UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Otimização da previsão de entrega de itens por meio de algoritmos de Machine Learning

Lucimara Lye da Silva

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Lucimara Lye da Silva

Otimização da previsão de entrega de itens por meio de algoritmos de Machine Learning

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Ricardo Rodrigues Ciferri

Versão original

São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

S586o

Silva, Lucimara Lye da
Otimização da previsão de entrega de itens por meio de algoritmos de Machine Learning / Lucimara Lye da Silva; orientador Ricardo Rodrigues Ciferri. -- São Carlos, 2024.
54 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Logística. 2. Machine Learning. 3. Análise de dados. I. Ciferri, Ricardo Rodrigues, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Lucimara Lye da Silva

Otimização da Previsão de entrega de itens por meio de algoritmos de Machine Learning

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Original version

São Carlos

2024

Dedico este trabalho à minha família,
que foi o alicerce na formação do meu caráter e personalidade,
e ao meu marido, cuja paciência e compreensão durante os meses
de dedicação a esta jornada foram inestimáveis.

"Nós só podemos ver um pouco do futuro,
mas o suficiente para perceber que há muito a fazer."
Alan Turing

RESUMO

SILVA, L.L. **Otimização da previsão de entrega de itens por meio de algoritmos de Machine Learning**. 2024. 54 p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

Este trabalho foca na otimização da previsão da data de entrega de itens em uma empresa localizada no interior do Brasil, onde desafios logísticos são comuns. Utilizando algoritmos de *Machine Learning*, busca-se aprimorar a eficiência operacional da empresa, proporcionando datas de entrega mais precisas e superando limitações dos métodos tradicionais. A metodologia inclui análise detalhada da base de dados, implementação dos algoritmos e avaliação do desempenho dos modelos. Os desafios enfrentados abrangem a heterogeneidade dos dados e variações regionais, sendo esperado que os algoritmos de *Machine Learning* possam superá-los. O estudo visa contribuir significativamente para a área logística, oferecendo soluções inovadoras que resultem em maior satisfação do cliente, redução de custos operacionais e aumento dos ganhos financeiros da empresa.

Os arquivos utilizados neste trabalho estão disponíveis no GitHub e podem ser acessados através do seguinte link: https://github.com/lucimaralye/TCC

Palavras-chave: Logística. *Machine Learning*. Análise de dados.

ABSTRACT

SILVA, L.L. **Otimização da previsão de entrega de itens por meio de algoritmos de Machine Learning**. 2024. 54 p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

This study focuses on optimizing the delivery date forecast for items in a company located in the interior of Brazil, where logistical challenges are common. Using Machine Learning algorithms, the aim is to improve the company's operational efficiency, providing more accurate delivery dates and overcoming the limitations of traditional methods. The methodology includes detailed analysis of the database, implementation of the algorithms and evaluation of the models' performance. The challenges faced include data heterogeneity and regional variations, which Machine Learning algorithms are expected to overcome. The study aims to make a significant contribution to the logistics area, offering innovative solutions that result in greater customer satisfaction, reduced operating costs and increased financial gains for the company.

The files used in this project are available on GitHub and can be accessed through the following link: https://github.com/lucimaralye/TCC

Keywords: Logistics. Machine Learning. Data analysis

LISTA DE FIGURAS

Figura 1 – Machine Learning 29
Figura 2 – Aprendizado Supervisionado 31
Figura 3 – Fórmula do erro médio absoluto (MAE) 35
Figura 4 – Fórmula do erro quadrático médio (MSE) 35
Figura 5 – Fórmula do coeficiente de determinação (R^2) 36
Figura 6 – Visão geral do processo de regressão realizado no software Orange . . . 43

LISTA DE TABELAS

Tabela 1 – Estudos selecionados e suas características 42
Tabela 2 – Cenários de execução dos algoritmos 47
Tabela 3 – Indicadores de execução do cenário 1 48
Tabela 4 – Indicadores de execução do cenário 2 48
Tabela 5 – Indicadores de execução do cenário 3 49
Tabela 6 – Indicadores de execução do cenário 4 49

LISTA DE ABREVIATURAS E SIGLAS

ANN	Redes Neurais Artificiais
ARIMA	Autoregressive Integrated Moving Average
BI	Business Intelligence
CNN	Redes Neurais Convulacionais
MAE	Erro médio absoluto
ML	Machine Learning
MSE	Erro Quadrático Médio
OTD	On-Time Delivery
RMSE	Raiz do Erro Quadrático Médio
RNA	Redes Neurais Artificais
RNN	Redes Neurais Recorrentes
R2	Coeficiente de Determinação
SVM	Máquinas de Vetores de Suporte
USP	Universidade de São Paulo
USPSC	Campus USP de São Carlos
WGLS	Wärtsilä Global Logistics

SUMÁRIO

1 INTRODUÇÃO 23
1.1 Contextualização 23
1.2 Metodologia, desafios e expectativas 24

2 FUNDAMENTAÇÃO TEÓRICA 27
2.1 A Logística e sua importância empresarial 27
2.2 A Importância da previsão de entrega na logística empresarial . . . 28
2.3 *Machine Learning* 29
2.4 *Machine Learning* aplicado em previsões de data de entrega 30
2.5 Base de dados e pré-processamento 33
2.5.1 Limpeza dos dados 33
2.5.2 Seleção de características 33
2.5.3 Divisão da base 33
2.5.4 Validação cruzada 34
2.6 Métricas para avaliação de desempenho 34
2.7 *Overfitting* e *Underfitting* 36

3 TRABALHOS RELACIONADOS 39
3.1 String de busca 39
3.2 Trabalhos selecionados 39
3.2.1 (Sousa,2022) 39
3.2.2 (Rokoss et al., 2024) 40
3.2.3 (Pereira; Oliveira, 2023) 40
3.2.4 (Cunha, 2023) 41
3.2.5 Características dos estudos selecionados 42

4 METODOLOGIA 43
4.1 Considerações iniciais 43
4.2 Metodologia 43
4.3 Coleta de dados 44
4.4 Pré-processamento 44
4.5 Seleção de colunas 44
4.6 Divisão de dados 45
4.7 Ambiente computacional 46
4.8 Regressão 46
4.9 Métricas de avaliação 46

5 DISCUSSÃO E ANÁLISE DE RESULTADOS 47
5.1 Cenário 1 47
5.2 Cenário 2 48
5.3 Cenário 3 48
5.4 Cenário 4 49
5.5 Análise dos resultados 49

6 CONCLUSÃO E TRABALHOS FUTUROS 51

REFERÊNCIAS 53

1 INTRODUÇÃO

Este capítulo apresenta a contextualização sobre a importância da Logística nas organizações, e como a otimização da previsão da data de entrega de itens em uma empresa que está situada no interior do Brasil, poderá impactar positivamente na eficiência operacional da empresa e na satisfação do cliente. Também será apresentada uma alternativa para solucionar os problemas logísticos enfrentados por esta empresa, através da aplicação de algoritmos de *Machine Learning*. A solução proposta necessita de uma análise da base de dados, implementação dos algoritmos de *Machine Learning* e a avaliação do desempenho desses modelos.

1.1 Contextualização

No contexto empresarial contemporâneo, a eficiência logística desempenha um papel crucial para o sucesso de uma organização (Christopher, 2023), especialmente em um país de dimensões continentais como o Brasil. Em particular, as empresas localizadas no interior do país muitas vezes enfrentam desafios logísticos severos (Assis; Marchetti; Dalto, 2017), decorrentes de procedimentos internos que podem ser otimizados para melhor atender às demandas do mercado. Nessa situação, uma previsão mais precisa da data de entrega de itens para os clientes torna-se uma peça-chave para a satisfação do cliente e o aprimoramento dos ganhos da empresa.

Este trabalho tem como objetivo principal a aplicação de algoritmos de *Machine Learning* para a previsão da data de entrega de itens de diferentes clientes em uma empresa situada no interior do Brasil. Ao abordar os problemas logísticos relacionados aos procedimentos internos, a pesquisa busca proporcionar uma contribuição significativa para a eficiência operacional da empresa, resultando em ganhos substanciais e, consequentemente, na satisfação aprimorada dos clientes.

A importância estratégica dessa pesquisa reside na capacidade de fornecer datas de entrega mais assertivas, superando as limitações dos métodos tradicionais. A aplicação de algoritmos de *Machine Learning* permitem uma análise mais detalhada das variáveis que podem influenciar no processo logístico, contribuindo para a otimização das operações e reduzindo possíveis atrasos. A melhoria na precisão da previsão de entrega não apenas fortalece a confiança do cliente, mas também estabelece a empresa como uma referência na eficiência logística em seu segmento (Christopher, 2023).

