



Previsão de Vendas no Comércio a Retalho Utilizando Algoritmos de Machine Learning

NATASHA KATHARINA ZIEGLER
setembro de 2024



PREVISÃO DE VENDAS NO COMÉRCIO A RETALHO UTILIZANDO ALGORITMOS DE MACHINE LEARNING

Natasha Katharina Ziegler

2024

Instituto Superior de Engenharia do Porto

Departamento de Engenharia Mecânica

isen

P.PORTO

PREVISÃO DE VENDAS NO COMÉRCIO A RETALHO UTILIZANDO ALGORITMOS DE MACHINE LEARNING

Natasha Katharina Ziegler

Estudante n.º 1220292

Dissertação apresentada ao Instituto Superior de Engenharia do Porto para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia e Gestão da Cadeia de Abastecimento, realizada sob a orientação do Doutor Carlos Manuel Abreu Gomes Ferreira e coorientação do Doutor Manuel Joaquim Pereira Lopes.

2024

Instituto Superior de Engenharia do Porto

Departamento de Engenharia Mecânica

isen

P.PORTO

AGRADECIMENTOS

Gostaria de manifestar a minha profunda gratidão a todos os que contribuíram para a realização deste projeto de investigação. A jornada foi grandemente enriquecida pelo apoio e empenho de muitos, sem os quais a concretização deste trabalho não teria sido possível.

Em primeiro lugar, gostaria de expressar um agradecimento especial ao Instituto Superior de Engenharia do Porto (ISEP) pela oportunidade de desenvolver este projeto. Este trabalho permitiu-me aplicar na prática os conhecimentos adquiridos e explorar técnicas avançadas num contexto real de mercado. A minha sincera apreciação vai também para o meu orientador científico, Doutor Carlos Manuel Abreu Gomes Ferreira, cuja orientação perspicaz e constante disponibilidade para esclarecer dúvidas foram fundamentais para o sucesso deste trabalho. O seu apoio incondicional foi crucial para orientar o projeto a um desfecho positivo.

Gostaria igualmente de agradecer ao Diretor de Curso, Engenheiro Manuel Pereira Lopes, pela sua visão inovadora e preparação cuidadosa, que nos prepararam para enfrentar os desafios futuros com confiança e competência. Agradeço também a todos os professores do curso, cuja dedicação e ensino de excelência foram decisivos para o meu crescimento académico e profissional. O conhecimento e a inspiração que transmitiram foram essenciais para a realização deste projeto.

Não posso deixar de mencionar os amigos e colegas que, ao longo deste percurso académico, estiveram sempre disponíveis para oferecer apoio e encorajamento, tornando a jornada mais leve.

Aos meus pais e à minha irmã, expresso a minha mais sincera gratidão pelo apoio incondicional e pela fé constante em mim, que foram fundamentais em cada fase da minha vida.

A todos os mencionados e a todos aqueles que, de alguma forma, contribuíram para este projeto, o meu obrigada!

RESUMO

O sector do retalho, avaliado em biliões de euros, é dinâmico e em constante evolução, enfrentando desafios na previsão precisa da procura devido à rápida mudança nas preferências dos consumidores, às tendências emergentes e à variedade de atributos dos produtos. A necessidade de planear as vendas com antecedência e a escassez de dados concretos aumentam ainda mais a complexidade. Métodos de previsão como os estatísticos, apreciados pela sua clareza, podem ser insuficientes com dados complexos, enquanto as técnicas de aprendizagem automática conseguem identificar padrões mais complexos, apesar de exigirem grandes volumes de dados.

Esta dissertação utiliza a metodologia CRISP-DM para explorar métodos estatísticos de previsão e a adaptabilidade dos algoritmos de aprendizagem automática na previsão de vendas no setor de retalho. A combinação destas técnicas avançadas com o processo CRISP-DM proporciona uma análise mais robusta e flexível, aumentando a eficiência e a precisão das previsões de procura. Isto permite otimizar a gestão de stocks e o planeamento estratégico, o que pode melhorar significativamente o desempenho operacional e a rentabilidade da empresa.

Os dados analisados pertencem à empresa norte-americana Walmart e abrangem as vendas de 10 lojas distribuídas por três estados: Califórnia, Texas e Wisconsin, ao longo de um período de 5 anos. Para prever as vendas futuras, foram desenvolvidos diversos modelos, incluindo um modelo global, treinado com todos os dados, e outros específicos por loja e por estado. O modelo global captura tendências gerais, enquanto os modelos específicos por estado e por loja identificam padrões únicos de cada região e unidade, proporcionando previsões mais ajustadas a nível local. Adicionalmente, cada modelo de aprendizagem automática foi treinado em 4 cenários distintos: utilizando dados totais, do último ano, dos últimos 3 meses e do último mês, com o objetivo de observar o impacto das variações temporais nas previsões. Foram também considerados fatores como feriados e dias da semana nos modelos, que têm um impacto significativo nas vendas.

Os resultados obtidos mostram que, entre os métodos estatísticos, o modelo agregado por loja com SARIMA foi o mais eficaz para a previsão de vendas em contextos sazonais, alcançando um MAE de 224 e um MAPE de 5,39%. No que diz respeito aos métodos de aprendizagem automática, o *Support Vector Machines* (SVM) destacou-se ao apresentar valores elevados de R^2 , variando entre 77% e 80% para os modelos agregados por estado e loja, com erros relativamente baixos. A análise também revelou que o modelo agregado global apresentou um desempenho inferior, refletindo uma capacidade limitada para capturar as particularidades regionais e locais. Estes resultados sublinham a importância de utilizar dados atualizados e ajustar os modelos às recentes variações do mercado, enfatizando a necessidade de uma abordagem adaptativa na modelagem preditiva. A comparação entre diferentes metodologias e a inclusão de hiperparâmetros foram cruciais para a precisão das previsões, com os resultados variando conforme o modelo e o contexto de aplicação.

Palavras-Chave: CRISP-DM, Previsão, Métodos Estatísticos, Machine Learning

ABSTRACT

The retail sector, valued in billions of euros, is dynamic and constantly evolving, facing challenges in accurately forecasting demand due to rapid changes in consumer preferences, emerging trends, and the variety of product attributes. The need to plan sales in advance and the lack of concrete data further increase the complexity. Forecasting methods like statistical ones, valued for their clarity, may be insufficient with complex data, while machine learning techniques can identify more complex patterns, although they require large volumes of data.

This dissertation uses the CRISP-DM methodology to explore statistical forecasting methods and the adaptability of machine learning algorithms in sales forecasting within the retail sector. Combining these advanced techniques with the CRISP-DM process provides a more robust and flexible analysis, enhancing the efficiency and accuracy of demand forecasts. This enables optimization of inventory management and strategic planning, potentially significantly improving operational performance and company profitability.

The analyzed data comes from the American company Walmart and covers sales from 10 stores across three states: California, Texas, and Wisconsin, over a period of 5 years. Various models were developed to forecast future sales, including a global model trained with all data, and other models specific to each store and state. The global model captures general trends, while the state-specific and store-specific models identify unique patterns of each region and unit, providing more localized forecasts. Additionally, each machine learning model was trained under 4 different scenarios: using total data, data from the last year, the last 3 months, and the last month, aiming to observe the impact of temporal variations on forecasts. Factors such as holidays and days of the week were also considered in the models, as they have a significant impact on sales.

The results showed that among statistical methods, the store-aggregated model with SARIMA was the most effective for sales forecasting in seasonal contexts, achieving an MAE of 224 and a MAPE of 5.39%. Regarding machine learning methods, *Support Vector Machines* (SVM) stood out by presenting high R^2 values, ranging from 77% to 80% for state and store-aggregated models, with relatively low errors. The analysis also revealed that the global aggregated model performed poorly, reflecting a limited ability to capture regional and local specifics. These results highlight the importance of using updated data and adjusting models to recent market variations, emphasizing the need for an adaptive approach in predictive modeling. Comparing different methodologies and including hyperparameters were crucial for forecast accuracy, with results varying depending on the model and application context.

Keywords: CRISP-DM, Forecasting, Statistical Methods, Machine Learning

ÍNDICE

ÍNDICE DE FIGURAS	IX
ÍNDICE DE TABELAS	XI
LISTAS DE SIGLAS E SÍMBOLOS.....	XV
1. INTRODUÇÃO	1
1.1. Enquadramento e pertinência	1
1.2. Questão e objetivos de investigação.....	3
1.3. Opções metodológicas	3
1.4. Apresentação da empresa.....	4
1.5. Estrutura do Documento.....	5
2. REVISÃO BIBLIOGRÁFICA.....	7
2.1. Gestão de Inventário	7
2.1.1. Métodos Tradicionais de Gestão de Inventário.....	9
2.1.2. Características dos Padrões de Procura.....	10
2.2. Previsão da Procura.....	12
2.3. Métodos Tradicionais de Previsão da Procura	14
2.3.1. Média Móvel Simples.....	15
2.3.2. Média Móvel Exponencial.....	16
2.3.3. ARIMA	16
2.3.4. SARIMA.....	17
2.3.5. <i>Holt-Winters</i>	18
2.3.6. ETS.....	19
2.4. Métodos Avançados de Previsão com <i>Machine Learning</i>	20
2.4.1. Regressão Linear Simples.....	20
2.4.2. Regressão Polinomial	21
2.4.3. <i>Decision Tree</i>	22
2.4.4. <i>XGBoost</i>	23
2.4.5. <i>Support Vector Machines</i>	24
2.5. Trabalhos Relacionados.....	24
2.6. Métricas de Avaliação	25
3. MÉTODOS E APLICAÇÃO	28
3.1. <i>Business Understanding</i>	29
3.2. <i>Data Understanding</i>	30
3.3. <i>Data Preparation</i>	37
3.4. <i>Data Modeling</i>	37
3.4.1. Métodos tradicionais de previsão de séries temporais	38
3.4.2. Métodos de previsão utilizando <i>Machine Learning</i>	41

3.5. <i>Evaluation</i>	45
4. RESULTADOS E DISCUSSÃO	47
4.1. Modelação baseada em Métodos Tradicionais.....	47
4.1.1. Resultados do Modelo SARIMA	47
4.1.2. Comparação entre as Abordagens Estatísticas	51
4.2. Modelação baseada em <i>Machine Learning</i>	53
4.2.1. <i>Support Vector Machines (SVM)</i>	53
4.2.2. Comparação entre as Abordagens de Aprendizagem Automática	59
5. CONCLUSÃO	61
5.1. Limitações e investigação futura.....	62
REFERÊNCIAS BIBLIOGRÁFICAS	63
ANEXO A	73

ÍNDICE DE FIGURAS

Figura 1: Gráfico de controlo de inventário	8
Figura 2: Procura estacionária.....	11
Figura 3: Procura com características de crescimento	12
Figura 4: Padrão de procura com sazonalidade	12
Figura 5: Etapas comuns para previsão da procura	14
Figura 6: Estrutura da Árvore de Decisão	22
Figura 7: Estrutura do XGBoost.....	23
Figura 8: Estrutura do modelo SVM	24
Figura 9: Metodologia CRISP-DM.....	28
Figura 10: Gráfico da distribuição das vendas	30
Figura 11: Distribuição das vendas por loja nos estados	31
Figura 12: Gráfico das séries temporais.....	32
Figura 13: Distribuição mensal das vendas	32
Figura 14: Distribuição anual das vendas.....	33
Figura 15: Distribuição semanal das vendas	33
Figura 16: Distribuição das vendas por tipo de feriado	34
Figura 17: Distribuição das vendas por data festiva	34
Figura 18: Influência dos dias festivos nas vendas.....	35
Figura 19: Boxplot do preço de venda dos produtos	36
Figura 20: Gráfico de comparação entre as vendas reais e as previsões utilizando SARIMA.....	50
Figura 21: Gráfico de dispersão entre as vendas reais e previstas com SVM.....	58
Figura A. 1: Gráfico de comparação entre as vendas reais e as previsões utilizando MMS.....	75
Figura A. 2: Gráfico de comparação entre as vendas reais e as previsões utilizando MME.....	78
Figura A. 3: Gráfico de comparação entre as vendas reais e as previsões utilizando ARIMA.....	81
Figura A. 4: Gráfico de comparação entre as vendas Reais e as previsões utilizando Holt-Winters	84
Figura A. 5: Gráfico de Comparação entre as Vendas Reais e as Previsões utilizando ETS	86
Figura A. 6: Gráfico de dispersão entre as vendas reais e previstas com Regressão Linear Simples	92
Figura A. 7: Gráfico de dispersão entre as vendas reais e previstas com Regressão Polinomial.....	98
Figura A. 8: Gráfico de dispersão entre as vendas reais e previstas com Decision Tree	104
Figura A. 9: Gráfico de dispersão entre as vendas reais e previstas com XGBoost	109

ÍNDICE DE TABELAS

Tabela 1: Distribuição total das vendas	30
Tabela 2: Resultados das métricas com SARIMA	48
Tabela 3: Resultados das previsões com SARIMA.....	49
Tabela 4: Comparação entre os modelos agregados com SARIMA	50
Tabela 5: Comparação entre os modelos estatísticos agregados.....	51
Tabela 6: Resultados das métricas do modelo global com SVM.....	53
Tabela 7: Resultados das previsões do modelo global com SVM	54
Tabela 8: Resultados das métricas do modelo por estado com SVM.....	54
Tabela 9: Resultados das previsões do modelo por estado com SVM.....	55
Tabela 10: Resultados das métricas do modelo por loja com SVM	56
Tabela 11: Resultados das previsões do modelo por loja com SVM.....	57
Tabela 12: Comparação entre os modelos agregados com SVM (Cenário 3).....	58
Tabela 13: Comparação entre os modelos agregados de aprendizagem automática (Cenário 3) ..	59
Tabela A. 1: Resultados das métricas com Média Móvel Simples	74
Tabela A. 2: Resultados das previsões com Média Móvel Simples.....	74
Tabela A. 3: Comparação entre os modelos agregados com MMS	75
Tabela A. 4: Resultados das métricas com MME	76
Tabela A. 5: Resultados das previsões com MME.....	77
Tabela A. 6: Comparação entre os modelos agregados com MME	78
Tabela A. 7: Resultados das métricas com ARIMA.....	79
Tabela A. 8: Resultados das previsões com ARIMA	80
Tabela A. 9: Comparação entre os modelos agregados com ARIMA.....	81
Tabela A. 10: Resultados das métricas com Holt-Winters	82
Tabela A. 11: Resultados das previsões com Holt-Winters.....	83
Tabela A. 12: Comparação entre os modelos agregados com Holt-Winters	84
Tabela A. 13: Resultados das métricas com ETS	85
Tabela A. 14: Resultados das previsões com ETS.....	85
Tabela A. 15: Comparação entre os modelos agregados com ETS	86
Tabela A. 16: Resultados das métricas do modelo global com Regressão Linear Simples	87
Tabela A. 17: Resultados das previsões do modelo global com Regressão Linear Simples.....	87
Tabela A. 18: Resultados das métricas do modelo por estado com Regressão Linear Simples	88
Tabela A. 19: Resultados das previsões do modelo por estado com Regressão Linear Simples	88
Tabela A. 20: Resultados das métricas do modelo por loja com Regressão Linear Simples.....	90
Tabela A. 21: Resultados das previsões do modelo por loja com Regressão Linear Simples	91
Tabela A. 22: Comparação entre os modelos agregados com Regressão Linear Simples (cenário 3)	92
Tabela A. 23: Resultados das métricas do modelo global com Regressão Polinomial	93
Tabela A. 24: Resultados das previsões do modelo global com Regressão Polinomial	93
Tabela A. 25: Resultados das métricas do modelo por estado com Regressão Polinomial.....	94
Tabela A. 26: Resultados das previsões do modelo por estado com Regressão Polinomial	94
Tabela A. 27: Resultados das métricas do modelo por loja com Regressão Polinomial – Parte 1 ..	95

Tabela A. 28: Resultados das métricas do modelo por loja com Regressão Polinomial – Parte 2 ..	96
Tabela A. 29: Resultados das previsões do modelo por loja com Regressão Polinomial – Parte 1.	96
Tabela A. 30: Resultados das previsões do modelo por loja com Regressão Polinomial – Parte 2.	97
Tabela A. 31: Comparação entre os modelos agregados com Regressão Polinomial (cenário 4) ...	97
Tabela A. 32: Resultados das métricas do modelo global com Decision Tree	99
Tabela A. 33: Resultados das previsões do modelo global com Decision Tree.....	99
Tabela A. 34: Resultados das métricas do modelo por estado com Decision Tree – Parte 1.....	99
Tabela A. 35: Resultados das métricas do modelo por estado com Decision Tree – Parte 2.....	100
Tabela A. 36: Resultados das previsões do modelo por estado com Decision Tree	100
Tabela A. 37: Resultados das métricas do modelo por loja com Decision Tree	101
Tabela A. 38: Resultados das previsões do modelo por loja com Decision Tree	102
Tabela A. 39: Comparação entre os modelos agregados com Decision Tree (cenário 3).....	103
Tabela A. 40: Resultados das métricas do modelo global com XGBoost	105
Tabela A. 41: Resultados das previsões do modelo global com XGBoost.....	105
Tabela A. 42: Resultados das métricas do modelo por estado com XGBoost	105
Tabela A. 43: Resultados das previsões do modelo por estado com XGBoost	106
Tabela A. 44: Resultados das métricas do modelo por loja com XGBoost.....	107
Tabela A. 45: Resultados das previsões do modelo por loja com XGBoost	108
Tabela A. 46: Comparação entre os modelos agregados com XGBoost (cenário 3).....	109

LISTAS DE SIGLAS E SÍMBOLOS

Lista de Siglas

CRISP-DM	<i>Cross Industry Standard Process for Data Mining</i>
ARIMA	<i>Autoregressive Integrated Moving Average</i>
SARIMA	<i>Seasonal Autoregressive Integrated Moving Average</i>
MMS	Média Móvel Simples
MME	Média Móvel Exponencial
ETS	<i>Error Trend and Seasonality</i>
MAE	<i>Mean Absolute Error</i>
MAPE	<i>Mean Absolute Percentage Error</i>
MSE	<i>Mean Squared Error</i>
RMSE	<i>Root Mean Squared Error</i>
R ²	Coeficiente de Correlação
SVM	<i>Support Vector Machine</i>
XGBoost	<i>Extreme Gradient Boosting</i>
CA	Califórnia
TX	Texas
WI	Wisconsin

1. INTRODUÇÃO

A área de Engenharia e Gestão da Cadeia de Abastecimento tem vindo a ganhar crescente relevância, com um foco crescente na otimização e eficiência dos processos logísticos. O avanço das metodologias nesta área é fundamental para melhorar a eficiência operacional, reduzir custos e promover a sustentabilidade ao longo de toda a cadeia de abastecimento. Com a evolução constante das necessidades do mercado e das tecnologias disponíveis, é imperativo que as práticas e abordagens na gestão da cadeia de abastecimento se adaptem para manter a competitividade e a excelência no setor.

