bem como em emissão de CO2 equivalente mensalmente.

Nesse estudo preliminar, em questão dos veículos, a instalação de um CDU se mostra vantajosa, principalmente no caso da utilização de uma moto, apresentando menor custo

mensal, menor emissão de poluentes e menor poluição sonora. Assim a região e a empresa podem se beneficiar com essa mudança, utilizando veículos compatíveis com as entregas urbanas, diminuindo a poluição local, tanto atmosférica quanto sonora.

Os valores considerados, foram retirados de websites com informações de 2019, em relação a preços de salários (Salários 2019). O mesmo foi feito para o preço da gasolina e diesel (PREÇO DOS COMBUSTÍVEIS - GASPASS, 2019). Para o cálculo do tempo, foi utilizada velocidade média, retirada da pesquisa da CET de monitoração de mobilidade (Mobilidade No Sistema Viário Principal Pesquisa De Monitoração Da Mobilidade, 2017). As taxas de emissão de poluentes de cada veículo foram extraídas de um relatório do IPEA. (RIBEIRO DE CARVALHO, 2011).

Tabela 6 - Resultados de Custos Simulações

Simulação	Entrega	Nº dias	Km percorrida por veículo (km)	Km para abastecer o CDU (km)	Km percorrida para entregas (km)
1 TOCO	semanal	4	305,879	NA	305,879
2 TOCOS em conjunto	semanal	3	171,523	NA	343,046
2 MOTOS em conjunto	semanal	4	191,127	21,4	382,254
3 MOTOS em conjunto	semanal	3	129,24	21,4	387,72
1 MOTO	diária	5	384,555	21,4	384,555
2 MOTOS	diária	5	207,625	21,4	415,25
3 MOTOS	diária	5	150,2	21,4	450,6
Simulação	**Tempo de espera total (h)**	**Tempo de percurso (h)**	**Tempo total (h)**	**Salário (R$/h)**	**Custo Salário (R$)**
1 TOCO	9,9	30,018	39:55:03	R$ 10,48	R$ 418,34
2 TOCOS em conjunto	7,65	16,832	24:28:57	R$ 10,48	R$ 256,58
2 MOTOS em conjunto	21,6	8,688	30:17:15	R$ 6,18	R$ 187,18
3 MOTOS em conjunto	14,6	5,875	20:28:28	R$ 6,18	R$ 126,53
1 MOTO	45	17,48	62:28:47	R$ 6,18	R$ 386,12
2 MOTOS	28	9,438	37:26:15	R$ 6,18	R$ 231,36
3 MOTOS	20	6,827	26:49:38	R$ 6,18	R$ 165,79
Simulação	**Preço combustível (R$/L)**	**Consumo combustível (km/L)**	**Custo combustível (R$)**	**Custo total/funcionário**	**Custo/mês**
1 TOCO	R$ 3,81	4,1	R$ 283,95	R$ 702,28	R$ 2.809,12
2 TOCOS em conjunto	R$ 3,81	4,1	R$ 318,45	R$ 575,02	R$ 4.600,19
2 MOTOS em conjunto	R$ 4,36	28	R$ 59,50	R$ 246,67	R$ 1.973,37
3 MOTOS em conjunto	R$ 4,36	28	R$ 60,35	R$ 186,88	R$ 2.242,53
1 MOTO	R$ 4,36	28	R$ 59,85	R$ 445,98	R$ 1.783,91
2 MOTOS	R$ 4,36	28	R$ 64,63	R$ 295,99	R$ 2.367,96
3 MOTOS	R$ 4,36	28	R$ 70,13	R$ 235,92	R$ 2.831,10

Simulação	Poluição (kg CO2 eq/km)	Poluição Total (kg CO2 eq)/mês	Poluição Sonora (dB)
1 TOCO	1,28	1566,12	103
2 TOCOS em conjunto	1,28	1756,4	206
2 MOTOS em conjunto	0,07	107,04	198
3 MOTOS em conjunto	0,07	108,56	297
1 MOTO	0,07	107,68	99
2 MOTOS	0,07	116,28	198
3 MOTOS	0,07	126,16	297

Fonte: Autores

5. CONSIDERAÇÕES FINAIS

O emprego de modelos de Aprendizado de Máquina cresceu em ritmo acelerado nos últimos anos devido à queda do custo computacional e ao aumento de informação produzida e armazenada, que podem ser usadas para treinar modelos. Com isso, computadores passaram a ser capazes de automatizar tarefas que, até então, apenas humanos poderiam fazer (TANNER, 2018).