1.2 Metodologia, desafios e expectativas

Assim, ao longo deste trabalho, serão explorados conceitos de *Machine Learning*, especialmente aplicados a previsão logística. Além disso, serão investigadas as particularidades dos problemas logísticos internos enfrentados pela empresa em questão, propondo soluções inovadoras baseadas em algoritmos de aprendizado de máquina. A meta final é disponibilizar prazos de entrega mais confiáveis, aumentando a satisfação do cliente, sua fidelização e, consequentemente, o crescimento financeiro para a organização.

A base de dados utilizada foi compilada entre os anos de 2018 e 2023, e engloba um vasto conjunto de informações, abrangendo a diversidade geográfica do Brasil e a complexidade da grande oferta de produtos da empresa. Com a presença de 825 itens distintos e a realização de entregas em 355 cidades e 26 estados, ela se configura como uma base extensamente diversificada, o que agrega desafios as abordagens convencionais de previsão. Por estas características, se justifica plenamente a aplicação de técnicas avançadas de *Machine Learning* para a análise desses dados complexos.

A metodologia adotada compreende a análise da base de dados, sua preparação para aplicação dos algoritmos de *Machine Learning* e a implementação desses algoritmos. Será dada atenção à divisão da base de dados em conjuntos de treinamento e teste, garantindo assim ser possível avaliar o desempenho na generalização dos modelos.

Os desafios a serem enfrentados incluem a heterogeneidade dos dados, a presença de sazonalidades e as possíveis variações climáticas em diferentes regiões. É esperado que a aplicação de algoritmos de *Machine Learning* possam superar esses desafios, proporcionando assim as previsões mais precisas e, consequentemente, melhorando a eficiência logística da empresa.

Este estudo busca contribuir para a área logística, oferecendo insights práticos e soluções inovadoras para aprimorar a previsão da data de entrega dos itens. O impacto esperado é o aumento da satisfação do cliente (o que poderia garantir sua fidelização), além da redução de custos operacionais e, claro, o incremento nos ganhos financeiros da empresa.

Com base nesta proposta, foi elaborada a seguinte questão de pesquisa:

Q1 “Como a aplicação de algoritmos de *Machine Learning* pode otimizar a previsão da data de entrega de itens, considerando uma base de dados tão abrangente e diversificada? ”

Para responder a essa questão, foram estabelecidos os seguintes objetivos:

- Realizar a revisão da literatura que abranja os algoritmos de *Machine Learning* que podem ser aplicados na previsão de dados.

- Implementar algoritmos de previsão em uma base de dados extraída do sistema empresarial, com a subsequente divisão dessa base em conjuntos de treinamento e testes.

- Analisar e interpretar os resultados obtidos, destacando a relevância prática e as implicações dos algoritmos de *Machine Learning*.

O próximo capítulo apresenta a fundamentação teórica deste estudo, que aborda conceitos essenciais sobre *Machine Learning* e sua aplicação na previsão temporal, bem como explora o aprendizado supervisionado nesse contexto. Além disso, serão discutidas métricas de desempenho relevantes e estratégias para o tratamento adequado da base de dados utilizada nesta pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo fornece a fundamentação teórica para contextualizar este trabalho, destacando a importância da logística empresarial e da previsão de entrega. Será explorado como o *Machine Learning* pode ser aplicado nesse contexto, discutindo métricas de avaliação, limpeza e pré-processamento da base de dados, e estratégias para enfrentar desafios na modelagem de previsão de entrega.

2.1 A Logística e sua importância empresarial

A logística é um componente de grande importância nas operações empresariais, desempenhando um papel fundamental na gestão do fluxo de bens, serviços e informações desde o ponto de origem até o ponto de consumo (Christopher, 2023). A princípio, a logística era vista somente como a responsável pelo transporte e o armazenamento dos produtos. Porém, a evolução e expansão comercial passou a impor maior complexidade ao processo. Sendo assim, a logística se tornou responsável pelo planejamento, implementação e controle de todas as etapas da cadeia de suprimentos. Isto quer dizer que ela é responsável pela aquisição, transporte, armazenamento, distribuição e a gestão dos estoques. E ela precisa garantir que os produtos certos estejam disponíveis no local certo, no momento certo e nas condições certas, enquanto minimiza os custos e otimiza a eficiência operacional.

Atualmente, podemos considerar que ela é essencial para o sucesso e a competitividade das empresas em um ambiente tão globalizado e dinâmico (Bowersox; Closs; Cooper, 2019). Dentre os motivos da importância atual da logística, podemos destacar:

- **Atendimento ao cliente:** As empresas podem atender às demandas dos clientes de forma rápida e confiável com um planejamento logístico eficiente. As entregas seguras e pontuais aumentam a satisfação do consumidor e os fidelizam à marca.

- **Redução de custos:** Quando a logística é bem planejada e executada, a empresa pode se beneficiar de uma redução nos custos operacionais. Otimizar rotas de transporte, gerenciar os estoques de forma inteligente e minizar o tempo de espera são exemplos de maneiras pelas quais as empresas podem economizar dinheiro por meio de práticas logísticas eficazes.

- **Eficiência operacional**: Uma logística eficiente consegue ter grande impacto positivo nas operações empresariais. Isso inclui a minimização de desperdícios, a redução de tempos de ciclo e a maximização da utilização de recursos e espaços, resultando em processos mais suaves e ágeis, que podem facilmente ser revertidos em ganhos financeiros.

• **Vantagem competitiva**: Empresas que investem em logística têm uma vantagem competitiva significativa. Elas são capazes de responder mais rapidamente às mudanças no mercado, as dificuldades de transporte específicas das diversas regiões, adaptar-se às demandas dos clientes e superar a concorrência por meio de serviços de entrega superiores e custos mais baixos.

2.2 A Importância da previsão de entrega na logística empresarial

A previsão de entrega envolve estimar o tempo necessário para que os itens sejam entregues aos clientes, desde o ponto de origem até o destino final, e é através desta previsão que as empresas devem montar o planejamento de suas operações. Ao estimar o tempo necessário para a entrega de itens, as empresas podem coordenar melhor os recursos necessários, como o transporte, a mão de obra e o espaço de armazenamento, reduzindo assim os custos operacionais e melhorando a eficiência geral.

Além disso, o cumprimento da estimativa de entrega é essencial na satisfação do cliente. Os consumidores modernos esperam receber seus produtos dentro de um prazo razoável após a compra (Chopra, 2019). Uma estimativa precisa do tempo de entrega permite que as empresas cumpram essas expectativas.

No entanto, a previsão de entrega apresenta uma série de desafios. Diversos fatores podem influenciar o tempo de entrega, incluindo distância, rota de transporte, condições climáticas, tráfego e capacidade de processamento nos centros de distribuição. A complexidade desses fatores torna a previsão de entrega uma tarefa desafiadora e muitas vezes imprecisa (Coyle *et al.*, 2016). Podemos inclusive, citar fatores muito específicos, como o furto de cargas frequentes em determinadas regiões, o que poderá impactar no planejamento, na rota e na entrega dos produtos (Cova, 2012).

Como solução para estes desafios, as empresas procuram uma variedade de tecnologias e ferramentas: algoritmos de previsão, análise de dados históricos, sistemas de rastreamento por GPS e previsão baseada em tempo real são apenas alguns exemplos de ferramentas usadas na previsão de entrega. Essas tecnologias atuam como auxílio para as empresas melhorarem a precisão de suas previsões e possibilitam uma resposta mais rápida às mudanças nas condições que possam afetar o tempo de entrega.

No entanto, é importante reconhecer que a previsão de entrega não é uma ciência exata. Incertezas nos dados, imprevisibilidade de eventos externos e complexidade nas redes de distribuição podem afetar a precisão das estimativas de entrega. Sendo assim, as empresas devem estar preparadas para lidar com variações e ajustar suas operações conforme necessário para atender às demandas dos clientes.

Uma vez que a entrega eficiente é essencial para a satisfação do cliente, podemos destacar que investir em um processo de logística poderá agregar benefícios importantes para o sucesso e crescimento da companhia.