Este capítulo introduz o problema de investigação e as questões que este trabalho se propõe a abordar. Inicialmente, realiza-se uma contextualização do tema, sublinhando a sua importância no contexto atual e a relevância para a comunidade científica e prática profissional. Em seguida, são apresentados os objetivos da investigação, tanto gerais quanto específicos, e é formulada a questão central da pesquisa. A metodologia a ser seguida é também descrita, detalhando as abordagens e técnicas que serão empregues. Finalmente, o capítulo oferece uma visão geral da empresa que serve de base para este projeto.

1.1. Enquadramento e pertinência

Definir a quantidade ideal de stock a ser alocada aos pontos de venda é uma decisão estratégica crucial para as empresas de retalho. Silva (2023) defende que esta escolha tem um impacto significativo, uma vez que a distribuição inadequada pode não só resultar na perda de oportunidades de venda, como também em custos elevados com armazenamento e na imobilização de capital. Para lidar com esta complexidade, Thomas (2019), ao estudar a previsão da procura na indústria de rolhas, destacou a importância de realizar previsões precisas, considerando vários factores, como flutuações sazonais e padrões de consumo. Neste contexto, a utilização de modelos de previsão, especialmente aqueles baseados em técnicas avançadas de aprendizagem automática, é essencial. Estes modelos oferecem uma base robusta para a tomada de decisões eficientes, assegurando assim a saúde financeira da empresa.

O setor de retalho enfrenta desafios consideráveis na previsão da procura, influenciados pela variação das preferências dos consumidores, sazonalidade e tendências emergentes. A complexidade surge da natureza não linear das interações entre produtos e da necessidade de adaptar-se rapidamente às mudanças nos padrões de consumo. Além disso, fatores sazonais, alterações nas preferências dos consumidores e influências externas complicam ainda mais o processo de previsão (Wacker & Lummus, 2002; Yelland, 2010). Estudos anteriores, como o de Stergiou (1991), que comparou modelos de suavização exponencial, ARIMA e modelos de regressão, demonstram que os métodos tradicionais, baseados em dados históricos de vendas, possuem limitações e requerem a adoção de abordagens mais flexíveis e inovadoras para assegurar previsões precisas e sustentáveis a longo prazo no setor de retalho.

Bayer (2012) verificou que a previsão da procura, quando se baseia apenas em métodos tradicionais de séries temporais e modelos estatísticos, enfrenta desafios significativos devido à escassez e incompletude dos dados. No setor de retalho, onde as preferências dos consumidores são voláteis, técnicas como suavização exponencial, ARIMA e modelos de regressão têm limitações (Pawan et al., 2019). Para superar essas limitações e planejar vendas com antecedência, Trapero (2013) cita que é essencial utilizar abordagens inovadoras, como algoritmos de aprendizagem automática e inteligência artificial. Métodos avançados, como Random Forest, Decision Trees e Máquinas de Vetores de Suporte (SVM), oferecem previsões mais precisas e adaptadas às necessidades do setor retalhista. Estes métodos melhoram a gestão de stocks, otimizam o surtimento e aumentam a rentabilidade, contribuindo para uma maior eficiência operacional.

A implementação de estratégias baseadas em aprendizagem automática enfrenta desafios significativos, especialmente na ausência de um histórico substancial de vendas. Pawan (2019) sugere que a combinação de modelos de aprendizagem automática com métodos estatísticos tradicionais pode ajudar a superar essas dificuldades, garantindo a fiabilidade das previsões. Para assegurar a precisão, é essencial realizar uma avaliação rigorosa do desempenho, utilizando técnicas como validação cruzada e análise de resíduos, além de métricas como Erro Médio Absoluto (MAE), Erro Quadrático Médio (MSE), Coeficiente de Determinação (R^2), entre outras. Esta abordagem metodológica robusta assegura uma análise completa e confiável dos modelos.

A complexidade do comportamento do consumidor, caracterizada pela sua variabilidade e não linearidade, destaca a importância dessas estratégias avançadas para previsões precisas. Trapero (2013) argumenta que técnicas aprimoradas são essenciais para melhorar a exatidão das previsões, fornecendo informações valiosas para otimizar as operações na cadeia de abastecimento. A previsão da procura é crucial para a gestão de inventário, a realização de encomendas e a adaptação às mudanças no comportamento do consumidor, contribuindo para uma maior eficiência operacional, rentabilidade e sustentabilidade no retalho.

Dentro deste contexto, a escolha entre um modelo global de previsão ou modelos específicos por loja e região é crucial para otimizar a gestão de inventário e a rentabilidade. Hyndman et al. (2018) argumentam que modelos globais, que utilizam dados agregados de todas as lojas e regiões, oferecem simplicidade e consistência, mas podem não capturar as nuances locais do comportamento do consumidor. Em contraste, modelos específicos permitem uma personalização detalhada, ajustando-se às características únicas de cada local e potencialmente proporcionando previsões mais precisas. Kumar e Reinartz (2016) confirmam em seu estudo que, enquanto os modelos específicos frequentemente resultam em previsões mais precisas devido à sua capacidade de refletir variações regionais, eles também envolvem maior custo e complexidade. A decisão sobre qual modelo adotar deve, portanto, levar em conta o contexto e as necessidades específicas da empresa. Modelos globais podem ser mais eficientes em termos de custo e implementação, enquanto modelos específicos podem justificar o investimento adicional pela maior precisão e adaptabilidade. É fundamental avaliar cuidadosamente as características e requisitos de cada situação para encontrar o equilíbrio ideal entre precisão, custo e complexidade, alinhando-se ao objetivo de aprimorar a gestão de inventário, maximizar a rentabilidade e melhorar a capacidade de resposta às variações no comportamento do consumidor (Guan et al. 2015).

1.2. Questão e objetivos de investigação

Diante do desafio de aprimorar a previsão da procura em pontos de venda do setor de retalho utilizando técnicas de aprendizagem automática, a questão central desta investigação é: "Como as abordagens tradicionais (estatísticas) e as baseadas em aprendizagem automática podem otimizar a previsão da procura no setor de retalho, considerando a complexidade do comportamento do consumidor e os desafios específicos do mercado? Além disso, qual é a estratégia de modelagem mais eficaz: utilizar um modelo global que agrega dados de todas as lojas e estados, ou adotar previsões segmentadas por loja e estado, treinadas com dados locais específicos?"

Para responder a esta questão, o objetivo geral desta pesquisa é melhorar a precisão das previsões de procura em pontos de venda do setor de retalho através da implementação de técnicas de aprendizagem automática e modelos estatísticos, avaliando quais métodos podem oferecer um desempenho superior para a empresa. Além disso, pretende-se investigar se a estratégia de modelagem tem um impacto significativo na precisão das previsões, comparando um modelo global com modelos específicos agregados por loja e por estado.

Para atingir o objetivo geral, os seguintes objetivos específicos foram delineados:

- Analisar e comparar diferentes algoritmos de aprendizagem automática para identificar a abordagem mais adequada à previsão da procura, considerando a utilização de modelos globais e modelos específicos agregados por loja e por estado;
- Analisar e comparar diferentes métodos tradicionais (estatísticos) de previsão da procura a partir da análise de métricas e identificar qual dos modelos agregados apresenta um melhor desempenho;
- Analisar o impacto de diferentes janelas de dados na performance dos modelos de aprendizagem de máquina, utilizando dados totais, dados do último ano, dados dos últimos 3 meses e dados do último mês;
- Analisar os resultados obtidos e escolher o melhor modelo, tendo em conta a estratégia de modelagem e a janela de dados utilizada.

Um estudo experimental de previsão da procura será conduzido com os dados de vendas da empresa norte-americana Walmart, disponibilizados em um banco de dados público.

1.3. Opções metodológicas

Neste estudo, as escolhas metodológicas foram criteriosamente selecionadas para garantir a eficácia na realização dos objetivos propostos. Dada a natureza da investigação sobre a previsão da procura em pontos de venda no setor do retalho, foi adotada uma abordagem quantitativa, focada na manipulação de variáveis numéricas e na utilização de métodos estatísticos, modelação matemática e técnicas de aprendizagem automática para formular previsões precisas e identificar padrões nos dados de vendas, conforme destacado por Knechtel em 2014.

A metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*) foi adotada para estruturar o processo de análise de dados de forma sistemática e eficaz. Vásquez (2024) explica que o CRISP-DM é uma abordagem robusta que orienta a execução de projetos de mineração de dados através de um ciclo de vida estruturado, que inclui as fases de compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implementação. A escolha do CRISP-DM permite uma abordagem iterativa e incremental, garantindo que os modelos desenvolvidos sejam ajustados e validados continuamente para cumprir os objetivos da investigação e fornecer soluções práticas e baseadas em dados para a previsão da procura.

Relativamente à abordagem de investigação, este projeto seguirá uma abordagem hipotético-dedutiva, combinando elementos indutivos e dedutivos. A estratégia consistirá em orientar a análise inicialmente pelos dados disponíveis da empresa, permitindo a identificação de padrões de forma indutiva. Posteriormente, será formulada uma hipótese com base na análise dos dados, seguindo uma abordagem dedutiva, e serão realizados testes para validar essa hipótese. Se necessário, serão feitos ajustes na teoria com base nos resultados obtidos. Este ciclo de ajustes entre a teoria e a evidência empírica contribuirá para o desenvolvimento de uma compreensão mais precisa da dinâmica da procura no setor do retalho, tal como proposto por Walliman (2011).

Relativamente à natureza da pesquisa, a abordagem escolhida é analítica, fundamentada na necessidade de construir uma solução quantificável que também forneça uma compreensão profunda e explicativa do seu funcionamento. O projeto visa não apenas prever a procura, mas também oferecer percepções que permitam uma adaptação eficaz no contexto real, tornando a pesquisa analítica uma abordagem sólida e alinhada com os objetivos propostos. Esta metodologia permite uma visão holística do comportamento do consumidor, capacitando a tomada de decisões para a otimização da cadeia de abastecimento e a maximização da eficiência operacional.

Por fim, o procedimento de pesquisa adotado para este trabalho é o estudo de caso, escolhido devido à necessidade de uma análise detalhada e específica do contexto no setor do retalho. Esta abordagem permite investigar relações causais e compreender as complexidades envolvidas na previsão da procura num ponto de venda de retalho. Ao concentrar-se num caso específico, o estudo de caso oferece compreensões valiosas e contribui para a aplicação prática das soluções propostas, alinhando-se com os objetivos do projeto e proporcionando uma compreensão mais rica e contextualizada dos desafios e oportunidades associados (Lüdke e André, 1999).

As ferramentas utilizadas neste trabalho incluíram a implementação em *Python* na plataforma *Visual Studio Code* com o *Jupyter Notebook*. Para construir a aplicação, foram empregadas várias bibliotecas essenciais: *pandas* e *numpy* para a manipulação e análise de dados; *matplotlib* e *seaborn* para a visualização e criação de gráficos; *scikit-learn* para a implementação de modelos de *machine learning*, incluindo classificação, regressão e métodos de *ensemble*; e *statsmodels* e *statistics* para análises e cálculos estatísticos.

1.4. Apresentação da empresa

A Walmart Inc., fundada em 1962 por Sam Walton, é uma das maiores cadeias de retalho do mundo, com sede em Bentonville, Arkansas. A empresa opera uma vasta rede de lojas, que inclui hipermercados, supermercados e clubes de desconto, destacando-se pela política de "preços

baixos todos os dias". Com mais de 10.000 lojas a nível global e milhões de funcionários, a Walmart oferece uma ampla gama de produtos, desde alimentos e vestuário até eletrónica e artigos para o lar. Além da sua forte presença física, a empresa dispõe de uma plataforma digital robusta que complementa as lojas físicas. A Walmart compromete-se com a sustentabilidade e a responsabilidade social, investindo em práticas para reduzir o impacto ambiental e apoiar as comunidades locais. A companhia continua a adaptar e inovar a sua estratégia para manter-se competitiva no dinâmico setor de retalho (Walmart, 2024).

Para a Walmart, uma cadeia de abastecimento eficiente e uma previsão precisa da procura são essenciais para o sucesso. A gestão eficaz da cadeia de abastecimento permite à empresa reduzir custos, otimizar a logística e garantir a disponibilidade de produtos. A previsão da procura ajuda a Walmart a equilibrar oferta e demanda, evitando excessos ou faltas de stock e mantendo a satisfação do cliente. Esta combinação é crucial para a empresa manter a sua estratégia de preços baixos e a sua posição de liderança no mercado global (Walmart, 2024).

1.5. Estrutura do Documento

A estrutura deste documento é composta por cinco capítulos. O primeiro capítulo, dedicado à introdução, apresenta uma análise do problema em estudo, detalha a escolha metodológica e fornece uma contextualização abrangente da empresa alvo da pesquisa.

No segundo capítulo, é realizada uma revisão sistemática da literatura, abordando e analisando estudos académicos relacionados com a previsão e gestão de inventário. Esta fase tem como objetivo identificar e compreender as técnicas de aprendizagem automática mais relevantes para a previsão da procura no setor de retalho, oferecendo uma base teórica sólida que orientará a escolha dos métodos mais adequados para a investigação.

O terceiro capítulo foca na seleção e preparação dos dados quantitativos necessários para alimentar os modelos de aprendizagem automática. Serão utilizados conjuntos de dados representativos do setor, incluindo variáveis como histórico de vendas, características dos produtos e fatores sazonais. Além disso, será feita uma análise geral do contexto atual da empresa.

No quarto capítulo, são apresentados e discutidos detalhadamente os resultados obtidos, permitindo uma análise quantitativa entre os diversos métodos estatísticos e de aprendizagem automática utilizados. Serão realizadas análises estatísticas para avaliar a precisão e eficácia dos modelos de aprendizagem automática em comparação com os métodos tradicionais de previsão. Métricas como erro médio absoluto (MAE), erro quadrático médio (MSE) e outras medidas estatísticas serão empregues para avaliar o desempenho dos modelos.

Por fim, o quinto capítulo resume as conclusões derivadas do trabalho e oferece uma perspetiva mais ampla sobre possíveis desenvolvimentos futuros, destacando áreas para futuras pesquisas e melhorias na compreensão da previsão e gestão de inventário em contextos específicos.

2. REVISÃO BIBLIOGRÁFICA

Neste capítulo, realiza-se uma revisão bibliográfica para contextualizar o trabalho, focando-se na previsão da procura em pontos de venda para uma empresa de retalho, através da análise de métodos estatísticos e da criação de algoritmos baseados em *Machine Learning*.