Neste trabalho de formatura, foi possível utilizar dados históricos de comercialização de um produto para criar modelos de Aprendizado Supervisionado e estimar o consumo de observações desconhecidas. Ao comparar seus resultados com o modelo convencional de estimativa de demanda adotado pela empresa, pode-se observar um melhor desempenho, sobretudo quando empregada com uma maior amostragem.

Em relação ao estudo logístico, de efeito preliminar, ou seja, seu propósito foi de mostrar a oportunidade da criação de novos modelos de entrega urbana, que podem beneficiar regiões urbanas e a própria empresa tem potencial de diminuir os custos de entrega. Além disso, esse modelo de entrega urbana, visou entregas semanais, que pode ser em geral utilizados para produtos perecíveis. Como estudo preliminar, ele não abrange outros custos existentes, como de instalação e operação da estrutura física do CDU, porém ele conseguiu mostrar que existem alternativas possivelmente viáveis que não utiliza um caminhão de grande porte em vias congestionadas dentro dos bairros, e que podem ser estudadas mais detalhadamente pela empresa.

6. REFERÊNCIAS

ALPAYDIN, E. Introduction to Machine Learning. 2nd edition. Cambridge: **The MIT press**. 2010.

BELLONI, L. Como o desperdício de alimentos afeta o Brasil e o seu bolso | **HuffPost Brasil**, 2018. Disponível em: <https://www.huffpostbrasil.com/2018/04/08/como-o-desperdicio-de-alimentos-afeta-o-brasil-e-o-seu-bolso_a_23375621/>. Acesso em: 01 jun. 2019.

BHATTACHARYYA, I. Linear Regression Part 1. **Medium**, 2018. Disponível em: <https://medium.com/coinmonks/linear-regression-bf5141ce9ac8>. Acesso em: 20 set. 2019.

BORGES, T. Até 14 horas de trabalho e 80 km pedalados por dia: conheça os entregadores por aplicativo - **Jornal CORREIO**, 2019. Disponível em: <https://www.correio24horas.com.br/noticia/nid/ate-14-horas-de-trabalho-e-80-km-pedalados-por-dia-conheca-os-entregadores-por-aplicativo/>. Acesso em: 10 set. 2019.

BRONSHTEIN, A. Train/Test Split and Cross Validation in Python - **Towards Data Science,** 2017. Disponível em: <https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6>. Acesso em: 20 set. 2019.

BROWNE, P. M. et al. **Urban Freight Consolidation Centres Final Report**. [sl; s.d].

CHAKURE, A. Random Forest Regression - **Towards Data Science,** 2019. Disponível em: <https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f>. Acesso em: 20 set. 2019.

CHOLLET, F.. Deep Learning with Python. 1st Edition. Greenwich: **Manning Publications**. 2017.

DABLANC, L. Goods transport in large European cities: Difficult to organize, difficult to modernize. **Transportation Research Part A: Policy and Practice**, [s. l.], v. 41, n. 3, p. 280–285, 2007.

DEO, S. R Tutorial : Residual Analysis for Regression. **Analytics Pro,** 2016. Disponível em: <http://analyticspro.org/2016/03/05/r-tutorial-residual-analysis-for-regression/>. Acesso em: 20 out. 2019.

Entenda as diferenças entre os caminhões “toco” e “trucado” e conheça as vantagens de cada modelo. **Blog Canal Volvista**, 2019. Disponível em: <https://seminovosvolvo.com.br/blog-canal-volvista/artigo/entenda-as-diferencas-entre-os-caminhoes-toco-e-trucado>. Acesso em: 10 out. 2019.

HAGERMAN, I. Residual Plots Part 1 — Residuals vs. Fitted Plot - Data Distilled - **Medium**, 2017. Disponível em: <https://medium.com/data-distilled/residual-plots-part-1-residuals-vs-fitted-plot-f069849616b1>. Acesso em: 20 set. 2019.

HAZELTON, M. L. Nonparametric Regression. **ScienceDirect Topics,** 2015. Disponível em: <https://www.sciencedirect.com/topics/psychology/nonparametric-regression/pdf>. Acesso em: 20 set. 2019.

HRUSCHKA, E.. **Aprendizado de Máquina - Regressão.** 2019. 69 Slides.

HYNDMAN, R. J. **The problem with Sturges’ rule for constructing histograms**. [s.l: s.n.].

JOSÉ, I. KNN (K-Nearest Neighbors) #1. **Towards Data Science**, 2018. Disponível em: <https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d>. Acesso em: 20 set. 2019.

K-Means Clustering in Python. **Mubaris**, 2017. Disponível em: <https://mubaris.com/posts/kmeans-clustering/>. Acesso em: 20 set. 2019.