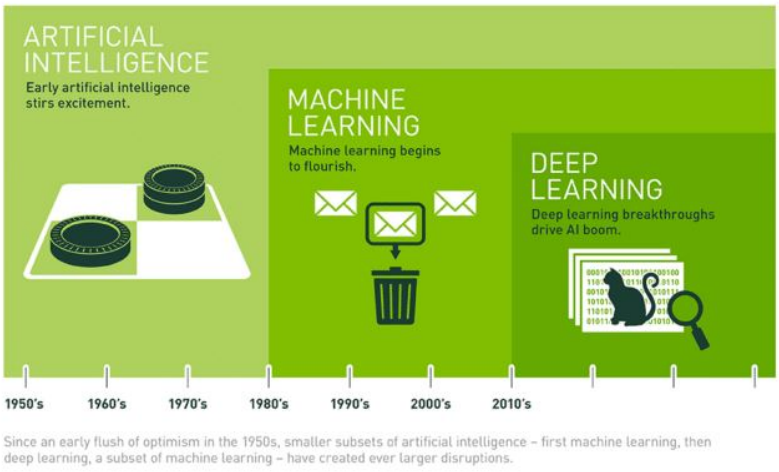
2.3 *Machine Learning*

Machine Learning (ML), ou Aprendizado de Máquina, é um subcampo da inteligência artificial (IA) focado no desenvolvimento de algoritmos e técnicas que permitem aos computadores aprender, fazer previsões ou tomar decisões a partir de bases de dados.

Diferentemente da programação convencional, onde algoritmos específicos são desenvolvidos para cada tarefa, o aprendizado de máquina capacita os computadores a reconhecer padrões subjacentes nos dados com base em exemplos e experiências anteriores. Utilizando métodos estatísticos, os algoritmos são treinados para fazer classificações ou previsões sem a necessidade de intervenção humana. Esta habilidade torna o ML uma ferramenta extremamente útil para lidar com quantidades grandes de dados e resolver problemas complexos (Murphy, 2012).

Na Figura 1, é possível observar a relação evolutiva entre IA, ML e Aprendizado Profundo (*Deep Learning*). Inicialmente, a IA surge como o conceito mais abrangente e primordial, representando a maior área. Posteriormente, dentro do campo da IA, desenvolve-se o ML, que foca em algoritmos e técnicas para permitir que sistemas aprendam a partir de dados. Por fim, dentro do Aprendizado de Máquina, emerge o Aprendizado Profundo, uma subárea mais recente e específica, que utiliza redes neurais profundas para modelar dados complexos e obter *insights* mais sofisticados.

Figura 1 – Machine Learning



Fonte: Copeland (2021)

Os algoritmos de *Machine Learning* podem ser categorizados em três tipos principais:

- **Aprendizado supervisionado**: Neste tipo, o modelo é treinado em um conjunto de dados rotulado, ou seja, onde a resposta desejada (rótulo) é conhecida. Exemplos incluem regressão linear, regressão logística, e redes neurais.

- **Aprendizado não supervisionado**: Aqui, o modelo é treinado em dados que não possuem rótulos. O objetivo é identificar padrões ou estruturas ocultas nos dados. Exemplos incluem *clustering* (agrupamento) e análise de componentes principais (PCA).

- **Aprendizado por reforço**: Este tipo envolve um agente que aprende a tomar decisões através de interações com o ambiente, recebendo recompensas ou punições com base em suas ações.

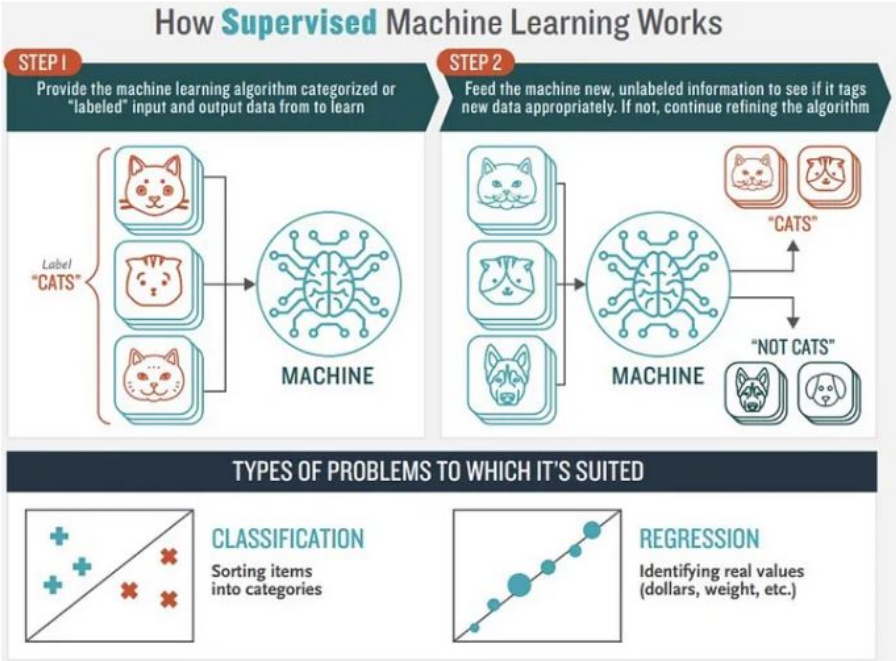
2.4 *Machine Learning* aplicado em previsões de data de entrega

Alpaydin (2010) discute a importância da previsão em *Machine Learning*, afirmando: "Uma vez que temos uma regra que se ajusta aos dados passados, se o futuro for semelhante ao passado, então podemos fazer previsões corretas para instâncias novas"(p. 5).

Em aprendizado supervisionado, um conjunto de dados de treinamento é composto por exemplos de entrada que podemos chamar de XX, associados a saídas desejadas que seriam as YY. Este paradigma pode ser utilizado em uma variedade de aplicações, incluindo previsão de séries temporais. Durante a fase de treinamento, o algoritmo trabalha ajustando seus parâmetros de forma iterativa, buscando minimizar a diferença entre as previsões do modelo e os rótulos verdadeiros. Após finalizar o treinamento, o modelo deverá ser avaliado em um conjunto de dados de teste, que foi previamente separado para verificar seu desempenho em dados que não foram vistos anteriormente, assim, é possível validar sua capacidade de generalização e sua precisão na previsão de valores futuros da série temporal.

A Figura 2 demonstra um diagrama ilustrando o processo de aprendizado supervisionado, onde os dados são inseridos já rotulados, e após o treinamento, novos dados são inseridos e o modelo consegue identificar a qual rótulo o novo dado pertence.

Figura 2 – Aprendizado Supervisionado



Fonte: Leonel (2018)

No contexto do aprendizado supervisionado, os dados de séries temporais são organizados de forma que cada entrada representa um ponto de dados no tempo, e a saída correspondente é o próximo valor na série temporal. Diferentes modelos podem ser utilizados nesta tarefa, incluindo a regressão linear, as redes neurais recorrentes (RNNs), as máquinas de Vetores de Suporte (SVM) ou os modelos ARIMA.

- **Regressão**: A regressão é uma técnica muito utilizada para prever variáveis contínuas (Bishop, 2006), como datas de entrega. Pode-se usar regressão linear, regressão polinomial ou outras formas mais avançadas de regressão, dependendo da natureza dos dados e da relação entre as variáveis.

- **Redes Neurais Artificiais (ANN)**: As redes neurais artificiais são eficientes em reconhecer padrões estatísticos nos dados (Bishop, 2006) e podem ser eficazes na previsão de datas de entrega. As redes neurais recorrentes (RNNs) e as redes neurais convolucionais (CNNs) são variantes que podem ser úteis para sequências temporais e dados de séries temporais, como as datas de entrega.

- **Máquinas de Vetores de Suporte (SVM)**: As SVMs são usadas para classificação e regressão (Bishop, 2006) e elas podem ser aplicadas para prever datas de entrega com base em características relevantes do pedido, como por exemplo, a localização do cliente, o tipo de transporte escolhido, o histórico de entrega, etc.

- ***Random Forest***: O *Random Forest* é um método de aprendizado de máquina amplamente utilizado para tarefas de classificação e regressão. Desenvolvido por Leo Breiman e Adele Cutler, o método funciona através da criação de várias árvores de decisão durante o processo de treinamento, e sua saída final é determinada pelo voto majoritário das classes (para classificação) ou pela média das previsões (para regressão) das árvores individuais. Esse modelo alia a simplicidade das árvores de decisão à robustez proporcionada pelo aprendizado em conjunto.

- **AdaBoost**: AdaBoost, que significa *Adaptive Boosting*, é um algoritmo de aprendizado em conjunto (*ensemble learning*) projetado para melhorar a performance de modelos de classificação. Introduzido por Yoav Freund e Robert Schapire em 1995, AdaBoost combina múltiplos classificadores fracos para formar um classificador forte, ajustando iterativamente os pesos dos exemplos de treinamento com base no desempenho do classificador anterior. Apesar do AdaBoost ser tradicionalmente um algoritmo de classificação, ele pode ser adaptado para tarefas de regressão, como a previsão da data de entrega, através de uma variante chamada AdaBoost.R. Esta adaptação permite que o algoritmo lide com problemas onde a variável de saída é contínua, tornando-o adequado para prever o tempo de entrega em dias.