Na primeira secção, aborda-se a gestão de inventário, destacando métodos tradicionais como o Modelo de Revisão Contínua, o Modelo de Revisão Periódica e a Classificação ABC (Christiansen, 2024). São explorados desafios comuns, como a variabilidade da procura, o lead time e a obsolescência de stock, fundamentais para a compreensão da gestão eficaz de inventário numa empresa (Singh, 2018).

Na segunda secção, exploram-se os fundamentos teóricos relacionados com os modelos estatísticos e os algoritmos de aprendizagem automática aplicados à previsão da procura. Conceitos-chave, como o aprendizado supervisionado e não supervisionado, são discutidos para compreender o desenvolvimento do algoritmo proposto. A revisão detalhada da literatura nesta área serve como uma base sólida para a formulação e implementação de técnicas avançadas de *Machine Learning* (Holzinger et al. 2018). Esta secção também irá explorar as métricas relevantes para a análise dos resultados e a seleção do melhor modelo de previsão para a empresa.

2.1. Gestão de Inventário

A gestão de inventário é uma peça-chave para as operações comerciais, exercendo um impacto significativo na eficiência das cadeias de abastecimento e na satisfação do cliente. Não se resume apenas a manter registos precisos dos produtos disponíveis; trata-se de encontrar o equilíbrio ideal entre ter os itens certos no momento certo, evitando excessos ou escassez (Singh, 2018). Este processo envolve a optimização de variáveis como tempo de reposição, custos associados à manutenção de inventário, previsão da procura, valoração de stock e visibilidade do inventário. Giannoccaro (2022) argumenta que alcançar um equilíbrio estratégico entre esses componentes é importante para garantir níveis ótimos de stock. Integrando estratégias de gestão de inventário na cadeia de abastecimento, procura-se não apenas atender às procuras dos clientes, mas também ampliar a variedade de produtos disponíveis e reduzir custos operacionais.

No cenário competitivo atual, onde as procuras do mercado podem ser voláteis, uma gestão eficaz de inventário oferece uma série de vantagens operacionais e estratégicas (Sarac et al., 2010). Vários estudos de casos reais destacam a importância de uma gestão de inventário sólida, como o trabalho de Zwaida (2021), que desenvolveu um modelo de Reinforcement Learning capaz de tomar decisões automáticas de reabastecimento de medicamentos para evitar a escassez na cadeia de abastecimento hospitalar, garantindo um atendimento de saúde ao cliente mais fiável e eficaz.

A gestão de inventário suaviza o fluxo de mercadorias, minimizando interrupções e permitindo uma adaptação ágil às mudanças nas procuras do mercado. Em essência, é a cola que mantém os elos da cadeia unidos e eficientes. O advento de tecnologias inovadoras tem sido um catalisador neste cenário, impulsionando a gestão de inventário para além da simples contabilidade de

produtos. Sistemas de automação, RFID e software de inteligência artificial oferecem visibilidade, precisão e automação, transformando a forma como as empresas gerem os seus stocks (Sarac et al., 2010). Um estudo de Baars et al. (2008) usou a combinação de RFID e inteligência de negócios para proporcionar visibilidade em tempo real do stock, permitindo análise detalhada dos padrões de procura e resposta ágil a mudanças. Os resultados destacaram uma redução nos custos operacionais, a partir da minimização de stocks obsoletos, e uma melhoria na otimização da gestão de inventário, o que contribuiu para a eficiência operacional geral da empresa, garantindo melhor utilização dos recursos e maior capacidade de atender às procuras dos clientes.

A implementação de uma gestão de inventário eficaz envolve práticas fundamentais, como análise de dados históricos de vendas, previsão de procura, monitorização dos níveis de stock e uso de métricas de desempenho. Além dos benefícios imediatos, essa eficiência constrói uma base para a sustentabilidade e competitividade a longo prazo, reduzindo desperdícios e minimizando o impacto ambiental (Silver, 1981). Um estudo realizado por Paam et al. (2019) examinou como otimizar a gestão de inventário em cadeias de abastecimento de maçãs para reduzir a perda de frutas e os custos de inventário e processamento. Eles analisaram o impacto da configuração de salas de armazenamento de alta tecnologia, usando uma variável de decisão baseada no intervalo de tempo entre a colheita e a entrega. Os resultados salientaram como práticas eficientes de gestão de stock podem contribuir para a sustentabilidade económica e ambiental da indústria.

Um gráfico de controlo de inventário é uma ferramenta essencial para a gestão eficaz dos níveis de stock ao longo do tempo. Nele, encontramos diversos elementos que desempenham papéis cruciais na manutenção de um equilíbrio adequado e eficiente no armazenamento de produtos, que podemos observar na Figura 1.

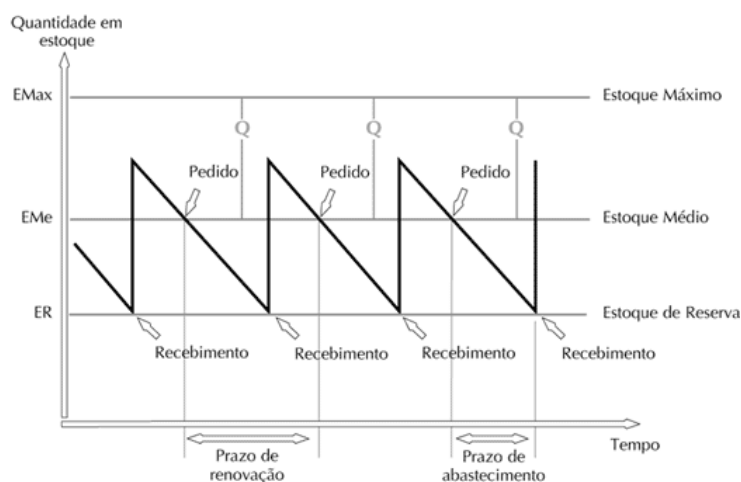


Figura 1: Gráfico de controlo de inventário
Fonte: Ávila, 2015

O stock máximo estabelece o limite superior desejado de unidades em stock, evitando excessos que resultem em custos desnecessários. Por outro lado, o stock mínimo define o ponto mais baixo aceitável para evitar escassez e garantir a disponibilidade de produtos. Quando o stock atinge o ponto de reordenação, geralmente definido a um nível intermédio, uma nova ordem de reposição é iniciada. Um conceito crítico associado a este gráfico é o tempo de reposição, que representa o período entre o início de uma nova ordem de reposição e a disponibilidade dos produtos para uso.

Além disso, é essencial considerar o stock de segurança ou buffer stock, que é uma quantidade adicional além do stock mínimo, destinada a lidar com incertezas, como flutuações na procura ou possíveis atrasos no tempo de reposição (Souza, 2020).

A interpretação de gráficos de gestão de inventário é fundamental para evitar custos desnecessários e garantir a satisfação do cliente, possibilitando uma administração eficaz dos custos relacionados com o armazenamento e pedidos. Além disso, a definição de Indicadores Chave de Desempenho (KPIs) é crucial para uma gestão eficaz do inventário em empresas de retalho, permitindo medir níveis de stock, velocidade de vendas e considerar factores como procura sazonal e tendências do consumidor. A compreensão desses elementos proporciona percepções valiosas para a tomada de decisões, evitando excessos ou faltas de stock e garantindo a adaptação ágil às mudanças no mercado de retalho (AQA, 2010).

Uma gestão eficiente de inventário entrega valor ao cliente e garante a lucratividade das organizações, envolvendo a optimização dos recursos para atender à procura de forma eficaz e evitar rupturas de stock. Para isso, é necessário o uso de algoritmos de previsão de vendas precisos, que considerem factores como sazonalidade e tendências de mercado. A capacidade de adaptação rápida a mudanças e a resolução proativa de desafios são fundamentais para o sucesso desta gestão e, conseqüentemente, da cadeia de abastecimento (Olson et al., 2010).

2.1.1. Métodos Tradicionais de Gestão de Inventário

De acordo com a *Rangel Logistics Solutions* (2023), os métodos tradicionais de gestão de inventário têm sido aplicados ao longo do tempo para equilibrar o desafio de manter níveis adequados de stock, o que influencia a capacidade de responder à procura do mercado enquanto optimiza os recursos disponíveis. Estes métodos oferecem estratégias para as empresas gerirem a aquisição, armazenamento e distribuição de produtos, visando garantir a disponibilidade adequada de mercadorias sem incorrer em custos excessivos. Desde a clássica abordagem da Quantidade Económica de Encomenda (EOQ) até aos modelos ABC e técnicas de revisão contínua, cada método tradicional possui características distintas que atendem a diferentes contextos e necessidades.

Entre os métodos tradicionais, destaca-se a Quantidade Económica de Encomenda (EOQ). Segundo Minhaz (2024), ao usar a previsão de procura como base para calcular a EOQ, as empresas podem optimizar os seus níveis de stock, encomendando a quantidade certa no momento certo para equilibrar os custos de encomenda e os custos de manutenção de stock. É uma abordagem mais eficaz em cenários com procura constante e preços estáveis, sendo por isso frequentemente combinada com outros métodos para uma gestão de stock mais abrangente. Num estudo conduzido por Aju et al. (2013), foi desenvolvido um modelo eficaz de previsão da procura de fertilizantes, utilizando tanto redes neurais como suavização exponencial, juntamente com a determinação da quantidade económica de encomenda. A implementação deste modelo resultou numa redução de 20% nas rupturas de stock e nos custos associados aos produtos mais vendidos.

A EOQ baseia-se num cálculo que determina a quantidade ideal a encomendar, minimizando os custos de gestão de inventário. Estes custos incluem tanto os de encomenda (processamento e transporte) como os de manutenção (armazenamento e deterioração), e o cálculo equilibra ambos

para garantir o menor custo total. Assim, as empresas evitam a falta ou excesso de stock, encomendando na quantidade e no momento certos (Harris, 1913).

Já os modelos ABC classificam os itens de inventário com base na sua importância, valor monetário e frequência de movimentação, permitindo uma alocação eficiente de recursos para itens críticos. A Classe “A” inclui itens de alta prioridade, geralmente de alto valor e alta rotatividade. A Classe “B” representa itens de prioridade moderada, enquanto a Classe “C” envolve itens de baixa prioridade, geralmente de baixo valor e menor rotatividade. Castanho (2023) defende que esta prática ajuda a otimizar recursos, reduzir custos e melhorar a eficiência geral da gestão de stock, já que concentra um maior esforço nos itens mais críticos. Hung et al. (2012) utilizaram um modelo ABC numa empresa de fechaduras na Tailândia para resolver o problema de escassez de stock de matéria-prima para o processo de fabricação. A técnica foi utilizada para identificar os produtos múltiplos mais importantes e, em seguida, foi estabelecido um modelo para definir a quantidade económica de encomenda para solucionar o problema de escassez.

Após a classificação, as empresas podem implementar políticas específicas de gestão para cada categoria. Teunter (2010) explica que para os itens da Classe A, adotam-se medidas de controlo rigorosas, como encomendas frequentes e um monitoramento contínuo para evitar roturas de stock, dada sua importância financeira. Os itens da Classe B recebem uma gestão moderada, com reposições e análises periódicas, enquanto os da Classe C, por sua menor relevância, são geridos com menos frequência e com políticas de reabastecimento mais simples. Esta segmentação permite que as empresas concentrem esforços e recursos nos itens que mais impactam o desempenho financeiro, melhorando a disponibilidade dos produtos cruciais sem desperdiçar recursos em itens de menor valor (Dhoka, 2013).

As técnicas de revisão contínua, como o Sistema de Reposição Contínua (ROP), monitorizam constantemente os níveis de stock, repondo-os quando atingem um ponto pré-determinado. Em contraste, o Sistema de Revisão Periódica realiza avaliações regulares em intervalos específicos para ajustar os níveis de stock. Ambas as abordagens têm as suas vantagens, sendo o ROP eficaz para itens de procura constante, enquanto a Revisão Periódica é mais adequada para produtos com procura variável. A escolha entre estas técnicas depende da natureza dos produtos e das necessidades da cadeia de abastecimento, visando uma gestão eficaz de inventário (Rosa, 2010).

Cada método tradicional possui características distintas que atendem a diferentes contextos e necessidades na gestão de inventário. No entanto, à medida que nos deparamos com sectores dinâmicos como o sector do retalho, marcado por padrões irregulares e alta variabilidade nas vendas, torna-se evidente a limitação destas abordagens convencionais. A complexidade do comportamento do consumidor, caracterizado pela sua variabilidade e não linearidade, destaca a necessidade de explorar métodos mais avançados e introduzir técnicas mais sofisticadas, como algoritmos de *machine learning* na previsão da procura, que permitirão uma abordagem mais proativa, eficiente e alinhada com as necessidades específicas do sector (Christiansen, 2024).

2.1.2. Características dos Padrões de Procura

Segundo Oliveira (2019), os sistemas de controlo de inventário enfrentam uma variedade de padrões de procura dos clientes, cada um requerendo previsões para uma eficiente política global

de controlo de stock. As séries temporais são o conjunto de dados organizados cronologicamente, nos quais as observações são registadas ao longo do tempo, representando uma variedade de fenómenos do mundo real, permitindo análises e previsões para compreender padrões e tendências temporais.

A tendência numa série temporal refere-se à direção geral em que os dados se movem ao longo do tempo, podendo ser ascendente, descendente ou estável. Identificar a tendência é crucial para antecipar possíveis mudanças a longo prazo e compreender o comportamento subjacente dos dados. Já a sazonalidade representa padrões que se repetem em intervalos fixos de tempo, como estações do ano, meses, semanas ou dias da semana (Silva, 2021).

A estacionaridade, no entanto, refere-se à estabilidade estatística dos dados ao longo do tempo. A procura estacionária é um padrão característico na gestão de inventário, representando uma condição em que as flutuações na procura dos clientes ocorrem ao longo do tempo, sem apresentar uma tendência de crescimento ou sazonalidade discernível a longo prazo. Nesse contexto, assume-se que, embora a procura por unidade de tempo possa variar, não há uma tendência subjacente aparente de expansão ou contração, facilitando análises estatísticas e previsões mais precisas. A Figura 2 mostra um padrão de procura estacionário sem tendência de crescimento ou decréscimo, e sem sazonalidade, mas com variações. (N. Li, 2015)

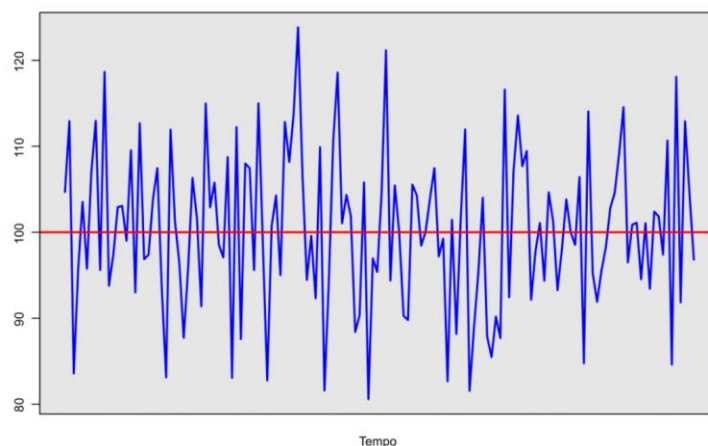


Figura 2: Procura estacionária
Fonte: Características das séries temporais (Oliveira, 2019)

Existem também padrões de procura que demonstram características de crescimento, como mostra a Figura 3, e é a partir desses que os modelos de previsão necessários se tornam mais complexos em comparação com situações de procura estacionária. Uma compreensão aprofundada desses padrões é crucial para desenvolver estratégias de previsão eficazes, considerando tanto o crescimento como possíveis declínios ao longo do tempo.

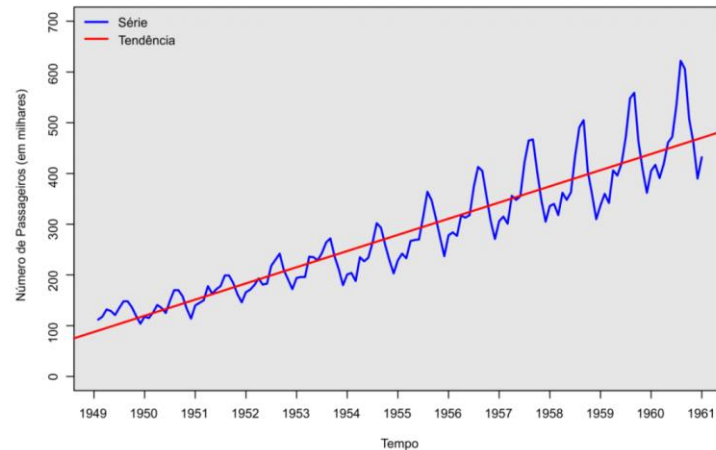


Figura 3: Procura com características de crescimento
 Fonte: Características das séries temporais (Oliveira, 2019)

Para Colin (1997), a procura com características sazonais, como mostrado na Figura 4, refere-se a padrões de procura que são influenciados pelas diferentes estações do ano e por eventos anuais específicos. Nestes casos, é possível avaliar o quanto a procura num período específico do ano difere da média típica. Os modelos de previsão que incorporam a sazonalidade procuram estabelecer essa relação para cada período do ano, utilizando factores de dessazonalização identificados durante esse processo para gerar previsões. Esta abordagem tem como objetivo otimizar a precisão no planeamento de inventário e na gestão da cadeia de abastecimento.