LI, T. et al. An introduction to reinforcement learning with AWS RoboMaker. **AWS Machine Learning Blog**, 2019. Disponível em: <https://aws.amazon.com/pt/blogs/machine-learning/an-introduction-to-reinforcement-learning-with-aws-robomaker/>. Acesso em: 20 set. 2019.

Mais de meia tonelada de alimentos são descartados em supermercados da Barra - **Jornal O Globo**, 2017. Disponível em: <https://oglobo.globo.com/economia/defesa-do-consumidor/mais-de-meia-tonelada-de-alimentos-sao-descartados-em-supermercados-da-barra-21577730>. Acesso em: 1 jun. 2019.

MILLER, M. The Basics: KNN for classification and regression - **Towards Data Science,** 2018. Disponível em: <https://towardsdatascience.com/the-basics-knn-for-classification-and-regression-c1e8a6c955>. Acesso em: 20 set. 2019.

MMA-Ministério do Meio Ambiente **PNIA-PAINEL NACIONAL DE INDICADORES AMBIENTAIS** Disponível em: <http://www.mcti.gov.br/index.php/content/view/326751.html>. Acesso em: 15 out. 2019.

Mobilidade no Sistema Viário Principal Pesquisa de Monitoração da Mobilidade.

Motoboy - Salários 2019 - **Quanto Ganha, Piso Salarial e Periculosidade**. Disponível em: <https://www.salario.com.br/profissao/motoboy/>. Acesso em:15 out. 2019.

MUNOZ, et al. Manhattan Distance. **ScienceDirect Topics**, 2009. Disponível em: <https://www.sciencedirect.com/topics/computer-science/manhattan-distance>. Acesso em: 20 set. 2019.

OLIVEIRA, L. K. De; CORREIA, V. de A. Proposta metodológica para avaliação dos benefícios de um centro de distribuição urbano para mitigação dos problemas de logística urbana. **Journal of Transport Literature**, [s. l.], v. 8, n. 4, p. 109–145, 2014.

Preço dos combustíveis - **Gaspass**, 2019. Disponível em: <https://precodoscombustiveis.com.br/pt-br/city/brasil/sao-paulo/sao-paulo/3830>. Acesso em: 30 out. 2019.

Procon apreende mais de 1 tonelada de alimentos em supermercados de Manaus | **Amazonas | G1**, 2019. Disponível em: <https://g1.globo.com/am/amazonas/noticia/2019/05/22/procon-apreende-mais-de-1-tonelada-de-alimentos-em-supermercados-de-manaus.ghtml>. Acesso em: 1 jun. 2019.

QUAK, H. **Sustainability of Urban Freight Transport**. [s.l: s.n.]. Disponível em: <http://repub.eur.nl/pub/11990/>

REALI COSTA, A. H. **Agentes Inteligentes**. 2019. 52 Slides.

Resolução CONAMA nº 252 de 07/01/1999. [s.d.]. Disponível em: <https://www.normasbrasil.com.br/norma/resolucao-252-1999_96054.html>. Acesso em: 30 out. 2019.

RIBEIRO DE CARVALHO, C. H. **EMISSÕES RELATIVAS DE POLUENTES DO TRANSPORTE URBANO**..

Salario Caminhoneiro 2019 - **Veja Quanto Ganha e Piso Salarial**. Disponível em: <https://www.salario.com.br/profissao/caminhoneiro-cbo-782505/>. Acesso em: 1 out. 2019.

SHARNA, N. Ways to Detect and Remove the Outliers - **Towards Data Science**, 2018. Disponível em: <https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba>. Acesso em: 20 set. 2019.

SHUGA, D. 5 Reasons why you should use Cross-Validation in your Data Science Projects. **Towards Data Science**, 2018. Disponível em: <https://towardsdatascience.com/5-reasons-why-you-should-use-cross-validation-in-your-data-science-project-8163311a1e79>. Acesso em: 20 set. 2019.

SINGH, S. Understanding the Bias-Variance Tradeoff - **Towards Data Science,** 2018. Disponível em: <https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229>. Acesso em: 20 set. 2019.

SOMA, S. Linear Regression Assumptions. **Medium**, 2018.. Disponível em: <https://medium.com/datadriveninvestor/linear-regression-assumptions-f2252b8e2912>. Acesso em: 20 set. 2019.

STEORTS, R.. **Tree Based Methods: Regression Trees.** 2017. 49 Slides.

TRANS, R. et al. **Best Practice Update 2007 / II Updated Handbook from Year 2003 Part II Intelligent Transport Systems (ITS)**.

VALA, K. Tree-Based Methods: Regression Trees - **Towards Data Science,** 2018. Disponível em: <https://towardsdatascience.com/tree-based-methods-regression-trees-4ee5d8db9fe9>. Acesso em: 20 set. 2019.