- ***Gradient Boosting***: *Gradient Boosting* é uma técnica de aprendizado de máquina utilizada tanto para classificação quanto para regressão. Ela pertence à família dos métodos de ensemble e é conhecida por sua alta performance e flexibilidade. O método consiste na construção de um modelo forte a partir de uma série de modelos fracos, geralmente árvores de decisão. Cada modelo subsequente é treinado para corrigir os erros residuais dos modelos anteriores. O *Gradient Boosting* pode ser usado de maneira eficaz para prever a data de entrega de pedidos, ajustando-se à tarefa de regressão. O processo envolve treinar o modelo para prever o tempo de entrega (em dias), com base em várias características do pedido.

Na previsão de entrega, os algoritmos de *Machine Learning* podem ser utilizados devido à sua capacidade de analisar grandes volumes de dados e identificar padrões complexos, possibilitando previsões mais precisas e confiáveis. Isso permite às empresas anteciparem a demanda, otimizarem rotas de entrega e melhorarem a eficiência do processo logístico como um todo. Com o uso de algoritmos de *Machine Learning* na previsão de entrega, é possível uma abordagem eficaz e escalável para melhorar a precisão das previsões, otimizando as operações logísticas e melhorando a satisfação do cliente.

Com base nessas características, podemos destacar algumas vantagens da utilização do ML na Logística empresarial:

• **Capacidade de lidar com grandes volumes de dados**: possibilidade de trabalhar com algoritmos capazes de processar grandes quantidades de dados de forma eficiente, o que é essencial no campo da logística, onde os dados são frequentemente volumosos e complexos.

• **Detecção de padrões complexos**: capacidade de identificar padrões e relações não lineares nos dados, possibilitando previsões mais precisas em cenários complexos.

• **Adaptação a mudanças nas condições**: algoritmos flexíveis e adaptáveis, que podem ajustar suas previsões com base em mudanças nas condições, como variações na demanda, clima, dificuldades de transporte para determinadas regiões geográficas ou tráfego.

• **Melhoria contínua**: os algoritmos de *Machine Learning* são capazes de aprender com os erros e ajustar suas previsões com o tempo. Eles podem ser atualizados regularmente com novos dados de entrega, o que leva a previsões cada vez mais precisas e confiáveis com o passar do tempo.

2.5 Base de dados e pré-processamento

2.5.1 Limpeza dos dados

Esta etapa inclui identificar e corrigir dados ausentes, inconsistentes ou duplicados que podem distorcer a análise. Por exemplo, se existirem registros de entrega com datas impossíveis ou faltando informações essenciais, como endereços de entrega incompletos, técnicas para corrigir ou remover esses registros são aplicadas, garantindo que apenas os dados mais confiáveis e relevantes alimentem o modelo.

2.5.2 Seleção de características

Nesta etapa, são selecionadas quais características dos dados são relevantes para serem utilizadas nos modelos de previsão. Isso envolve avaliar as características com base em sua importância, correlação com a variável alvo (data de entrega) e potencial para melhorar o desempenho do modelo. Neste caso, técnicas estatísticas podem ser utilizadas, como análise de correlação ou testes de hipóteses, para identificar características altamente correlacionadas com a data de entrega.

2.5.3 Divisão da base

Na fase final da preparação dos dados, a base é dividida em duas partes, onde cada uma tem sua participação no aprimoramente e validação do modelo: conjuntos de treinamento e conjunto de testes. O conjunto de treinamento é utilizado para ensinar o modelo, é onde ele absorve os padrões e os relacionamentos entre as caracteristicas da

base com a data de entrega. O conjunto de testes é utilizado para validar o modelo, em dados que ele ainda não teve acesso. Esta divisão permite avaliar o modelo de forma mais objetiva, buscando alcançar uma generalização para diferentes cenários de entrega.

2.5.4 Validação cruzada

Validação cruzada é um processo onde o conjunto de dados é dividido em vários subconjuntos, chamados de folds. O modelo é treinado em um subconjunto e avaliado em outro, repetindo esse processo para todos os subconjuntos. Essa abordagem ajuda a reduzir o viés e a variabilidade na avaliação do desempenho do modelo, garantindo que a performance observada seja uma estimativa mais precisa da capacidade de generalização do modelo para dados não vistos.

No método K-Fold Cross-Validation, o conjunto de dados é dividido em K subconjuntos iguais. O modelo é treinado e avaliado K vezes, com cada subconjunto servindo como conjunto de teste uma vez e como parte do conjunto de treinamento K-1 vezes. As métricas de desempenho são então agregadas para fornecer uma avaliação geral do modelo.

2.6 Métricas para avaliação de desempenho

Avaliar de modo preciso os modelos é crucial para assegurar a eficácia e a confiabilidade das previsões que são efetuadas. Nesta seção, são destacadas as métricas de desempenho que costumam ser utilizadas na avaliação dos modelos de regressão.

• **Erro Médio Absoluto (MAE):** O MAE é uma métrica utilizada para avaliar a precisão de um modelo de regressão. Ele mede a média das diferenças absolutas entre as previsões feitas pelo modelo e os valores reais observados. Em outras palavras, o MAE calcula a média dos erros em que cada erro é o valor absoluto da diferença entre a previsão e o valor real (Cosio, 2021). Sendo assim, um MAE baixo significa que em média, as previsões do modelo foram muito perto do valor real, e um MAE alto significa que em média, os resultados do modelo estão longe do valor real, demonstrando assim, que o modelo tem um desempenho insatisfatório. Na figura 3, podemos ver a fórmula para calcular o MAE.

Figura 3 – Fórmula do erro médio absoluto (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Onde:

- n é o número de observações.
- y_i é o valor real da i-ésima observação.
- $\hat{y}_i$ é o valor previsto pelo modelo para a i-ésima observação.
- $|y_i - \hat{y}_i|$ é o valor absoluto do erro para a i-ésima observação.

Fonte: adaptada de Cosio (2021)

• **Erro Quadrático Médio (MSE)**: O MSE é uma métrica que calcula a média dos quadrados das diferenças entre as previsões do modelo e os valores reais (Cosio, 2021). Isso significa que grandes discrepâncias entre previsões e valores reais têm um impacto desproporcional no MSE. Como resultado, o MSE é útil para identificar modelos que têm erros grandes, mesmo que ocorra em apenas alguns casos.

Figura 4 – Fórmula do erro quadrático médio (MSE)

Fórmula:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Onde:

- y_i é o valor real da i-ésima observação.
- $\hat{y}_i$ é o valor previsto pela modelagem para a i-ésima observação.
- n é o número total de observações.

Fonte: adaptada de Cosio (2021)

• **Coeficiente de determinação (R^2):** O R^2 é uma métrica estatística que indica a proporção da variabilidade dos dados que é explicada pelo modelo (Kuhn; Johnson, 2013). No caso da previsão de entrega, o R^2 pode ser interpretado como a proporção da variabilidade nas datas de entrega que é explicada pelo modelo. Um R^2 próximo de 1

indica um ajuste perfeito do modelo aos dados, enquanto um R^2 próximo de 0 indica que o modelo não explica a variabilidade dos dados apresentados.

Figura 5 – Fórmula do coeficiente de determinação (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Onde:

- y_i são os valores reais.
- $\hat{y}_i$ são os valores previstos pelo modelo.
- $\bar{y}$ é a média dos valores reais.
- n é o número de observações.

Fonte: adaptada de Cosio (2021)

2.7 *Overfitting* e *Underfitting*

No desenvolvimento dos modelos de previsão de entrega, é importante dar atenção aos fenômenos de *overfitting* e *underfitting*, que podem impactar significativamente na capacidade do modelo de generalização para os novos dados.

- ***Overfitting*:** Quando o modelo se ajusta excessivamente aos dados de treinamento, capturando não apenas os padrões verdadeiros, mas também o ruído aleatório ou ocorrências irrelevantes nos dados, chamamos de *overfitting* (Bashir *et al.*, 2020). Por exemplo, se o modelo de previsão de data de entrega é treinado com um conjunto de dados históricos que inclui variáveis muito específicas e únicas para um determinado período de tempo, como condições climáticas extremas ou eventos sazonais incomuns, o modelo pode aprender a correlacionar essas variáveis específicas com as datas de entrega de forma excessiva. Isso pode resultar em um modelo que se comporta bem nos dados de treinamento, mas que falha em generalizar adequadamente para novos cenários ou períodos de tempo, podendo levar a previsões excessivamente otimistas ou muito pessimistas, que não refletem com precisão a realidade das operações de entrega.