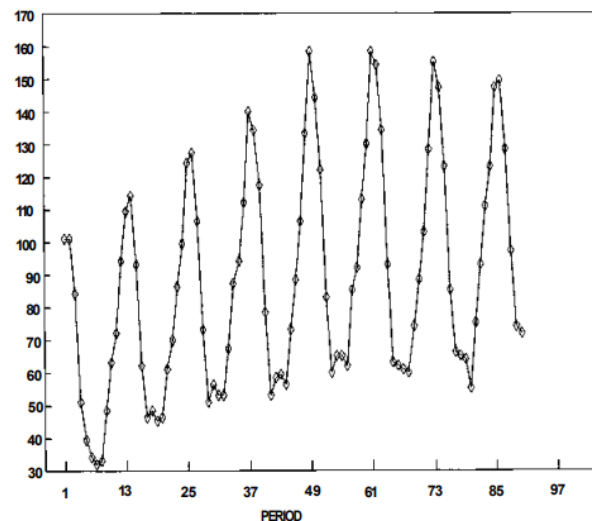


Figura 4: Padrão de procura com sazonalidade
 Fonte: Demand Forecasting and Inventory Control (Colin, 1997)

2.2. Previsão da Procura

A previsão da procura influencia diretamente o planeamento das atividades empresariais. Em indústrias de retalho, caracterizadas por dinâmicas únicas, como lançamentos frequentes de novos produtos e procura imprevisível, previsões inadequadas podem resultar em desafios como escassez

de stock, obsolescência e ineficiência nos recursos. Segundo Nenni et al., em 2013, perante este cenário, as empresas têm orientado as suas estratégias de gestão para criar estruturas que permitam a produção com base na procura em tempo real, incorporando também tecnologias avançadas de previsão.

Em empresas de retalho, é mais comum adotar a estratégia "*make-to-stock*" (MTS). Nesta abordagem, como Segal mencionou em 2022, os produtos são produzidos em massa com base em previsões de procura e mantidos em stock para pronta entrega aos clientes, permitindo que as empresas respondam rapidamente às necessidades do mercado e garantam elevados níveis de disponibilidade de produtos aos consumidores. No entanto, para que esta estratégia seja eficaz, é essencial uma previsão precisa da procura, visto que os pontos de venda dependem também da capacidade de antecipar as tendências de compra dos consumidores. Jenkins, em 2022, destacou que isto envolve a análise de dados históricos de vendas, a consideração da sazonalidade, as tendências de mercado e até fatores externos, como eventos promocionais ou mudanças nas preferências dos consumidores.

Portanto, na busca por aprimorar a previsão de procura, Nalini (2024) salientou que é fundamental que as empresas invistam em tecnologias avançadas, promovendo a colaboração entre as equipas de previsão, marketing e vendas, e mantendo um constante monitoramento e adaptação aos sinais do mercado. Compreender profundamente estes elementos capacita as empresas a antecipar e responder eficientemente às mudanças no mercado, permitindo ajustes estratégicos contínuos para maximizar a eficácia operacional e atender de forma mais eficaz às necessidades dos clientes.

Anusha et al. (2014) realizaram um estudo sobre a previsão da procura numa grande empresa de retalho farmacêutico na Índia, que enfrentava desafios relacionados com o excesso ou falta de stock. Utilizaram modelos estatísticos como a Média Móvel, a Suavização Exponencial e *Holt-Winters*, concluindo que o método *Holt-Winters* é confiável para produtos sazonais, enquanto a Média Móvel de 6 meses se revelou eficaz para produtos não sazonais. Por outro lado, o trabalho de Priyadarshi et al. (2019) teve como objetivo selecionar o melhor modelo de previsão da procura no retalho para vegetais selecionados. Foram propostos e aplicados vários modelos, incluindo técnicas de aprendizagem de máquina como redes LSTM e SVR. Os resultados mostraram que esses algoritmos superaram outros modelos de previsão em termos de desempenho.

Apesar da existência de numerosos estudos nesta área, a previsão da procura pode enfrentar diversos desafios nas empresas. A falta de dados adequados é um obstáculo comum, resultando em previsões imprecisas, assim como a superdependência de dados passados pode levar a projeções inadequadas, uma vez que as condições de mercado evoluem constantemente. A limitada visibilidade para o futuro também dificulta a antecipação da procura, e muitos modelos acabam por ser inexatos, contribuindo para resultados imprecisos. A escassez de recursos, como analistas de dados qualificados, e vários fatores externos imprevisíveis, como mudanças na economia, complicam ainda mais a previsão precisa da procura. Esses desafios destacam a complexidade e a necessidade de abordagens estratégicas na previsão da procura, como salientado por Danielsson (2023).

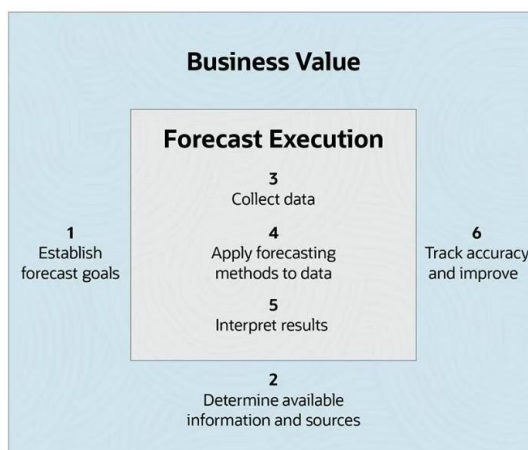


Figura 5: Etapas comuns para previsão da procura
Fonte: Demand Forecasting: Everything You Need to Know (Jenkins, 2023)

A previsão da procura não segue uma abordagem única universalmente aceita, uma vez que diferentes contextos requerem diferentes processos. Em alguns casos, a limpeza e validação de dados são passos críticos, enquanto em outros, os responsáveis pela previsão podem obter dados de alta qualidade de outras áreas da empresa, prontos para uso imediato. Contudo, existem características consistentes em esforços bem-sucedidos de previsão contínua, e seis passos essenciais aplicam-se a quase toda equipa de previsão de procura, como pode ser observado na Figura 5 (Jenkins, 2023).

O primeiro passo é identificar claramente os objetivos da previsão, fornecendo orientação para decisões subsequentes. Definir o tipo de previsão, o horizonte temporal e outros parâmetros ajuda a moldar o design do processo de previsão. O segundo passo consiste em determinar quais informações estão disponíveis e podem ser obtidas, sendo crucial identificar fontes de dados, sejam internas, externas, qualitativas ou quantitativas (Jenkins, 2023). Em seguida, é necessário estabelecer e executar um plano de coleta de dados, que pode envolver pesquisas, grupos focais e coordenação com outros departamentos. Após a coleta de dados, é importante aplicar métodos apropriados de previsão, sejam qualitativos ou quantitativos, dependendo dos objetivos estabelecidos (Jenkins, 2023).

Na interpretação dos resultados, é importante relatar não apenas os resultados quantitativos, mas também o que se acredita que ocorrerá, a precisão estimada, a margem de erro, pressupostos críticos e informações ausentes (Jenkins, 2023). Por fim, é essencial acompanhar os resultados para identificar possíveis falhas e oportunidades de melhoria. Manter registos das previsões, avaliar a precisão ao longo do tempo e buscar oportunidades de aprimoramento são passos fundamentais para aperfeiçoar os processos de previsão da procura (Jenkins, 2023).

2.3. Métodos Tradicionais de Previsão da Procura

Segundo Stoll (2022), os métodos tradicionais de previsão da procura são fundamentais para a análise de padrões temporais e para a realização de projeções. Estas técnicas têm sido aplicadas em setores como finanças, economia, meteorologia e gestão de stocks, fornecendo percepções valiosas para a tomada de decisões estratégicas. Karim et al. (2013) realizaram um estudo sobre

previsão de energia elétrica para equilibrar a oferta e a procura numa empresa e descobriram que, devido à natureza estacionária da série temporal e à pequena variação no consumo de energia ao longo do tempo, os métodos de alisamento exponencial e média móvel aplicados foram eficazes, não sendo necessário, neste caso, implementar métodos mais complexos.

Ao utilizar informações históricas, os métodos estatísticos de previsão da procura procuram captar tendências, sazonalidades e comportamentos cíclicos, fornecendo uma base sólida para estimativas a longo prazo. Os métodos tradicionais incluem abordagens simples, como a média móvel simples e a média móvel ponderada, que são eficazes para suavizar flutuações de curto prazo e destacar tendências subjacentes. Além disso, técnicas mais avançadas, como o alisamento exponencial e os modelos ARIMA e SARIMA, oferecem maior complexidade, permitindo a modelação de séries temporais com padrões mais intrincados (Stoll, 2020).

Os métodos utilizados neste trabalho foram a média móvel simples, média móvel exponencial, ARIMA, SARIMA, *Holt-Winters* e ETS.

2.3.1. Média Móvel Simples

Johnston (1999) afirma que a utilização da Média Móvel Simples (MMS) na previsão da procura no setor do retalho pode enfrentar desafios significativos devido à natureza dinâmica e imprevisível deste mercado. A MMS é uma técnica estatística que calcula a média de um conjunto de dados ao longo de um período fixo, sendo amplamente utilizada para suavizar flutuações e identificar tendências em séries temporais. A fórmula matemática para calcular a MMS é dada por:

$$MMS_t = \frac{1}{n} \sum_{i=0}^{n-1} x_{t-i} \quad (1)$$

Onde MMS_t é a média móvel simples no período t , x_{t-i} são os valores das observações passadas e n é o número de períodos na janela da média móvel (Box, 2015).

No entanto, embora seja uma ferramenta útil para análises básicas, a sua aplicação em mercados voláteis, como o retalho, pode ser limitada. Hansun (2013) observa que, no contexto do retalho, a MMS pode não ser suficientemente ágil para captar as rápidas mudanças nas preferências dos consumidores ou responder de forma eficaz à introdução constante de novos produtos e promoções. A volatilidade das vendas, influenciada por fatores sazonais, tendências de mercado, e eventos específicos, pode reduzir a precisão das previsões feitas com base na MMS. Esta técnica, ao atribuir o mesmo peso a todos os valores no período de análise, pode subestimar as variações recentes e, conseqüentemente, produzir previsões menos precisas em mercados onde as condições mudam rapidamente.

Nau (2014) destaca que, por outro lado, a simplicidade da Média Móvel Simples é uma das suas maiores vantagens. É fácil de implementar e interpretar, o que a torna uma escolha comum para previsões de curto prazo em ambientes onde as condições de mercado são relativamente estáveis. Um estudo realizado por Hanggara (2021) explorou a eficácia da MMS na previsão da procura por automóveis na Indonésia, com o objetivo de evitar o excesso de produção. O estudo mostrou que, embora a MMS possa fornecer resultados satisfatórios em determinados contextos, a sua eficácia é limitada quando comparada com modelos mais sofisticados, como a regressão linear ou métodos

de séries temporais mais avançados, que conseguem captar melhor a complexidade das variáveis envolvidas.

2.3.2. Média Móvel Exponencial

Caixeta (2023) defende que a Média Móvel Exponencial (MME) é uma técnica estatística que suaviza séries temporais e prevê tendências futuras, atribuindo maior peso aos dados mais recentes, ao contrário da MMS, que trata todos os dados igualmente. Utilizando um fator de suavização, α , a MME ajusta a média de forma a que as observações recentes influenciem mais fortemente o resultado, proporcionando uma análise mais sensível às mudanças rápidas. Em um estudo, dos Reis et al. (2015) realizaram a previsão da procura em um mercado de retalho de baterias automotivas, e concluíram que dentre os modelos de média móvel simples, média móvel exponencial e regressão linear, a MME apresentou menor erro absoluto, demonstrando ser o melhor modelo para previsão das baterias. A fórmula matemática para calcular a MME é dada por:

$$MME_t = \alpha x_t + (1 - \alpha)MME_{t-1} \quad (2)$$

Onde MME_t é a média móvel exponencial no período t , x_t é o valor da observação no período t , MME_{t-1} é a média móvel exponencial no período anterior e α é o fator de suavização ($0 < \alpha < 1$)

Embora a MME seja eficaz para captar alterações rápidas nas tendências e seja útil em mercados dinâmicos e voláteis, tem suas limitações. O fator α deve ser cuidadosamente escolhido para evitar previsões excessivamente reativas ou lentas em relação às mudanças. Além disso, em ambientes com alta volatilidade e padrões sazonais complexos, a MME pode ser menos eficaz, podendo exigir a utilização de métodos adicionais para otimizar as previsões (dos Reis, 2015).

2.3.3. ARIMA

Mondal (2014) cita que o modelo ARIMA (AutoRegressive Integrated Moving Average) é uma técnica estatística amplamente utilizada para modelar e prever séries temporais. O ARIMA combina três componentes principais: autoregressivo (AR), integrado (I) e média móvel (MA). Cada um desses componentes desempenha um papel específico na modelagem e previsão de séries temporais (Hyndman, 2018):

1. *Componente Autoregressivo (AR)*: Este componente captura a dependência linear entre as observações passadas e o valor atual da série temporal. A parte AR do modelo é representada por um parâmetro p , que indica o número de defasagens ou períodos passados incluídos no modelo. Em termos matemáticos, o modelo AR de ordem p é expresso como (Hyndman, 2018):

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \epsilon_t \quad (3)$$

onde ϕ_1 são os parâmetros autoregressivos e ϵ_t é o erro branco.

2. *Componente Integrado (I)*: O componente integrado realiza a diferenciação da série temporal para torná-la estacionária, ou seja, para remover tendências e sazonalidades. A diferenciação é representada por um parâmetro d , que indica o número de vezes que a série deve ser diferenciada para alcançar a estacionariedade. A diferenciação é calculada como (Hyndman, 2018):

$$\Delta^d X_t = (1 - B)^d X_t \quad (4)$$

Onde B é o operador de defasagem e Δ é o operador de diferenciação.

3. *Componente de Média Móvel (MA)*: Este componente ajusta a média dos erros de previsão passados para melhorar a precisão das previsões atuais. O modelo MA é representado por um parâmetro q , que indica o número de períodos passados de erros incluídos no modelo. Em termos matemáticos, o modelo MA de ordem q é expresso como (Hyndman, 2018):

$$X_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q} \quad (5)$$

onde θ_i são os parâmetros de média móvel e ϵ_t é o erro branco.

A combinação destes três componentes resulta no modelo ARIMA (p, d, q), que pode ser ajustado para capturar diferentes características das séries temporais. O ajuste dos parâmetros p , d e q é essencial para a eficácia do modelo ARIMA. O processo de ajuste envolve a identificação e estimação dos valores ótimos destes parâmetros para minimizar o erro de previsão (Hyndman, 2018). No estudo de Ediger e Akar (2007), foi utilizado o modelo ARIMA para prever a demanda primária de energia na Turquia de 2005 a 2020. O ARIMA mostrou-se mais confiável do que a soma das previsões individuais das fontes de energia, comprovando ser um método eficiente.

Khan (2020) explica que para aplicar o ARIMA, é necessário ajustar os parâmetros adequados para cada componente: o número de termos autoregressivos (p), o número de diferenciações (d) e o número de termos de média móvel (q). Embora o ARIMA seja eficaz para séries temporais complexas e não estacionárias, pode exigir pré-processamento adicional para garantir a estacionariedade da série. Além disso, o ARIMA pode ser menos eficaz em séries com padrões sazonais fortes, podendo necessitar de variantes como o SARIMA (*Seasonal ARIMA*) para capturar essas características adicionais.

2.3.4. SARIMA

O modelo SARIMA (*Seasonal AutoRegressive Integrated Moving Average*), segundo Walter (2013), é uma extensão do modelo ARIMA. Enquanto o ARIMA é eficaz para séries temporais não estacionárias e não sazonais, o SARIMA incorpora elementos adicionais para capturar as variações sazonais que ocorrem em intervalos regulares, como dias, meses ou trimestres. O SARIMA combina os componentes do ARIMA com elementos adicionais que capturam a sazonalidade na série temporal. O modelo SARIMA é representado pela notação $ARIMA(p,d,q)(P,D,Q)_s$, onde P é o número de termos autorregressivos sazonais, D é o número de diferenciações sazonais, Q é o número de termos de média móvel sazonais e S é o período sazonal (Wei, 2006).

Ainda segundo Walter (2013), o SARIMA combina os componentes do ARIMA com componentes sazonais específicos. Estes componentes sazonais são divididos em autoregressivos sazonais (SAR), integrais sazonais (SI) e médias móveis sazonais (SMA). O componente SAR captura a relação entre os valores passados da série com um período sazonal específico, o componente SI trata da diferenciação sazonal para estabilizar a série temporal, e o componente SMA ajusta a média dos erros sazonais passados (Wei, 2006). O modelo SARIMA é expresso pela combinação dos componentes sazonais e não sazonais e pode ser calculado utilizando a seguinte fórmula geral:

$$\Phi(B^S)\Delta_S^D X_t = \Theta(B^S) \epsilon_t \quad (6)$$

Onde $\Phi(B^S)$ é o polinómio autorregressivo sazonal, Δ_S^D é o operador de diferenciação sazonal, $\Theta(B^S)$ é o polinómio de média móvel sazonal e ϵ_t são os erros do modelo (Wei, 2006).

Divisekara et al. (2020) aplicaram o modelo SARIMA para prever os preços das lentilhas vermelhas no Canadá com base em dados semanais, demonstrando que o modelo foi eficaz na captura da sazonalidade e volatilidade dos preços. Segundo Wu et al. (2021), a principal vantagem do SARIMA reside na sua capacidade de modelar séries temporais com padrões sazonais complexos, o que aumenta a precisão das previsões em contextos onde a sazonalidade é um fator predominante. No entanto, Wu et al. (2021) também aponta que o ajuste do SARIMA pode ser mais desafiante devido ao elevado número de parâmetros e à necessidade de identificar corretamente o período sazonal. Além disso, em séries temporais com múltiplos padrões sazonais ou comportamentos não lineares, o SARIMA pode necessitar de ser combinado com outras técnicas para alcançar resultados mais precisos.

2.3.5. Holt-Winters

Chatfield (1978) cita que o modelo *Holt-Winters* é uma técnica de previsão de séries temporais que é particularmente eficaz para lidar com dados que exibem tendências e sazonalidades. Desenvolvido por Charles Holt e para o ajuste da sazonalidade, este modelo é uma extensão do modelo de suavização exponencial simples e é frequentemente utilizado em contextos onde as séries temporais apresentam padrões sazonais e tendências que variam ao longo do tempo. No estudo de Panda (2020), foi utilizado o modelo de suavização exponencial *Holt-Winters* para prever a propagação da COVID-19 na Índia e em cinco de seus estados e a avaliação do modelo mostrou que ele é eficaz para prever a disseminação da doença, auxiliando no planejamento estratégico.

A principal vantagem do modelo *Holt-Winters*, ainda segundo Chatfield (1978), é a sua capacidade de capturar tanto a tendência quanto a sazonalidade nas séries temporais, o que o torna útil para previsões em contextos onde essas características são evidentes. No entanto, a eficácia do modelo depende da escolha adequada dos parâmetros de suavização e da natureza dos dados. Em casos onde a série temporal exibe padrões não lineares complexos ou mudanças abruptas, pode ser necessário considerar modelos mais avançados ou técnicas complementares para melhorar a precisão das previsões.

Canela (2019) explica que o modelo é calculado a partir de três componentes principais: o nível, a tendência e a sazonalidade, ajustados por três parâmetros de suavização. O parâmetro α (alfa) ajusta o nível da série, β (beta) ajusta a tendência e γ (gama) controla a sazonalidade. A depender

da natureza da sazonalidade da série, o modelo pode ser aditivo, quando a variação sazonal é constante, ou multiplicativo, quando essa variação aumenta ou diminui proporcionalmente ao nível da série (Canela, 2019).

O cálculo do modelo é iterativo, ajustando o nível com base no valor observado, a tendência e o componente sazonal da última observação. A tendência é ajustada pela diferença entre o nível atual e o anterior, e a sazonalidade é recalculada com base na diferença entre o valor observado e o nível ajustado. Esses ajustes são feitos de maneira ponderada, considerando os parâmetros de suavização, o que permite que o modelo responda a mudanças recentes nos dados sem ser afetado por pequenas oscilações. As previsões futuras são geradas usando o nível e a tendência ajustados, somados ao componente sazonal correspondente ao período previsto, o que garante a capacidade do modelo de capturar variações sazonais e tendências ao longo do tempo (Canela, 2019).

2.3.6. ETS

O modelo ETS (Erro, Tendência, Sazonalidade), conforme descrito por Panigrahi (2021), é uma técnica de previsão de séries temporais conhecida pela sua flexibilidade em captar diversos padrões dos dados. O modelo ETS decompõe uma série temporal em três componentes principais: erro, tendência e sazonalidade, refletidos na sigla ETS. A principal vantagem deste modelo é a sua capacidade de ajustar-se a diferentes padrões de dados, tornando-o especialmente útil para séries que apresentam variações sazonais e tendências não lineares. Além disso, o modelo ETS é fácil de interpretar e ajustar, o que o torna uma ferramenta valiosa para análises e previsões. Contudo, a eficácia do modelo depende da escolha adequada dos componentes e da configuração dos parâmetros, sendo crucial especialmente para séries com comportamentos complexos ou mudanças abruptas (Panigrahi, 2021).

O cálculo do modelo ETS envolve a decomposição da série temporal em três componentes: erro, tendência e sazonalidade. Cada um destes componentes é modelado e ajustado para refletir as características da série temporal observada. O erro, que pode ser aditivo ou multiplicativo, representa a parte da série que não é explicada pelos componentes de tendência e sazonalidade. A tendência captura a direção geral do movimento da série ao longo do tempo e pode ser constante ou variável. A sazonalidade captura padrões que se repetem em intervalos regulares, podendo ser aditiva ou multiplicativa (Martinello, 2021).

Ainda segundo Martinello (2021), para calcular o modelo ETS, utiliza-se a suavização exponencial para ajustar cada componente. O nível é atualizado com base na observação atual e nos componentes anteriores. A tendência é ajustada considerando as diferenças entre o nível atual e o nível anterior. A sazonalidade é ajustada com base na diferença entre a observação atual e o nível ajustado, tendo em conta o padrão sazonal. Os parâmetros de suavização são ajustados para minimizar o erro de previsão, e as previsões futuras são calculadas combinando o nível, a tendência e a sazonalidade, conforme o modelo escolhido. Esta abordagem garante que o modelo ETS possa adaptar-se de forma eficaz a diferentes padrões de dados e proporcionar previsões precisas.

2.4. Métodos Avançados de Previsão com *Machine Learning*

No campo da Inteligência Artificial, a aprendizagem de máquina (*Machine Learning*) destaca-se pela sua eficácia na previsão de resultados práticos, através do treino de modelos com dados históricos no âmbito do aprendizado supervisionado. Esta abordagem permite a avaliação da precisão dos modelos, complementando os métodos estatísticos tradicionais. Além disso, a aprendizagem de máquina não se limita a problemas lineares, sendo particularmente útil na análise de séries temporais com padrões não lineares, como ciclos e transições entre períodos de estabilidade e volatilidade (Sharda, 1994).

No contexto dos algoritmos de *Machine Learning*, é importante considerar a sua hierarquia e categorização. Os algoritmos de regressão são utilizados para prever valores contínuos, como a estimativa do preço de um imóvel com base nas suas características (Bishop, 2006). Algoritmos de classificação são empregados para categorizar dados em classes distintas, por exemplo, identificar e-mails como spam ou não spam. Por outro lado, o reinforcement learning (aprendizagem por reforço) é uma abordagem em que um agente aprende a tomar decisões interativas com o ambiente, visando maximizar uma recompensa cumulativa ao longo do tempo (Sutton & Barto, 2018). Cada uma dessas categorias de algoritmos é aplicada com base na natureza específica do problema e nas características dos dados disponíveis, permitindo uma gama diversificada de soluções para desafios distintos.

A aprendizagem de máquina tornou-se essencial na análise de grandes volumes de dados gerados diariamente na era digital. Holzinger et al. (2018) destacam que a quantidade de informação disponível, como transações financeiras e interações em redes sociais, ultrapassa a capacidade humana de processamento. Nesse cenário, a aprendizagem de máquina surge como uma ferramenta poderosa para a extração de conhecimento, identificação de padrões complexos e realização de previsões precisas em diversas áreas. De acordo com Holzinger et al. (2018), a União Europeia tem direcionado crescentes investimentos para a investigação em Inteligência Artificial, reconhecendo a importância desta área para o avanço tecnológico e económico. Contudo, apesar dos avanços significativos, persistem desafios, como a falta de transparência em modelos de alto desempenho, frequentemente descritos como "caixas-pretas", o que dificulta a sua interpretação e aplicação prática. Assim, apesar dos investimentos na área de Inteligência Artificial e aprendizagem de máquina, é imperativo abordar essas questões para garantir que essas tecnologias sejam empregadas de maneira ética e responsável.

Os métodos de aprendizagem de máquina utilizados neste trabalho foram a Regressão Linear Simples, Regressão Polinomial, *Decision Tree*, *XGBoost*, e *Support Vector Machines* (SVM).

2.4.1. Regressão Linear Simples

Segundo Galton (1886), a regressão linear simples é uma técnica estatística fundamental que ele introduziu para modelar a relação entre duas variáveis quantitativas através de uma relação linear. Galton, um dos pioneiros da estatística, desenvolveu este conceito ao estudar a herança de características físicas, como a altura, e observou que os filhos de pais muito altos ou baixos tendem

a ter alturas mais próximas da média da população do que os seus pais. Este fenómeno, que ele chamou de "regressão à média", levou à formulação da equação de regressão linear simples. A técnica procura ajustar uma linha reta que melhor representa a relação entre a variável dependente Y e a variável independente X . A equação da regressão linear simples é dada pela fórmula X:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (7)$$

onde β_0 é o intercepto da linha com o eixo Y , β_1 é o coeficiente angular que representa a inclinação da linha, e ε é o termo de erro, que captura a variação não explicada pela relação linear.

No estudo de Bianco et al. (2009), modelos de regressão linear simples foram utilizados para prever o consumo anual de eletricidade na Itália, e ele defende que essa abordagem é amplamente utilizada devido à sua simplicidade e interpretabilidade. A regressão linear simples permite prever o valor da variável dependente com base nos valores da variável independente, e oferece uma visão clara da força e direção da relação entre ambas. A análise da regressão linear simples inclui a estimativa dos parâmetros β_0 e β_1 através do método dos mínimos quadrados, que minimiza a soma dos quadrados das diferenças entre os valores observados e os valores preditos pela equação. Este método é essencial em diversas áreas de pesquisa e prática, proporcionando uma base sólida para a modelagem e a análise de dados.

2.4.2. Regressão Polinomial

Segundo Bowley (1920), a regressão polinomial é uma técnica estatística que expande o conceito da regressão linear para modelar a relação entre duas variáveis quantitativas quando a relação entre elas é não linear. A abordagem original de regressão linear simples, desenvolvida por Galton (1886), ajusta uma linha reta à relação entre as variáveis, mas em muitos casos, essa abordagem não é suficiente para capturar a complexidade das relações entre as variáveis. Para superar essa limitação, Bowley introduziu a regressão polinomial, que permite ajustar uma curva polinomial de grau n à relação entre a variável dependente Y e a variável independente X . A equação geral de um modelo de regressão polinomial de grau d é dada pela fórmula:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_d X^d + \varepsilon \quad (8)$$

onde os coeficientes $\beta_0, \beta_1, \beta_2 \dots \beta_d$ são estimados através do método dos mínimos quadrados, e ε é o termo de erro que captura as variações não explicadas pelo modelo.

No estudo de Wi et al. (2011), foi utilizada a regressão polinomial para prever a carga elétrica em feriados, e eles concluíram que o método é útil quando a relação entre as variáveis não pode ser adequadamente descrita por uma linha reta. A regressão polinomial permite ajustar a curva ao comportamento dos dados, oferecendo uma maior flexibilidade na modelagem de tendências e padrões complexos. Contudo, a adição de termos polinomiais pode também aumentar o risco de sobreajuste (overfitting), especialmente se o grau do polinómio for muito elevado. O sobreajuste ocorre quando o modelo se ajusta demasiado bem aos dados de treino, capturando ruídos e variações aleatórias, o que pode comprometer a sua capacidade de generalização para novos

dados. Portanto, é crucial escolher o grau do polinômio de forma cuidadosa e validar o modelo com dados independentes para assegurar a robustez e a validade das previsões (Wibisana, 2024).

2.4.3. Decision Tree

Segundo Breiman et al. (1984), a árvore de decisão (*Decision Tree*) é uma técnica amplamente utilizada em estatística e aprendizagem automática para modelar e visualizar decisões e as suas possíveis consequências. Este método constrói um modelo em forma de árvore que prevê o valor de uma variável alvo com base em várias variáveis de entrada. Breiman e seus colegas introduziram e detalharam a técnica de árvores de decisão em seu trabalho seminal, estabelecendo a base para muitos dos métodos modernos. A estrutura da árvore, conforme visualizado na Figura 6, é composta por nós, que representam testes ou decisões, e ramos, que indicam os resultados desses testes. Cada nó interno da árvore corresponde a um teste numa característica dos dados, enquanto cada folha representa uma previsão ou decisão final (Breiman et al, 1984).

O processo começa na raiz da árvore, onde o algoritmo realiza um teste sobre o atributo da variável de entrada e, em seguida, segue o ramo correspondente ao resultado do teste para o próximo nó. Esse processo é repetido de forma recursiva, com o algoritmo a verificar os valores dos atributos e a mover-se através dos nós internos até chegar a um nó folha. Liu (2017) explica que o critério de divisão em cada nó é baseado em métricas como o índice de Gini, a entropia ou a redução da variância, dependendo se o problema é de classificação ou regressão. Uma das principais contribuições do trabalho de Breiman et al. (1984) foi a formalização da técnica e a sua aplicabilidade a diversos tipos de problemas, tornando-a uma ferramenta poderosa e versátil. Além disso, a árvore de decisão oferece a vantagem de ser facilmente interpretável e de lidar com relações não lineares. No entanto, um desafio associado a esta técnica é o sobreajuste (*overfitting*), especialmente quando se lida com muitos atributos e amostras pequenas. (Liu, 2017).

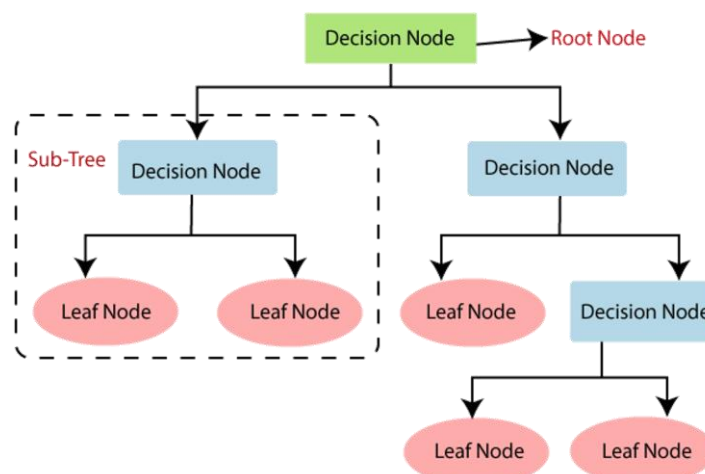


Figura 6: Estrutura da Árvore de Decisão
Fonte: AlmaBetterBytes (2024)

2.4.4. XGBoost

O *XGBoost*, abreviação de "*Extreme Gradient Boosting*", é uma poderosa técnica de aprendizagem automática baseada em árvores de decisão, amplamente utilizada para problemas de classificação e regressão. Introduzido por Chen e Guestrin (2016), o *XGBoost* é uma implementação sofisticada do método de *boosting*, que cria um conjunto de modelos de árvore de decisão de forma sequencial. Cada nova árvore é treinada para corrigir erros das árvores anteriores, proporcionando alta eficiência computacional e robustez, especialmente em grandes volumes de dados e múltiplas variáveis. A principal vantagem do *XGBoost* em relação a outras técnicas de *boosting* é a sua eficiência computacional, robustez e capacidade de lidar com grandes volumes de dados e múltiplas variáveis. A estrutura do modelo *XGBoost* pode ser visualizada na Figura 7.

Chen e Guestrin (2016) descrevem que o *XGBoost* funciona através de uma abordagem de aprendizagem aditiva, onde as árvores de decisão são construídas uma após a outra. A cada iteração, o modelo tenta minimizar uma função de perda, como o erro quadrático médio ou a entropia, ajustando-se aos erros residuais das árvores anteriores. Uma das características distintivas do *XGBoost* é a utilização do regularizador L1 e L2, que ajuda a evitar o sobreajuste (overfitting) ao controlar a complexidade do modelo, penalizando árvores excessivamente complexas. Além disso, o *XGBoost* suporta a paralelização de operações durante o processo de treino, o que o torna extremamente rápido, mesmo em conjuntos de dados muito grandes (Rahman, 2022).