- ***Underfitting*:** Ocorre quando o modelo é muito simples e não consegue capturar a complexidade dos dados utilizados (Bashir *et al.*, 2020). Isso poderá resultar em um modelo que não consegue incorporar os padrões subjacentes nos dados de treinamento e, consequentemente, tem um desempenho ruim tanto nos dados de treinamento quanto nos dados de teste. No caso da previsão de entrega, o *underfitting* poderá levar a previsões

imprecisas e inconsistentes, que não fornecem uma base confiável para o planejamento e a tomada de decisões.

No próximo capítulo, exploraremos os trabalhos relacionados que abordam temas similares aos discutidos nesta fundamentação teórica. A revisão dos trabalhos relacionados nos auxilia a obter *insights* adicionais e orientações para o desenvolvimento e a análise dos nossos resultados.

3 TRABALHOS RELACIONADOS

Neste capítulo, apresentaremos trabalhos relacionados ao mesmo estudo abordado neste TCC, focando na aplicação de algoritmos de *Machine Learning* para a previsão de tempos de entrega em diferentes contextos logísticos.

3.1 String de busca

Para a pesquisa de trabalhos relacionados a este estudo, foi utilizada uma string de busca com o objetivo de encontrar estudos que utilizassem algoritmos de *Machine Learning* para prever dados relacionados à entrega de itens ou a sitemas de logística. Diante disso, foram combinadas palavras utilizando os combinadores lógicos AND e OR, resultando em uma String de pesquisa em Inglês e Português, que foi utilizada na plataforma *Google Scholar*. A sintaxe de busca utilizada na plataforma foi: ((Machine Learning OR Aprendizado de Maquina) AND (previsão de entrega OR delivery OR logistics)).

3.2 Trabalhos selecionados

3.2.1 (Sousa,2022)

(Sousa, 2022) abordou em seu trabalho, desenvolvido na *Metropolia University of Applied Sciences*, a aplicação de *Machine Learning* (ML) para prever entregas pontuais na *Wärtsilä Global Logistics* (WGLS). Em seu estudo, ela destaca que a previsão de entrega pontual *(On-Time Delivery* - OTD) é fundamental para melhorar a satisfação do cliente e otimizar os processos operacionais. Ela utilizou uma combinação de métodos de pesquisa qualitativos e quantitativos, incluindo análise de documentos internos, observações e entrevistas para coletar dados qualitativos, além de dados quantitativos para desenvolver o algoritmo de ML. Este método misto enfatiza a profundidade e a aplicabilidade da pesquisa em um contexto empresarial real, proporcionando uma base robusta para o desenvolvimento tecnológico.

Sua pesquisa empregou vários algoritmos de ML como *Random Forest* e Redes Neurais demonstrando uma abordagem abrangente para identificar a ferramenta mais eficaz. O uso de múltiplas técnicas destacou a complexidade e os desafios na previsão de OTD, demonstrando a importância de adaptar e ajustar os modelos de acordo com as especificidades dos dados e do contexto operacional da WGLS.

Como resultado, a implementação do modelo de ML mostrou-se uma ferramenta promissora na redução do número de entregas atrasadas, um benefício direto para a

eficiência operacional e satisfação do cliente. Além disso, a integração do algoritmo em ferramentas de BI (*Business Intelligence*), como o Power BI, facilitou a análise contínua e a tomada de decisão baseada em dados.

3.2.2 (Rokoss et al., 2024)

(Rokoss *et al.*, 2024) exploram em seu trabalho, o uso de técnicas de aprendizado de máquina para prever prazos de entrega de forma mais precisa no início do processo de pedido. O estudo destaca a importância de integrar a data desejada de entrega do cliente como uma característica no modelo preditivo, uma abordagem que se mostrou eficaz para antecipar os prazos de entrega logo após o recebimento de um pedido de oferta. Esta metodologia permite às empresas ajustar suas capacidades de produção de forma proativa, uma estratégia que pode ser adaptada para melhorar os processos de planejamento da empresa.

Os autores aplicam a metodologia CRISP-DM adaptada especificamente para a previsão de tempos de entrega. E então, empregaram diversos algoritmos de aprendizado de máquina: XGBoost (*Extreme Gradient Boosting*), que é conhecido por sua eficiência e flexibilidade em tarefas de regressão e classificação. Redes Neurais Artificiais (ANNs) foram aplicadas para modelar relações complexas entre as variáveis. Máquinas de Vetores de Suporte (SVM) e Florestas Aleatórias (*Random Forests*) também foram usadas, ambas renomadas por sua robustez e capacidade de gerar modelos altamente precisos. Adicionalmente, Árvores de Decisão foram implementadas para proporcionar uma interpretação mais intuitiva dos modelos. Por fim, a Regressão Linear foi utilizada devido à sua simplicidade e eficácia em prever valores contínuos.

Por fim, a pesquisa de Rokoss et al. fornece *insights* valiosos sobre como os conjuntos de dados e as variáveis selecionadas impactam a precisão da previsão. Os resultados obtidos por eles demonstram que a inclusão de variáveis específicas do domínio, como as datas desejadas de entrega e detalhes do processo de manufatura, podem aumentar significativamente a precisão das previsões.

3.2.3 (Pereira; Oliveira, 2023)

O trabalho de (Pereira; Oliveira, 2023), desenvolvido na Faculdade de Computação e Informática da Universidade Presbiteriana Mackenzie, denominado "Previsão e Controle de Tempos das Entregas em Plataformas de Serviços com Inteligência Artificial", aborda a aplicação de modelos de aprendizado de máquina para otimizar os tempos de entrega em uma plataforma de distribuição de gás de cozinha.

Este estudo destacou a necessidade de adaptação das operações comerciais às novas tecnologias digitais e oferece um caso prático de como a inteligência artificial pode ser

empregada para melhorar a precisão das previsões e a gestão da cadeia de suprimentos. A abordagem metodológica adotada por Pereira e Oliveira, envolveu a análise comparativa de diversos modelos de aprendizado de máquina, onde podemos citar: D*ecision Tree, Random Forest, Extremely Randomized Trees, Support Vector Machines* (SVM), AdaBoost, *Gradient Boosting, Histogram-based Gradient Boosting, Bagging, e Recursive Feature Elimination* (RFE).

Esses modelos foram testados e avaliados por sua eficácia em prever os tempos de entrega, com o algoritmo *Extremely Randomized Trees* emergindo como o mais preciso, superando os métodos tradicionalmente utilizados pela empresa envolvida no estudo.

3.2.4 (Cunha, 2023)

A precisão na previsão de tempos de entrega tornou-se um fator crítico no sucesso das operações de *e-commerce*, especialmente em um cenário global afetado por flutuações significativas de demanda, como observado durante a pandemia de COVID-19. O trabalho de (Cunha, 2023) , intitulado "Uso de Aprendizado de Máquina para Especificação do Tempo de Entrega em Vendas Via *E-commerce*", busca resolver este problema, explorando a viabilidade de modelos de aprendizado de máquina avançados para prever tempos de entrega com mais precisão. Utilizando o "*Brazilian E-Commerce Public Dataset*"disponibilizado pela plataforma Kaggle, Costa Cunha implementa e avalia dois modelos distintos:

1. Redes Neurais Artificiais (RNAs): Este modelo foi escolhido devido à sua habilidade em captar a não-linearidade dos dados através de múltiplas camadas e neurônios. Costa Cunha adaptou uma rede neural profunda para prever o tempo de entrega com base em variáveis como dados demográficos do cliente, detalhes do produto e histórico de transações.

2. *Random Forest*: Reconhecido por sua robustez em face da variabilidade dos dados e menos propenso ao sobreajuste, o modelo de *Random Forest* foi utilizado para identificar as características mais influentes no prazo de entrega. Este modelo funciona por meio de uma combinação de várias árvores de decisão, cada uma treinada com uma amostra ligeiramente diferente dos dados, para produzir uma estimativa agregada que é geralmente mais precisa do que qualquer árvore individual poderia fornecer.

Os resultados indicam que o modelo de *Random Forest* superou as técnicas tradicionais usadas pelas empresas de e-commerce incluídas no estudo, demonstrando uma maior precisão nas previsões de tempo de entrega. Este achado é especialmente relevante, pois sugere que o *Random Forest* pode ser uma ferramenta valiosa para os operadores de *e-commerce* ajustarem seus processos logísticos, potencialmente levando a uma melhor satisfação do cliente e eficiência operacional.

Esta abordagem ilustra o potencial de técnicas de aprendizado de máquina em um campo aplicado, fornecendo insights significativos para futuras pesquisas e práticas industriais.