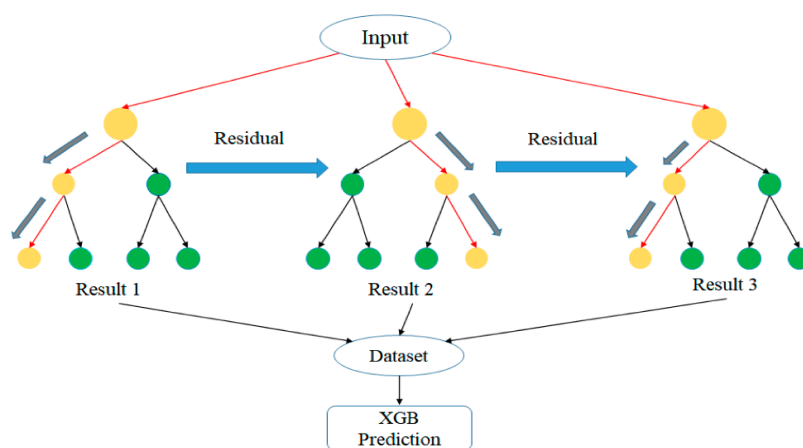


Figura 7: Estrutura do XGBoost
Fonte: Wang et al. (2022)

Outra característica importante do *XGBoost* é a capacidade de lidar com dados esparsos e variáveis com alta cardinalidade de forma eficiente. O algoritmo implementa técnicas de tratamento de valores ausentes e otimização de memória, tornando-o particularmente eficaz em cenários do mundo real, onde os dados nem sempre são perfeitos. (Rahman, 2022)

2.4.5. Support Vector Machines

A Máquina de Vetores de Suporte (SVM) é explicada por Huang (2005) como sendo uma técnica avançada de aprendizagem automática usada principalmente para classificação, embora também seja aplicada em regressão. O SVM funciona identificando o hiperplano que melhor separa os dados em diferentes classes, maximizando a margem entre os pontos mais próximos de cada classe, chamados vetores de suporte. Este processo reduz o risco de sobreajuste e cria modelos generalizáveis, tornando o SVM eficaz para problemas de alta dimensionalidade (Kurani, 2023). A ilustração do modelo SVM está representada na Figura 8.

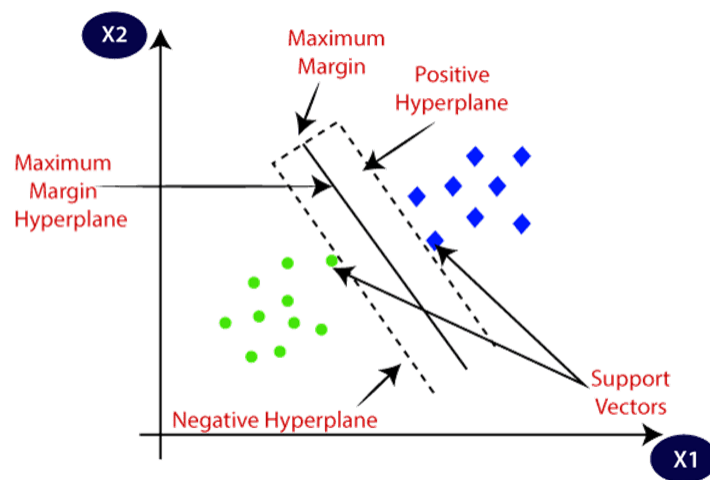


Figura 8: Estrutura do modelo SVM
Fonte: Soylev, 2023

Uma das principais vantagens do SVM é a sua capacidade de tratar problemas de classificação não linear através do "truque do kernel", que transforma os dados em um espaço dimensional superior para permitir uma separação linear das classes. Os kernels comuns, como o linear, polinomial e a função de base radial (RBF), são adaptáveis a diferentes tipos de dados. No entanto, o SVM pode ser computacionalmente intensivo e seu desempenho depende da escolha correta dos parâmetros do kernel e da regulação dos parâmetros de penalização (Kurani, 2023).

2.5. Trabalhos Relacionados

O estudo de Singh et al. (2019) destaca que, inicialmente, modelos baseados em árvores, como o *XGBoost*, mostraram um desempenho superior comparado aos modelos de redes neurais, como LSTM e MLP, na previsão da procura por novos itens de moda. Contudo, à medida que o volume de dados cresceu e os parâmetros das redes neurais foram ajustados, esses modelos neurais passaram a apresentar um desempenho significativamente melhor. Isso sugere que, embora técnicas tradicionais como o *XGBoost* sejam eficazes, as redes neurais têm um potencial crescente à medida que são refinadas.

Complementando esta visão, Vashishtha (2020) analisou a eficácia dos modelos de previsão baseline, como *XGBoost* e Redes Neurais Artificiais, comparando-os com modelos que consideram a idade dos produtos. Os resultados indicam que os modelos que incorporam a idade dos produtos e a sazonalidade são mais precisos, pois capturam melhor a semelhança entre itens e a sazonalidade, resultando em previsões superiores para categorias específicas, como roupas infantis e vestidos.

Por outro lado, Hamza (2021) abordou um desafio diferente ao investigar o uso de Redes Bayesianas para prever vendas durante promoções. Aqui, a combinação de fatores qualitativos e quantitativos das BNs mostrou-se mais eficaz do que os modelos quantitativos tradicionais, que muitas vezes falham em capturar as rápidas variações impulsionadas por promoções. As BNs ofereceram previsões mais precisas de vendas semanais e promocionais, destacando a necessidade de modelos que integrem diferentes tipos de dados.

Swami (2020) focou na previsão de vendas futuras para cada produto e loja, utilizando técnicas *XGBoost* e LSTM. A avaliação revelou que o *XGBoost* superou o LSTM devido à sua capacidade de lidar melhor com dados esparsos, reforçando a importância de escolher a técnica adequada com base na natureza dos dados disponíveis.

Finalmente, o estudo de Huber (2020) investigou a previsão de procura para uma cadeia de padarias, especialmente em dias especiais. Comparando redes neurais artificiais e árvores de decisão com boosting, os resultados mostraram que abordagens de classificação são mais eficazes do que as de regressão para previsões em grande escala. Este estudo sublinha a vantagem de métodos de *Machine Learning* na previsão de procura no retalho.

2.6. Métricas de Avaliação

Para Chicco (2021), as métricas de avaliação são fundamentais para entender o desempenho dos modelos de regressão e incluem várias abordagens para quantificar a precisão das previsões. Entre as principais, destacam-se o R^2 , MAE, MSE, RMSE e MAPE, cada uma com suas características e fórmulas específicas. O R^2 (coeficiente de determinação) mede a proporção da variabilidade total dos dados que é explicada pelo modelo. É calculado pela fórmula:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (9)$$

Onde SS_{res} é a soma dos quadrados dos resíduos e SS_{tot} é a soma dos quadrados totais. Um valor de R^2 próximo de 1 indica que o modelo explica bem a variabilidade dos dados, enquanto valores próximos de 0 sugerem que o modelo não está explicando muito bem a variabilidade.

Neste trabalho, foram utilizadas quatro métricas de erro: Erro Médio Absoluto (MAE), Erro Quadrático Médio (MSE), Raiz do Erro Quadrático Médio (RMSE) e Erro Percentual Absoluto Médio (MAPE). O MAE mede o erro médio absoluto entre as previsões e os valores reais, fornecendo uma visão geral da precisão sem considerar erros quadráticos. O MSE calcula a média dos quadrados das diferenças, penalizando mais severamente erros grandes. O RMSE, raiz quadrada do MSE, é sensível a grandes erros, sendo útil para penalizar discrepâncias significativas. O MAPE expressa o erro como uma percentagem dos valores reais. Essas métricas oferecem diferentes perspectivas sobre o

desempenho do modelo, auxiliando na identificação de pontos fortes e fracos e na melhoria da precisão preditiva (Chicco, 2021). As fórmulas estão disponíveis abaixo.

$$MAE = \frac{1}{n} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (10)$$

$$RMSE = \sqrt{\sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{N}} \quad (11)$$

$$MAPE = \frac{100}{n} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (12)$$

$$MSE = \frac{1}{n} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (13)$$

Onde y_i é o valor real, \hat{y}_i é o valor previsto e n é o número total de observações.

3. MÉTODOS E APLICAÇÃO

Nesta dissertação, a metodologia adotada para conduzir o projeto segue o modelo *Cross Industry Standard Process for Data Mining* (CRISP-DM). Este modelo estruturado é amplamente reconhecido pela sua eficácia na gestão de projetos de mineração de dados e é composto por seis etapas principais como evidenciado na Figura 9: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation* e *Deployment*. A escolha do CRISP-DM é particularmente adequada para esta pesquisa, uma vez que se adapta bem às complexidades e exigências de precisão do projeto (Vásquez, 2024).

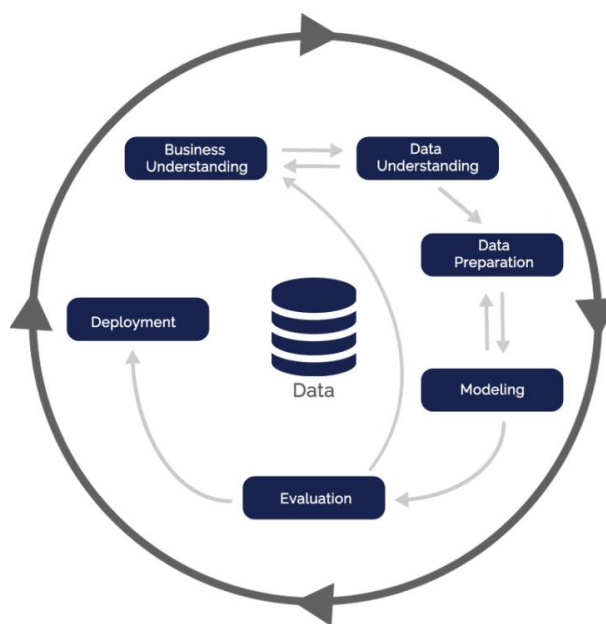


Figura 9: Metodologia CRISP-DM

Fonte: Anirudh, 2020

A fase de *Business Understanding* envolveu a identificação das necessidades específicas da investigação e a definição dos objetivos do estudo. Nesta fase, foi fundamental compreender o contexto do problema e alinhar os objetivos da pesquisa com as expectativas do negócio. A etapa de *Data Understanding* consistiu na análise dos dados disponíveis, incluindo a sua origem, qualidade e relevância para o projeto. Esta etapa é crucial para garantir que os dados utilizados sejam adequados e capazes de fornecer compreensões significativas.

Na fase de *Data Preparation*, foram realizadas várias tarefas de limpeza e transformação dos dados para garantir que estivessem prontos para a modelagem. Esta etapa envolveu a remoção de dados inconsistentes, a normalização e a transformação dos dados em formatos apropriados. A fase de *Modeling* utilizou diferentes modelos tradicionais (estatísticos) e algoritmos de aprendizagem automática para construir e treinar modelos que pudessem prever a procura com precisão. Durante esta fase, foram testados diversos modelos para identificar quais ofereciam o melhor desempenho em relação aos objetivos estabelecidos.

A *Evaluation* focou na análise dos resultados obtidos com os modelos e na seleção do que melhor se ajustava aos objetivos da pesquisa. Para essa análise, foram utilizadas métricas de erro, como o Erro Médio Absoluto (MAE), o Erro Quadrático Médio (MSE) e a Raiz do Erro Quadrático Médio (RMSE), bem como o Coeficiente de Determinação (R^2). Essas métricas permitiram uma avaliação detalhada da eficácia dos modelos de regressão, ajudando a identificar o modelo que melhor equilibrava precisão e desempenho. É importante destacar que, para este estudo, a fase de *Deployment* não foi realizada, uma vez que o principal objetivo foi o desenvolvimento e a avaliação dos modelos para gerar percepções para o negócio.

Para a execução deste projeto, foi utilizada a plataforma *Visual Studio Code*, conhecida por ser um ambiente de desenvolvimento integrado (IDE) flexível e altamente personalizável. A linguagem *Python* foi escolhida devido à sua robustez e às suas bibliotecas avançadas para análise de dados e aprendizagem automática. O *Jupyter Notebook*, utilizado em conjunto com o *Python*, permitiu a criação de um ambiente interativo e intuitivo para análise e visualização de dados, facilitando a experimentação e a documentação dos processos. Esse conjunto de tecnologias proporcionou uma base sólida para a realização das diversas etapas do CRISP-DM, permitindo a manipulação eficaz dos dados e a aplicação de algoritmos avançados de aprendizagem automática.

Este enfoque metodológico e tecnológico garante uma abordagem sistemática e detalhada para a previsão da procura, assegurando a validade e a precisão dos resultados obtidos e contribuindo para o sucesso da investigação.

3.1. Business Understanding

O conjunto de dados analisado pertence à Walmart, uma proeminente empresa norte-americana de retalho com uma vasta rede de lojas. Abrange um período de cinco anos, de 29 de janeiro de 2011 a 24 de abril de 2016, e inclui informações detalhadas sobre as vendas unitárias de 3.049 produtos distribuídos em três categorias principais: *Hobbies*, Alimentos e Produtos Domésticos. Estes produtos estão organizados em 7 departamentos e são vendidos em 10 lojas situadas em três estados dos EUA: Califórnia (CA), Texas (TX) e Wisconsin (WI).

A Walmart disponibilizou estes dados com o intuito de melhorar a previsão das vendas nas suas lojas e otimizar as estratégias de gestão de inventário e planeamento promocional. Com base nestas informações, é possível desenvolver modelos de previsão que ajudam a identificar padrões de consumo e a antecipar a procura de forma mais precisa. Este processo inclui a construção de modelos globais, que capturam tendências gerais, e modelos específicos por loja e estado, que permitem uma análise mais detalhada das particularidades regionais.

Para além da precisão das previsões, a análise dos dados visa descobrir percepções valiosas para o negócio. Estas compreensões podem revelar padrões e tendências que não são imediatamente evidentes, proporcionando à Walmart a capacidade de adaptar as suas estratégias de vendas e promoções de maneira mais eficaz. A consideração de variáveis adicionais como preços dos produtos, eventos promocionais, dias da semana e feriados especiais ajuda a refinar ainda mais as previsões, ajustando-as às condições e comportamentos específicos que impactam as vendas. Em suma, a exploração e modelação dos dados visam não só melhorar a precisão das previsões de

vendas, mas também fornecer conhecimento relevante para a Walmart, permitindo a otimização das suas operações e estratégias no mercado de retalho.

3.2. Data Understanding

A análise inicial da base de dados indicou que não foram encontrados valores ausentes ou duplicados, e a formatação das variáveis categóricas segue um padrão consistente. No entanto, a análise identificou a presença de *outliers*, com algumas vendas a variar de quantidades muito baixas a valores superiores a 300 unidades. Apesar da existência desses *outliers*, optou-se por mantê-los no modelo para assegurar uma previsão mais precisa que reflita a variação real nas vendas.

Durante o período analisado, foram registradas vendas totais de 65.695.409 produtos. A distribuição das vendas revelou que a categoria de Alimentos representa 68,63% do total, Produtos Domésticos 22,04% e Hobbies 9,32%. A Califórnia destaca-se como o estado com a maior parte das vendas, representando 43,65% do total, seguida pelo Texas com 28,77% e Wisconsin com 27,58%. Dentro da Califórnia, as lojas CA_3 e CA_1 têm uma contribuição significativa, com 17,03% e 11,72% das vendas totais, respetivamente. No Texas, a loja TX_2 lidera com 38,2% das vendas, enquanto em Wisconsin, a loja WI_2 tem a maior participação com 36,1%.

Estes resultados, disponíveis na Figura 10 e Tabela 1, oferecem uma visão detalhada da distribuição das vendas e permitem identificar padrões e tendências cruciais. Esta análise é fundamental para o desenvolvimento de modelos de previsão mais eficazes, ajudando a otimizar a gestão de stock e a estratégia de vendas com base nas variações e concentrações de vendas por categoria e localização.