3.2.5 Características dos estudos selecionados

A tabela 1 demonstra um comparativo entre os trabalhos relacionados identificados na literatura e o presente trabalho:

Tabela 1 – Estudos selecionados e suas características

ESTUDO	Objetivo	Abrangência	Algoritimos Utilizados	Resultados Obtidos
Sousa (2022)	Prever a probabilidade de uma entrega ser pontual ou atrasada.	Mundial	Random Forest, Neural Network, Bagging Classifier, XGBoos e Support Vector Machine	O modelo com o melhor desempenho foi o Random Forest com 72,6% de precisão. Neural Network e Bagging Classifier foram os segundos melhores, ambos com uma precisão de 71,6%.
Rokoss et al. (2024)	Prever a data de Entrega assim que um pedido é recebido, considerando a data de entrega desejada pelo cliente como uma característica.	Mundial	Random Forest, Neural Network, Decision Tree, XGBoost e Support Vector Machine	O modelo com melhor desempenho foi o Random Forest, alcançando um R2 de 0,56 e um MAE de 5,82. Em seguida, foram identificados os modelos de Neural Network e Support Vector Machine com melhores resultados.
Oliveira e Pereira (2023)	Previsão do tempo de entrega em uma plantaforma de venda e entrega de gás.	Brasil	Random Forest, Bagging Classifier, Decision Tree, XGBoost e Support Vector Machine e Extremely Randomize Trees	O algoritmo Extremely Randomized Trees foi identificado como o que fez previsões com maior acurácia, com um MAPE de 0,164471. Em seguida, veio o Random Forest com um MAPE de 0,161812.
Cunha (2023)	Previsão do tempo de entrega de compras online	Brasil	Random Forest e Neural Network	Comparando os resultados dos modelos com as previsões da ferramenta atual da empresa, constatou-se que o Random Forest reduziu o erro em 78,48%, enquanto a Rede Neural reduziu o erro em 78,71%.
Proposta de Pesquisa	Previsão da data de Entrega de Itens no Brasil	Brasil	Support Vector Machine, Random Forest, Neural Network, AdaBoost, Gradient Boosting, Regressão Linear	Documentado nos próximos capítulos

Fonte: autoria própria

O próximo capítulo descreverá a metodologia adotada para este estudo. Iremos detalhar os procedimentos e técnicas utilizados para a coleta e análise dos dados, bem como a implementação dos modelos preditivos e as estratégias para a avaliação de seu desempenho.

4 METODOLOGIA

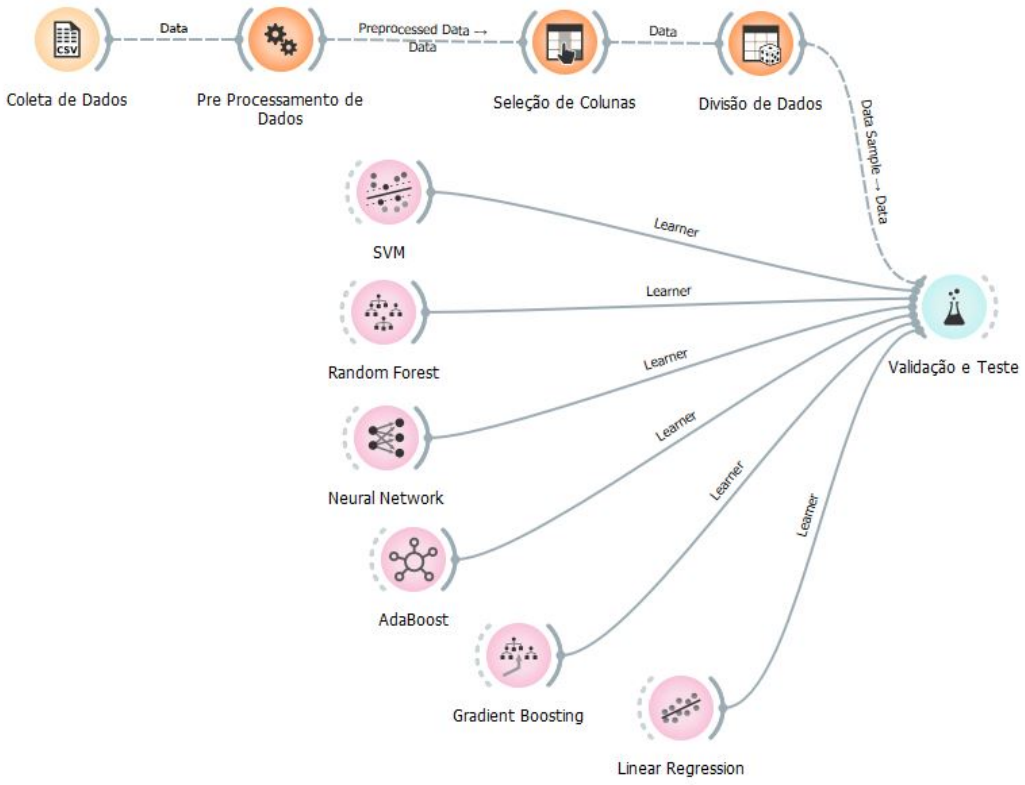
4.1 Considerações iniciais

Este capítulo tem como objetivo detalhar a proposta desta pesquisa. Durante este capítulo, detalharemos a base de dados que está sendo utilizada na pesquisa, sua forma de extração e o tratamento dos dados. Posteriormente, será detalhado o ambiente computacional utilizado, os modelos selecionados na pesquisa e quais as métricas de desempenho foram selecionadas para analisar o resultado do trabalho.

4.2 Metodologia

A figura 6 mostra um fluxo detalhado criado no software Orange, com as etapas envolvidas na Mineração e Preparação dos Dados, para a Aplicação dos Algoritmos de Regressão:

Figura 6 – Visão geral do processo de regressão realizado no software Orange



Fonte: autoria própria, gerada através do software Orange 3.36.2

4.3 Coleta de dados

Os dados utilizados neste projeto foram extraídos de um banco de dados Progress OpenEdge, oriundos das operações de um sistema ERP da empresa, e transformados em um arquivo CSV. A extração de dados focou especificamente nas vendas realizadas, abrangendo as seguintes informações:

- **Código do item:** Identificador único de cada produto vendido.

- **Cidade e Estado de entrega:** Localização geográfica onde cada item foi entregue.

- **Data de implantação do pedido:** Data em que o pedido foi registrado no sistema.

- **Data de entrega do pedido:** Data em que o item foi efetivamente entregue ao cliente.

- **Tempo de entrega:** Número de dias decorridos entre a implantação do pedido e a entrega do item.

Os dados estão disponibilizados no projeto do GitHub: https://github.com/lucimaralye/TCC.

4.4 Pré-processamento

O pré-processamento da base de dados foi realizado para remover inconsistências e erros que poderiam comprometer a análise. Este processo incluiu a identificação e a exclusão de registros com dados divergentes, como:

- **Datas de entrega anteriores às datas de implantação do pedido:** Esses registros foram removidos por estarem evidentemente incorretos.

- **Pedidos sem data de entrega:** Pedidos sem a data de entrega registrada também foram excluídos, pois esses dados são essenciais para calcular o tempo de entrega e realizar análises precisas.

Esse pré-processamento foi fundamental para garantir a integridade e a confiabilidade dos dados utilizados na análise subsequente.

4.5 Seleção de colunas

Nesta etapa, as colunas da base de dados foram analisadas e classificadas quanto à sua importância para o treinamento dos modelos de previsão. Esse processo é crucial para

identificar quais variáveis podem influenciar significativamente os resultados. Por exemplo, uma análise detalhada das colunas pode revelar que o mês de implantação do pedido tem uma influência considerável na previsão do tempo de entrega.

Para realizar essa análise, foram seguidas as seguintes etapas de forma manual:

- **Avaliação individual das colunas:** Cada coluna foi avaliada individualmente para determinar sua relevância e qualidade dos dados. Isso incluiu a verificação de consistência e a presença de valores ausentes ou errôneos.

- **Identificação de padrões sazonais:** Foi examinado se o mês de implantação do pedido influencia o tempo de entrega. Este tipo de análise pode revelar padrões sazonais ou tendências que são vitais para melhorar a precisão dos modelos de previsão.

Essa análise criteriosa das colunas não só melhora a qualidade dos dados para o treinamento dos modelos, mas também ajuda a identificar quais variáveis devem ser incluídas para otimizar as previsões.

4.6 Divisão de dados

Os dados foram segregados em subconjuntos para treinamento e teste, a fim de garantir a validade e a robustez dos modelos de previsão desenvolvidos. Este processo envolveu a divisão da base de dados em dois grupos distintos, cada um com diferentes percentuais de amostras:

- **Conjunto de treinamento:** Uma parte maior dos dados foi alocada para o treinamento dos modelos. Este subconjunto é utilizado para ajustar os parâmetros dos modelos e aprender os padrões presentes nos dados.

- **Conjunto de teste:** Uma porção menor dos dados foi reservada para testar a performance dos modelos. Este subconjunto é utilizado para avaliar a precisão e a capacidade de generalização dos modelos em novos dados não vistos durante o treinamento.

Além da divisão básica dos dados em conjuntos de treinamento e teste, foram empregados diferentes cenários de validação cruzada (*cross-validation*) para uma avaliação mais criteriosa dos modelos. A validação cruzada é uma técnica que envolve a divisão dos dados em múltiplos subconjuntos, permitindo que cada subconjunto seja utilizado como conjunto de teste em diferentes iterações enquanto os outros são usados para treinamento. Isso ajuda a obter uma avaliação mais robusta do desempenho do modelo e a reduzir o viés associado a uma única divisão dos dados.

A segregação dos dados e a aplicação desses métodos de validação cruzada garantiram que os subconjuntos fossem representativos do comportamento geral dos dados e possibilitaram uma avaliação mais completa da eficácia dos modelos de previsão. Essa abordagem ajuda a ajustar os parâmetros dos modelos conforme necessário, melhorando assim a precisão e a capacidade de generalização das previsões.

4.7 Ambiente computacional

Os modelos foram executados em um notebook Dell, com as seguintes configurações:

- Processador 13th Gen Intel(R) Core(TM) i7-13650HX 2.60GHz
- 16,00 GB RAM
- Windows 11 Home Single Language 64 bits
- Visual Studio Code Versão 1.89.1

4.8 Regressão

Seis algoritmos de regressão foram treinados com os dados desta pesquisa: SVM, *Random Forest, Neural Network*, AdaBoost, *Gradient Boosting* e *Linear Regression*.

4.9 Métricas de avaliação

Três métricas de avaliação foram utilizadas para analisar a performance dos modelos: MAE, MSE e R^2.

No próximo capítulo, apresentaremos os resultados obtidos a partir da aplicação da metodologia descrita. Focaremos na análise dos desempenhos dos modelos de previsão de entrega, discutindo os achados principais e interpretando as implicações dos dados coletados.

5 DISCUSSÃO E ANÁLISE DE RESULTADOS

Neste capítulo, apresentamos os resultados obtidos a partir da execução dos algoritmos em quatro cenários diferentes. Cada cenário foi projetado para avaliar o desempenho dos modelos sob diferentes configurações de validação e divisão dos dados. Os cenários incluem variações no número de folds e nas proporções de dados utilizados para treinamento e teste.

Os arquivos utilizados neste trabalho estão disponíveis no GitHub e podem ser acessados através do seguinte link: https://github.com/lucimaralye/TCC

Neste estudo, utilizamos dois cenários onde o percentual de dados alocados para treinamento e teste variou (um cenário com 70% e um com 80%), e dois cenários de validação cruzada k-fold, com variação no número de folds (um cenário com 5 folds e outro com 10 folds), e que são apresentados na Tabela 2.

Tabela 2 – Cenários de execução dos algoritmos

	Número de *folds*	**Percentual utilizado**
Cenário 1	Não Aplicável	70%
Cenário 2	Não Aplicável	80%
Cenário 3	5	Não Aplicável
Cenário 4	10	Não Aplicável

Fonte: autoria própria

5.1 Cenário 1

No cenário 1, os dados foram divididos em um conjunto de treinamento com 70% das amostras e um conjunto de teste com 30%. As métricas médias de desempenho são apresentadas na tabela 3.

Um ponto de destaque neste cenário foi a agilidade na execução dos modelos, mesmo lidando com um arquivo que continha mais de 120 mil linhas.

Tabela 3 – Indicadores de execução do cenário 1

Indicador	SVM	Random Forest	Neural Network	AdaBoost	Gradient Boosting	Linear Regression
MSE	178,88	179,99	47,45	1606,57	431,6	32413,4
MAE	3,64	3,56	1,93	36,91	14,63	4,38
R2	0,62	0,62	0,90	-2,34	0,10	-66,45
Tempo Treinamento	6m 25.9s	8m 52.5s	7m 22.9s	4m 10.s	8m 9.9s	21,3s

Fonte: autoria própria

5.2 Cenário 2

Neste cenário, 80% dos dados foram usados para treinamento e 20% para testes. As métricas médias de desempenho obtidas são apresentadas na tabela 4.

Neste cenário, observamos um aumento no tempo de execução dos modelos comparado ao cenário 1. No entanto, houve um ganho de desempenho nas métricas para o SVM, a *Random Forest* e a *Neural Network*.

Tabela 4 – Indicadores de execução do cenário 2

Indicador	SVM	Random Forest	Neural Network	AdaBoost	Gradient Boosting	Linear Regression
MSE	164,09	170,08	40,09	1731,85	438,16	8043,1
MAE	3,34	3,31	1,89	38,52	14,69	2,30
R2	0,66	0,65	0,91	-2,56	0,09	-15,55
Tempo Treinamento	8m 8.0s	10m 45.5s	8m 24.6s	4m 44.2s	9m 22.5s	18,3s

Fonte: autoria própria

5.3 Cenário 3

Neste cenário, os dados foram divididos em 5 folds iguais. Cada modelo foi treinado e avaliado 5 vezes, com cada fold servindo como conjunto de teste uma vez e como parte do conjunto de treinamento nas demais iterações. Na tabela 5, podemos analisar as métricas médias de desempenho obtidas para os modelos.

Neste caso, observamos que o uso dessa técnica com mais iterações resultou em um aumento substancial no tempo de execução de todos os modelos. Além disso, notamos

uma piora no desempenho dos modelos SVM e *Random Forest*, enquanto que o *Gradient Boosting* apresentou um aumento significativo no desempenho.

Tabela 5 – Indicadores de execução do cenário 3

Indicador	SVM	Random Forest	Neural Network	AdaBoost	Gradient Boosting	Linear Regression
MSE	333,63	331,47	85,76	729,52	151,32	59025,06
MAE	7,98	12,45	2,14	23,18	4,55	4,32
R2	0,32	0,32	0,87	-0,48	0,69	-119,22
Tempo Treinamento	683m 42.0s	1277m 44.6s	112m 19.4s	3 min 12.5s	72 m 15.4s	24m 16.2s

Fonte: autoria própria

5.4 Cenário 4

Para este cenário, os dados foram divididos em 10 folds. Cada modelo foi avaliado em 10 iterações, com cada fold servindo como conjunto de teste uma vez. As métricas de desempenho para este cenário são apresentadas na tabela 6.

Com essas informações, podemos concluir que, quanto maior o número de iterações, mais tempo será necessário para executar os modelos. No entanto, em alguns casos, os resultados obtidos foram significativamente melhores do que os anteriores.

Tabela 6 – Indicadores de execução do cenário 4

Indicador	SVM	Random Forest	Neural Network	AdaBoost	Gradient Boosting	Linear Regression
MSE	325,74	298,4	43,23	728,11	143,63	198721,87
MAE	7,76	10,51	2,02	23,35	4,35	6,55
R2	0,34	0,39	0,93	0,50	0,71	-404,67
Tempo Treinamento	830m 32.4s	2948m 25.6s	280m 40.1s	8m 22,5s	88m 14.3s	48m 23.9s

Fonte: autoria própria

5.5 Análise dos resultados

Os resultados mostraram que o ***Neural Network*** se destacou como o melhor modelo em termos de precisão, apresentando consistentemente os menores valores de MAE e MSE, além do maior R2. Isso indica que o modelo foi capaz de capturar a complexidade

dos dados, lidando eficazmente com padrões não lineares e interações entre variáveis que outros modelos não captaram com a mesma eficiência.

Em comparação:

- ***Random Forest*:** apresentou bom desempenho, especialmente em cenários com menor variabilidade nos dados (quando não utilizamos a validação cruzada k-fold). No entanto, seu desempenho foi inferior ao *Neural Network*, especialmente em termos de R2, sugerindo que ele teve dificuldades em capturar as interações mais complexas presentes no conjunto de dados.

- ***Gradient Boosting*:** demonstrou resultados satisfatórios após a aplicação da validação cruzada k-fold, que ajudou a aumentar a precisão do modelo. No entanto, apesar do desempenho competitivo, o *Gradient Boosting* ainda ficou atrás do Neural Network.

- **SVM:** teve um desempenho modesto, mostrando-se eficaz em cenários onde não utilizamos a validação k-fold.