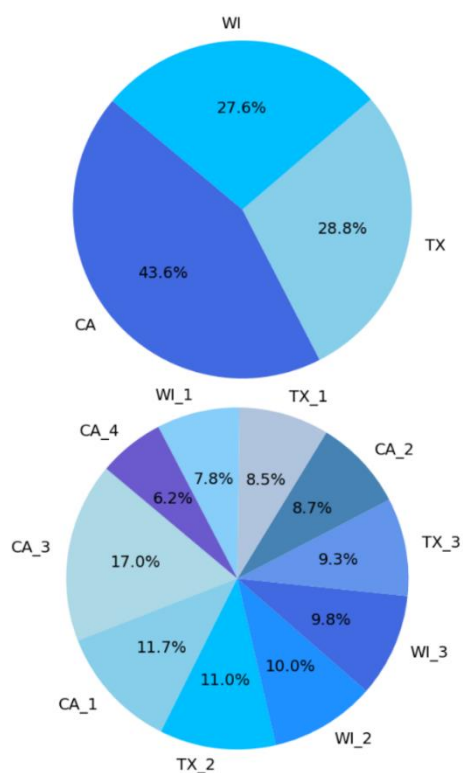


Figura 10: Gráfico da distribuição das vendas

Tabela 1: Distribuição total das vendas

Categoria	Total de Vendas	Porcentagem
<i>Foods</i>	45.089.939	68.63
<i>Household</i>	14.480.670	22.04
<i>Hobbies</i>	6.124.800	9.32
Estado	Total de Vendas	Porcentagem
CA	28.675.547	43.65
TX	18.899.006	28.77
WI	18.120.856	27.58
Loja	Total de Vendas	Porcentagem
CA_3	11.188.180	17.03
CA_1	7.698.216	11.72
TX_2	7.214.384	10.98
WI_2	6.544.012	9.96
WI_3	6.427.782	9.78
TX_3	6.089.330	9.27
CA_2	5.685.475	8.65
TX_1	5.595.292	8.52
WI_1	5.149.062	7.84
CA_4	4.103.676	6.25

As vendas no estado da Califórnia estão significativamente concentradas na loja CA_3, que representa 39% do volume total de vendas no estado, seguida pela loja CA_1, com 26,8%, CA_2 com 19,8% e CA_4 com 14,13%. Este padrão de concentração pode ser atribuído a fatores como a localização estratégica das lojas e a variabilidade na disponibilidade de produtos. No estado do Texas, a distribuição das vendas entre as lojas é relativamente equilibrada. A loja TX_2 apresenta o maior volume de vendas, correspondendo a 38,2% do total, seguida pela loja TX_3 com 32,3% e TX_1 com 29,6%. Por outro lado, no estado de Wisconsin, as vendas estão distribuídas de forma ainda mais uniforme entre as lojas. A loja WI_2 lidera com 36,1% das vendas, enquanto as lojas WI_3 e WI_1 contribuem com 35,5% e 28,4%, respectivamente. Os resultados da distribuição de vendas por loja estão disponíveis na Figura 11.

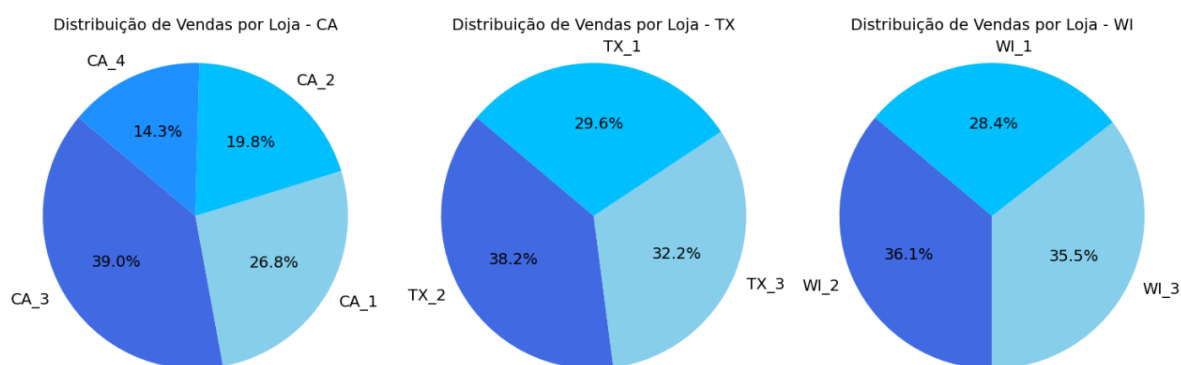


Figura 11: Distribuição das vendas por loja nos estados

A Figura 12 apresenta uma série temporal que representa as vendas da Walmart ao longo do período de 2011 a 2016. A série temporal analisada revela uma tendência crescente nas vendas ao longo do tempo, indicando um aumento geral no desempenho das vendas. Este crescimento pode ser atribuído a vários fatores, como estratégias de marketing mais eficazes, expansão da base de clientes, melhorias no produto ou alterações favoráveis no mercado. O aumento constante sugere que, de forma geral, o negócio está a expandir-se e a tornar-se mais bem-sucedido.

Adicionalmente, observa-se uma clara sazonalidade, com padrões repetitivos em intervalos fixos, sugerindo que as vendas flutuam de acordo com ciclos específicos do ano. Esta sazonalidade implica que as vendas aumentam ou diminuem em determinados períodos, refletindo a influência de fatores sazonais. Essa flutuação regular das vendas pode ser influenciada por fatores externos, como clima, datas comemorativas ou promoções sazonais, que afetam o comportamento dos consumidores de maneira previsível. O resíduo, por sua vez, representa a variação aleatória que resta após a remoção dos efeitos da tendência e da sazonalidade, capturando as flutuações inexplicáveis e o ruído nos dados. Assim, a análise destaca uma combinação de crescimento contínuo, padrões sazonais distintos e variações aleatórias nos dados de vendas.

A análise dos dados de vendas revelou várias observações importantes que oferecem uma visão detalhada dos padrões de compras ao longo do tempo, bem como das influências sazonais e dos eventos especiais. Em termos de distribuição mensal, constatou-se que as vendas estão relativamente bem distribuídas ao longo do ano, com picos significativos nos meses de julho, agosto, setembro e outubro. Este padrão pode ser atribuído a fatores sazonais, como o aumento

das compras durante as férias escolares e as promoções de regresso às aulas. A exclusão de 2016 foi uma medida necessária para garantir a precisão dos resultados, visto que a ausência de dados completos para o final desse ano poderia ter afetado a análise dos meses com maior volume de vendas, concentrando-se mais nos meses iniciais. Os resultados da análise do comportamento das vendas ao longo dos meses podem ser visualizados na Figura 13.

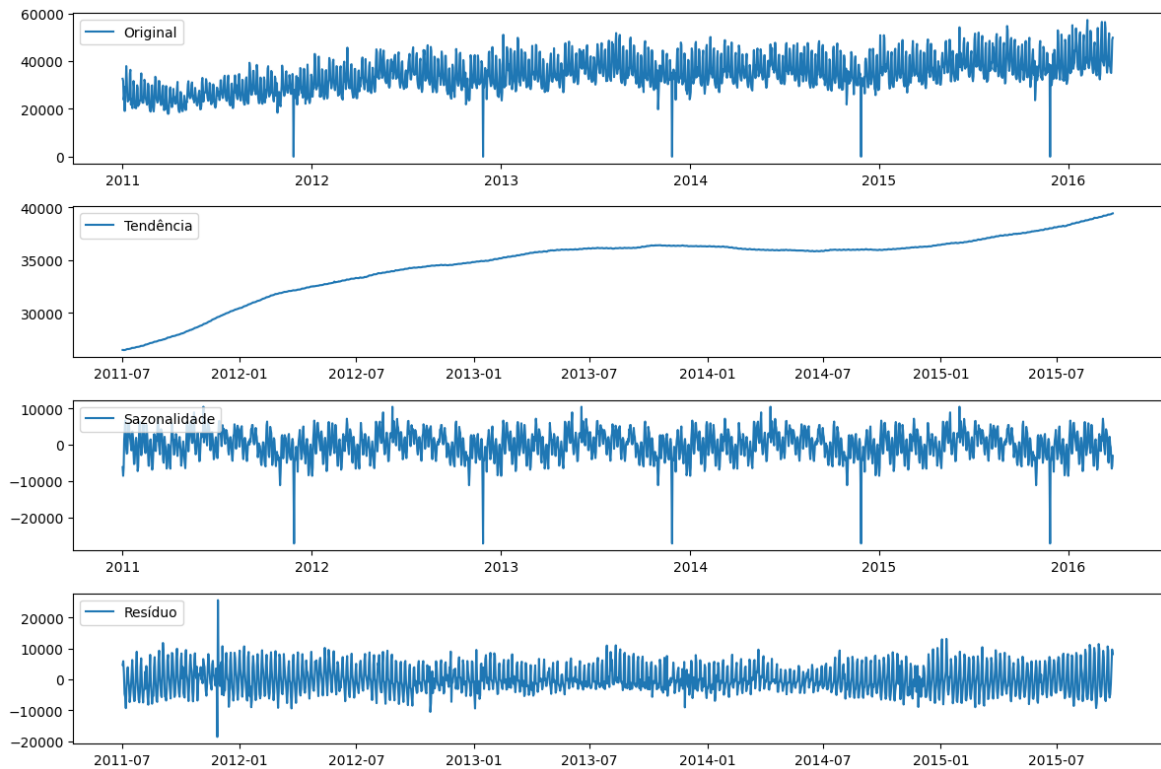


Figura 12: Gráfico das séries temporais

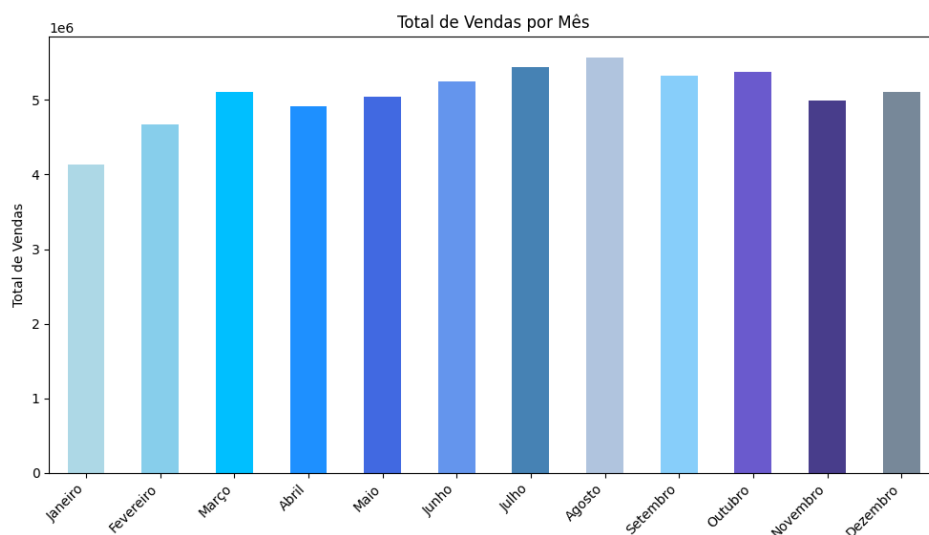


Figura 13: Distribuição mensal das vendas

Além disso, observou-se um crescimento significativo no volume de vendas ao longo dos anos, conforme demonstrado na Figura 14, com um aumento total de 55,83% de 2011 a 2015. Este

crescimento contínuo pode refletir uma expansão no mercado, melhorias nas estratégias de vendas ou alterações no comportamento dos consumidores.

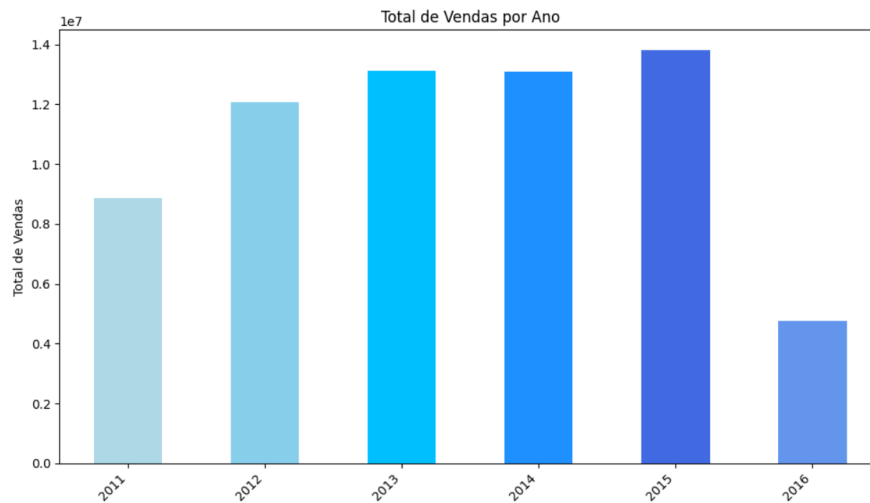


Figura 14: Distribuição anual das vendas

Analisando os dias da semana, verificou-se que os sábados e domingos são os dias com o maior número de vendas, o que pode estar relacionado com a maior disponibilidade dos consumidores durante o fim de semana. Em contraste, a quinta-feira e a quarta-feira apresentaram os menores volumes de vendas, sugerindo que os consumidores tendem a realizar compras mais frequentemente durante o fim de semana. Os resultados da análise do comportamento das vendas ao longo dos dias da semana podem ser visualizados na Figura 15.

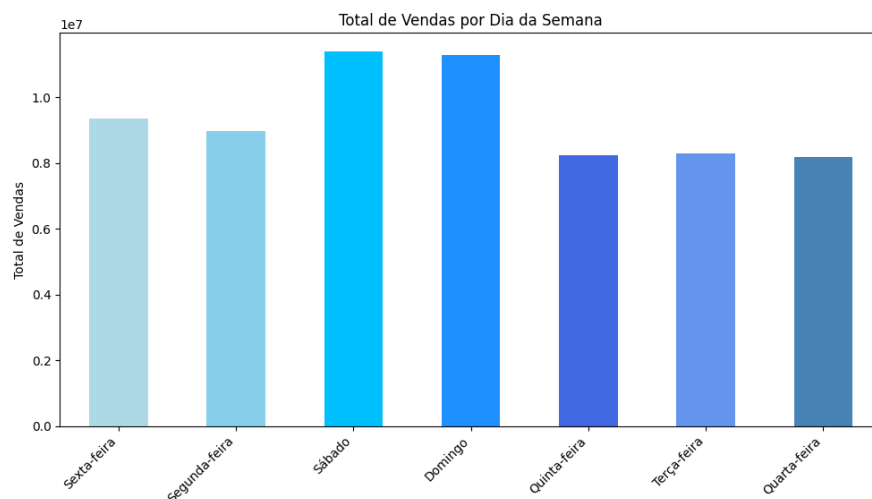


Figura 15: Distribuição semanal das vendas

A análise dos eventos também revelou que os eventos religiosos têm o maior impacto nas vendas, seguidos por eventos nacionais e culturais, enquanto os eventos desportivos têm menor influência, conforme representado na Figura 16. Entre os eventos religiosos mais significativos estão o *Chanukah* (festividade judaica que celebra a dedicação do Templo Sagrado), o *Eid al-Fitr* (celebração muçulmana que marca o fim do jejum do Ramadão), o *Purim End* (festividade judaica que comemora a salvação do povo judeu na antiga Pérsia) e o *Eid al-Adha* (festividade muçulmana

que celebra a disposição de Abraão em sacrificar o seu filho). Os feriados nacionais, como o *Independence Day* (comemoração da independência dos EUA), o *Labor Day* (dia dedicado aos trabalhadores nos EUA) e o *Columbus Day* (celebração da chegada de Cristóvão Colombo à América), também desempenham um papel importante, enquanto eventos desportivos, como as *NBA Finals* (finais do campeonato de basquetebol da NBA) e o *Superbowl* (final do campeonato de futebol americano), embora menos frequentes, ainda influenciam as vendas. A distribuição das vendas de acordo com o tipo de evento está representado na Figura 17.