- **AdaBoost:** apresentou um dos desempenhos mais modestos entre os modelos testados. No entanto, observou-se uma melhora significativa em seu desempenho na última simulação, quando o número de iterações foi aumentado, sugerindo que o modelo pode se beneficiar de ajustes mais finos de hiperparâmetros para alcançar resultados mais competitivos.

- ***Linear Regression*:** apresentou resultados abaixo do esperado em todas as simulações realizadas, não conseguindo atingir um desempenho satisfatório, destacando sua inadequação para o tipo de dados logísticos analisados neste estudo.

A validação cruzada k-fold revelou-se uma técnica crucial para melhorar a performance dos modelos, especialmente para o *Neural Network* e o *Gradient Boosting*, que mostraram maior capacidade de generalização e redução de *overfitting*. No entanto, essa técnica também aumentou significativamente o tempo de execução, evidenciando um equilíbrio necessário entre precisão e eficiência computacional que deve ser considerado na prática.

No próximo capítulo, serão discutidas as conclusões finais deste trabalho, juntamente com sugestões para futuros estudos e pesquisas que podem ser desenvolvidas com base nos resultados deste projeto.

6 CONCLUSÃO E TRABALHOS FUTUROS

Ao longo das análises realizadas, o algoritmo de *Neural Network* destacou-se como o de melhor desempenho, apresentando os menores valores de MAE e MSE, além do maior valor de R^2, quando comparado aos demais modelos testados, em todos os cenários utilizados no experimento.

No entanto, é importante ressaltar que a utilização da validação cruzada k-fold contribuiu significativamente para melhorar a eficiência de outros modelos. Tanto o *Neural Network* quanto o *Gradient Boosting* apresentaram resultados satisfatórios com essa abordagem. Contudo, também foi notável o aumento no tempo de execução dos modelos ao empregar essa técnica, o que deve ser considerado ao optar por sua aplicação em cenários práticos.

Do ponto de vista prático, os resultados obtidos na pesquisa deste TCC, sugerem que a aplicação de técnicas avançadas de *Machine Learning*, pode significativamente melhorar a precisão na previsão de datas de entrega, contribuindo para o aumento da satisfação do cliente e a eficiência operacional da empresa. Tais melhorias são especialmente relevantes em contextos logísticos complexos, como o da empresa estudada, onde as variações regionais e sazonais, e os processos industriais internos, apresentam desafios adicionais.

Entretanto, é importante destacar algumas limitações do estudo. A base de dados utilizada, embora abrangente, pode conter variáveis que não foram totalmente exploradas, o que pode influenciar no resultado das previsões. Além disso, a dependência de dados históricos implica que mudanças abruptas ou imprevistas no contexto operacional da empresa podem não ser adequadamente capturadas pelos modelos.

Embora os resultados deste estudo sejam promissores, há várias oportunidades para aprofundar e expandir a pesquisa. A seguir, são apresentadas algumas sugestões para trabalhos futuros:

Incorporação de novas variáveis: A inclusão de variáveis adicionais, como dados meteorológicos, condições de tráfego em tempo real e variáveis econômicas, poderia enriquecer os modelos preditivos e fornecer insights mais detalhados sobre fatores que influenciam os prazos de entrega.

Automatização e escalabilidade dos modelos: Outra possibilidade seria a automatização do processo de modelagem e previsão, permitindo a aplicação dos algoritmos em

tempo real. Além disso, a pesquisa pode ser expandida para contextos com bases de dados maiores, visando testar a escalabilidade dos modelos.

Estudo de outras técnicas de pré-processamento: Investigar técnicas de pré-processamento mais avançadas, como a seleção automática de características e o uso de métodos de imputação de dados faltantes, poderia melhorar a qualidade dos dados e, consequentemente, a precisão das previsões.

Aplicação em diferentes contextos logísticos: Por fim, seria interessante aplicar os modelos desenvolvidos em outros contextos logísticos, como e-commerce, transporte de cargas pesadas ou distribuição urbana, para verificar a generalização dos resultados obtidos e adaptar as técnicas conforme necessário.

Para concluir, este trabalho abre caminho para o uso avançado de *Machine Learning* na logística, demonstrando seu potencial para transformar a gestão de entregas e reforçando a importância de continuar explorando novas abordagens para otimizar esse processo.

REFERÊNCIAS

ALPAYDIN, E. **Introduction to Machine Learning**. [*S.l.: s.n.*]: MIT Press, 2010. 712 p. 4th edition.

ASSIS, A. C. V.; MARCHETTI, D. d. S.; DALTO, E. J. Panoramas setoriais 2030: Desafios e oportunidades para o brasil. Rio de Janeiro, p. 225, 2017. Artigo de Logistica, páginas 173 a 190. Disponível em: https://www.bndes.gov.br/wps/wcm/connect/site/48dedb93-fb01-4b58-92de-4a5735669c86/BNDES_PANORAMAS+SETORIAIS+2030_completo.pdf?MOD=AJPERES&CVID=m3.O69v. Acesso em: 23 Setembro 2024.

BASHIR, D. *et al.* **An Information-Theoretic Perspective on Overfitting and Underfitting**. [*S.l.: s.n.*], 2020. (Lecture Notes in Computer Science). Disponível em: https://arxiv.org/pdf/2010.06076.pdf. Acesso em: 01 abril 2024.

BISHOP, C. M. **Pattern Recognition and Machine Learning**: Information science and statistics. [*S.l.: s.n.*]: Springer, 2006. 778 p.

BOWERSOX, D. J.; CLOSS, D. J.; COOPER, M. B. **Supply chain logistics management**. [*S.l.: s.n.*]: McGraw-Hill Education, 2019. 960 p. 5th edition.

CHOPRA, S. **Supply chain management: Strategy, planning, and operation, Global Edition**. [*S.l.: s.n.*]: Pearson, 2019. 540 p. 7th edition.

CHRISTOPHER, M. **Logistics supply chain management**: Praise for logistics supply chain management. [*S.l.: s.n.*]: FT Publishing International, 2023. 360 p. 6th edition.

COPELAND, M. **Qual é a Diferença entre Inteligência Artificial, Machine Learning e Deep Learning?** 2021. Disponível em: https://blog.nvidia.com.br/blog/qual-e-a-diferenca-entre-inteligencia-artificial-machine-learning-e-deep-learning/. Acesso em: 20 abril 2024.

COSIO, N. A. L. **Métricas en regresión**. 2021. Disponível em: https://medium.com/@nicolasarrioja/m%C3%A9tricas-en-regresi%C3%B3n-5e5d4259430b. Acesso em: 23 Setembro 2024.

COVA, C. **Logística Empresarial**: Material didático. [*S.l.: s.n.*], 2012. Volume 3. Disponível em: https://canal.cecierj.edu.br/012016/a5c5d29ef5fb9421f11f3d3d9c50a2f5.pdf. Acesso em: 10 jan. 2024.

COYLE, J. J. *et al.* **Supply chain management: a logistics perspective**. [*S.l.: s.n.*], 2016. Nelson Education.

CUNHA, V. B. C. Uso de aprendizado de máquina para especificação do tempo de entrega em vendas via e-commerce. 2023. Disponível em: http://www.monografias.ufop.br/handle/35400000/5808. Acesso em: 18 abril 2024.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. [*S.l.: s.n.*]: Springer, 2013. 600 p. 2013th edition.

LEONEL, J. **Supervised Learning**. 2018. Disponível em: https://medium.com/@jorgesleonel/supervised-learning-c16823b00c13. Acesso em: 20 abril 2024.

MURPHY, K. P. **Machine Learning: A Probabilistic Perspective**. [*S.l.: s.n.*], 2012. 1067 p.

PEREIRA, G. B.; OLIVEIRA, R. Previsão e controle de tempos das entregas em plataformas de serviços com inteligência artificial. 2023. Disponível em: https://adelpha-api.mackenzie.br/server/api/core/bitstreams/13013e58-6106-479f-8574-0fd133c64dba/content. Acesso em: 16 abril 2024.

ROKOSS, A. *et al.* Case study on delivery time determination using a machine learning approach in small batch production companies. **Journal of Intelligent Manufacturing**, 2024. Disponível em: https://link.springer.com/article/10.1007/s10845-023-02290-2. Acesso em: 15 abril 2024.

SOUSA, D. G. **Using Machine Learning to Predict On-Time Delivery**. 2022. 78 p. Dissertação (Degree Programme in Business Informatics) — Metropolia University of Applied Sciences, 2022. Disponível em: https://www.theseus.fi/handle/10024/784410. Acesso em: 15 abril 2024.