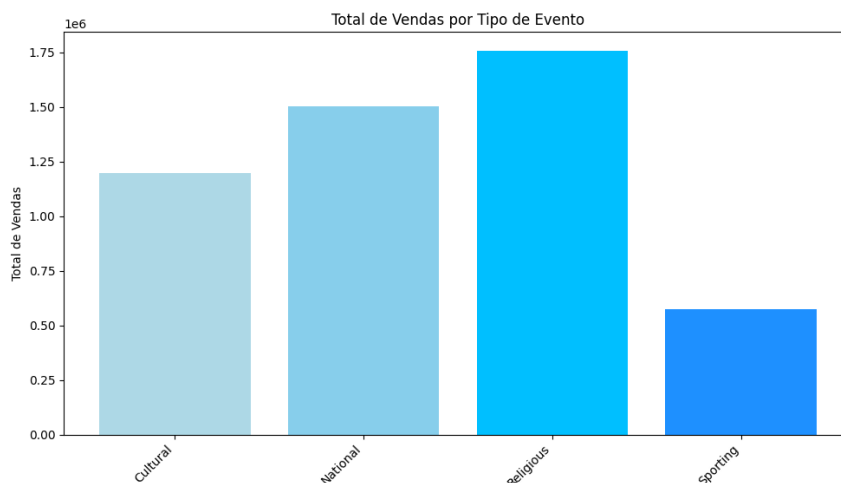


Figura 16: Distribuição das vendas por tipo de feriado

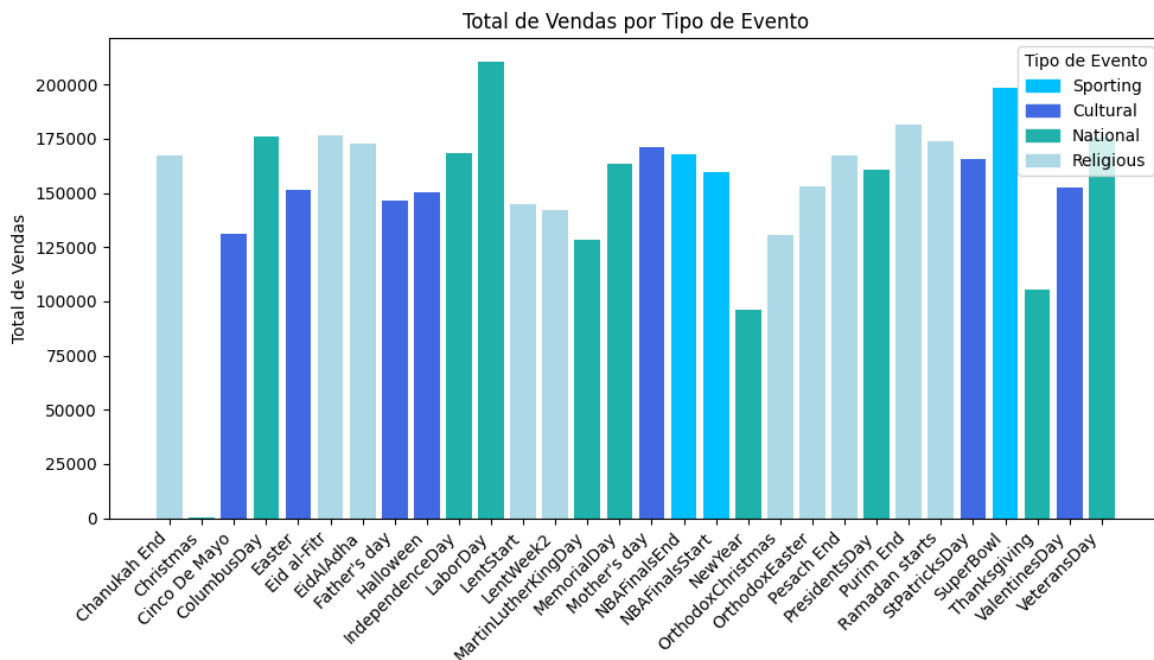


Figura 17: Distribuição das vendas por data festiva

As vendas são predominantemente maiores durante feriados religiosos, representando 34,91% do total das vendas em datas festivas. Este resultado é esperado, dado o maior número de feriados religiosos ao longo do ano (55) em comparação com os feriados desportivos (18), que são menos frequentes. A predominância de feriados religiosos na análise indica que estes períodos têm um

impacto significativo nas compras dos consumidores. A Figura 18 apresenta uma análise detalhada das vendas ao longo de diferentes períodos, destacando a influência dos dias festivos e dos fins de semana nas tendências de consumo. As barras em azul mais escuro representam dias festivos, permitindo uma comparação direta com dias normais.

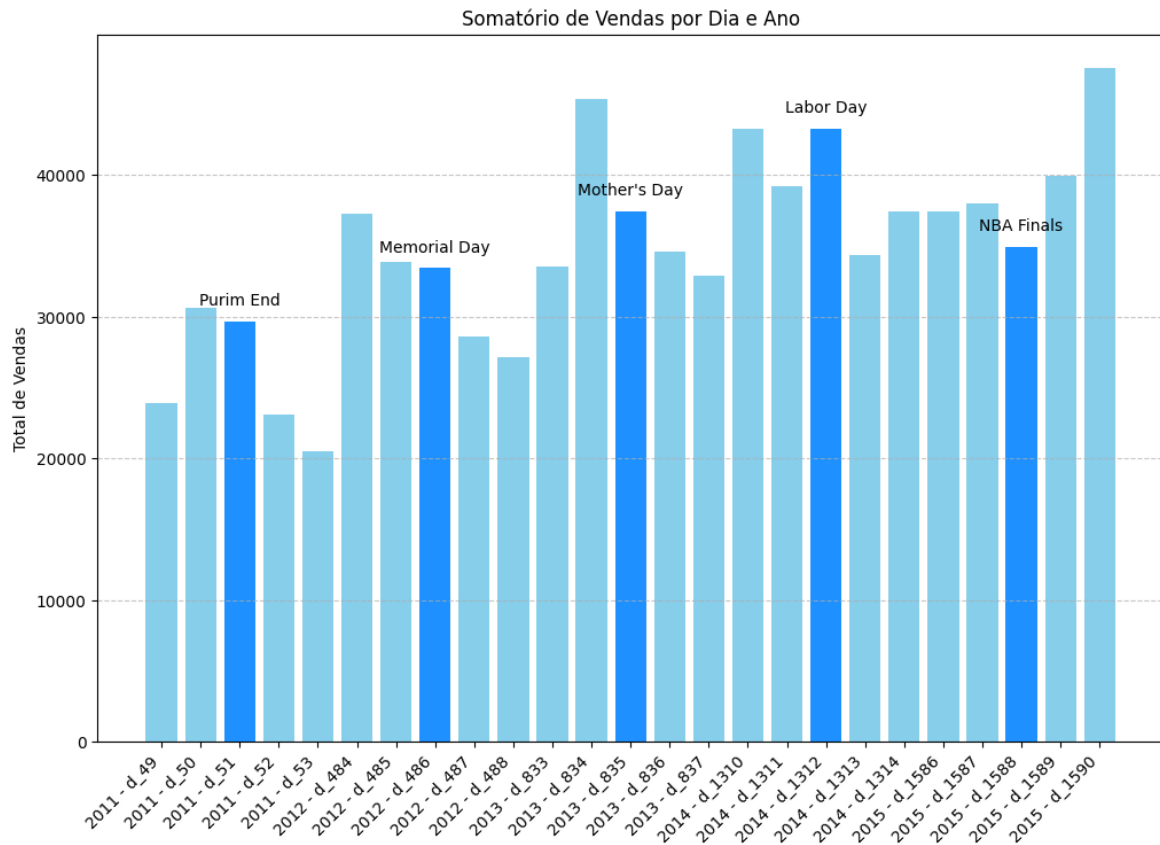


Figura 18: Influência dos dias festivos nas vendas

Analisando os primeiros dias do gráfico, especificamente de d_49 até d_53, verifica-se que os dias com maior volume de vendas foram d_50 e d_51, correspondendo a um sábado e a um domingo. Estes dias, tradicionalmente associados a um maior tempo livre dos consumidores, resultam num aumento natural das vendas. Curiosamente, o dia d_51, que coincide com o feriado religioso de *Purim End*, não demonstrou um impacto significativo nas vendas, sugerindo que, apesar da importância cultural do feriado, este não impulsionou as compras de forma notável.

No período de d_484 a d_488, observa-se outro cenário interessante. O dia d_486, que marca o *Memorial Day*, um feriado nacional nos Estados Unidos, ocorreu numa segunda-feira. Nesse caso, o gráfico revela que as vendas foram surpreendentemente altas para um dia útil, aproximando-se dos volumes típicos de fim de semana. Este comportamento pode ser explicado pelo facto de muitos consumidores aproveitarem o feriado prolongado para realizar compras que normalmente fariam durante o fim de semana.

Outro período relevante é o de d_833 a d_837, onde o d_835 representa o Dia da Mãe, celebrado num domingo. As vendas nesse dia foram substancialmente elevadas em comparação com a segunda e terça-feira seguintes, mas curiosamente não superaram as vendas de sábado. Isto

indica que a maioria dos consumidores preferiu antecipar as compras para a véspera do feriado, talvez para garantir a disponibilidade de produtos ou evitar a correria do próprio Dia da Mãe.

Mais adiante, entre d_1310 e d_1314, encontra-se o feriado do Dia do Trabalho (*Labor Day*), que caiu numa segunda-feira (d_1312). A análise das vendas revela um padrão semelhante ao observado no *Memorial Day*, com volumes de vendas bastante altos, comparáveis aos de um fim de semana, reforçando a ideia de que feriados prolongados incentivam os consumidores a realizar compras em dias que normalmente seriam mais calmos.

Finalmente, no período de d_1586 a d_1590, o dia d_1588 marcou o início das finais da NBA (*NBA Finals Start*), um evento desportivo significativo, que ocorreu numa quinta-feira. As vendas nos dias anteriores foram notavelmente mais altas, sugerindo que os consumidores se prepararam com antecedência para o evento, comprando os produtos necessários para aproveitar o festival desportivo. Como esperado, o sábado seguinte (d_1590) também registou um aumento nas vendas, consistentemente com o padrão observado ao longo de todo o gráfico.

Em relação ao preço de venda dos produtos, observa-se que a maioria dos valores se situa entre 0 e 10, com a presença de alguns outliers, como é visível na Figura 19. No entanto, esses *outliers* não foram removidos da análise, uma vez que se tratam de preços de venda, e, dependendo da categoria do produto, haverá produtos mais caros e outros mais baratos.

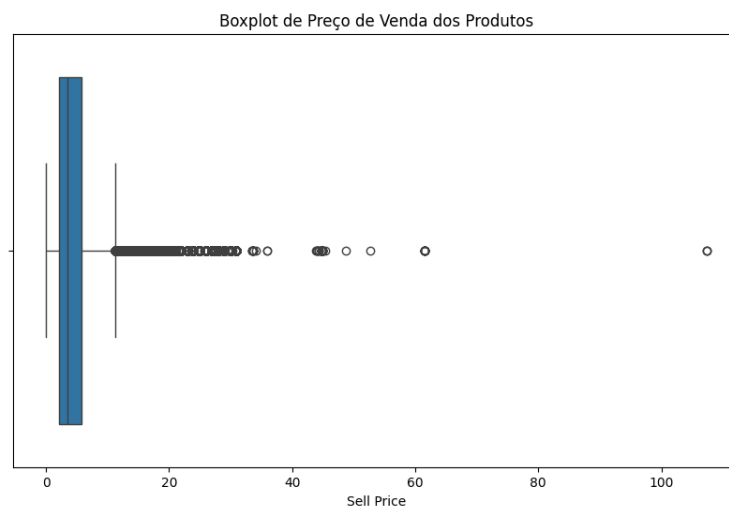


Figura 19: Boxplot do preço de venda dos produtos

A análise dos dados durante a etapa do *Data Understanding* permitiu identificar um aumento nas vendas durante os fins de semana e feriados, com variações influenciadas pelo tipo de evento e pela antecipação dos consumidores. Este padrão evidencia a necessidade de ajustar as estratégias de marketing para não só explorar o impacto do dia festivo em si, mas também os dias anteriores, quando as compras tendem a intensificar-se. A sazonalidade e a ocorrência de eventos desempenham um papel central nos hábitos de consumo, oferecendo uma oportunidade para prever melhor a procura. Essas percepções são fundamentais para afinar a gestão de inventários e planejar promoções de forma mais eficaz, permitindo às empresas ajustarem-se às mudanças do mercado e melhorarem a eficiência das operações.

3.3. Data Preparation

Na fase de preparação dos dados, foram realizados vários processos para garantir a adequação e a eficiência do conjunto de dados para a análise subsequente. Primeiramente, foi verificado que não era necessária a limpeza dos dados, uma vez que não foram identificados valores em falta e os tipos de dados estavam devidamente formatados. A única exceção referiu-se aos feriados no calendário, que estavam ausentes, mas foram codificados como 1 ou 0, o que facilita a previsão em séries temporais.

Os campos *event_name_1* e *event_name_2* foram removidos do conjunto de dados do calendário, pois, apesar de serem informativos, não têm impacto significativo no resultado geral da análise. Apenas as categorias associadas a estes eventos foram mantidas, sendo consideradas relevantes para a análise.

Em seguida, procedeu-se à codificação dos dias da semana, atribuindo o valor 1 aos dias de fim de semana (sábado e domingo) e 0 aos restantes dias. Este novo campo, denominado *weekend*, foi então adicionado ao conjunto de dados. De seguida, foram eliminados os campos *wm_yr_wk*, *wday*, *month*, *year* e *weekday*, considerados redundantes para a análise, reduzindo assim a complexidade do conjunto de dados.

Adicionalmente, os campos *event_type_1* e *event_type_2* foram transformados em variáveis numéricas, utilizando uma codificação ordinal onde os eventos de tipo '*National*' foram atribuídos o valor 4, '*Religious*' o valor 3, '*Cultural*' o valor 2, e '*Sporting*' o valor 1. Valores ausentes foram substituídos por 0, indicando a ausência de evento. Finalmente, as variáveis foram convertidas para o tipo inteiro para garantir a consistência dos dados e facilitar a análise.

Estes passos foram fundamentais para a transformação e simplificação dos dados, garantindo que as variáveis mais relevantes fossem corretamente preparadas para a modelagem subsequente.

3.4. Data Modeling

Neste estudo, com o intuito de identificar o método mais eficiente para prever as vendas dos próximos sete dias, correspondentes aos dias da semana, foram analisadas duas abordagens distintas:

- Modelação utilizando métodos tradicionais de previsão de séries temporais;
- Modelação baseada em algoritmos de *machine learning*.

Para atingir este objetivo, foram desenvolvidos e testados diversos métodos clássicos de previsão de séries temporais e algoritmos de *machine learning*, amplamente utilizados na área. Cada modelo foi configurado segundo as melhores práticas para garantir a sua eficácia. Além disso, para assegurar uma comparação justa entre os modelos, as configurações dos diferentes cenários e a divisão dos dados de treino e teste foram mantidas iguais para todos os métodos.

Cada modelo estatístico foi desenvolvido para prever os próximos 7 dias de vendas

Neste capítulo, serão apresentados os algoritmos adotados, juntamente com as respectivas configurações e estratégias de treino. Os resultados obtidos serão posteriormente sintetizados e analisados no capítulo seguinte, onde se discutirá o desempenho de cada abordagem e a sua aplicabilidade prática na previsão de vendas.

3.4.1. Métodos tradicionais de previsão de séries temporais

Os modelos tradicionais aplicados no trabalho – Média Móvel Simples (MMS), Média Móvel Exponencial (MME), ARIMA, SARIMA, *Holt-Winters* e ETS –apresentaram pontos fortes e limitações distintas, permitindo uma comparação detalhada sobre a sua eficácia.

3.4.1.1 Média Móvel Simples

Os modelos de previsão com média móvel simples foram configurados com uma janela de 7 dias e aplicados de forma global, por estado e por loja. O modelo global agregou as vendas diárias de todas as lojas, enquanto os modelos regionais desagregaram as vendas para cada estado e loja individualmente. Para calcular a média móvel, foram utilizados os dados de vendas disponíveis até o dia d_{1906} , com o conjunto de treino abrangendo as vendas de d_1 a d_{1906} . A média móvel foi calculada aplicando uma janela de 7 dias sobre esses dados históricos. O conjunto de teste, composto pelos dados de vendas de d_{1907} a d_{1913} , foi reservado exclusivamente para a avaliação do desempenho do modelo. Para cada dia do período de teste, as previsões foram feitas com base na média móvel dos últimos 7 dias de vendas históricas disponíveis, simulando como o modelo se comportaria em cenários futuros.

O *backtesting* foi crucial para validar a precisão das previsões do modelo de média móvel simples. As previsões para o período de d_{1907} a d_{1913} foram comparadas com os valores reais das vendas desses mesmos dias, que não foram utilizados na construção do modelo. Esse processo permitiu simular o desempenho do modelo em cenários futuros e avaliar a eficácia das previsões. As métricas de erro, como MAE (Erro Absoluto Médio), RMSE (Raiz do Erro Quadrático Médio), MAPE (Erro Percentual Absoluto Médio) e MSE (Erro Quadrático Médio), foram aplicadas durante o *backtesting* para medir a precisão e robustez das previsões, ajudando a identificar falhas ou ajustes necessários. No modelo por estado e por loja, a abordagem seguiu o mesmo princípio, mas ajustou as previsões às vendas agregadas de cada região ou loja individualmente. O *backtesting* foi novamente aplicado, testando a capacidade do modelo em prever os últimos 7 dias de vendas de cada estado e loja separadamente, garantindo que as previsões fossem adequadas a cada cenário específico.

3.4.1.2 Média Móvel Exponencial

O modelo de média móvel exponencial também foi configurado com uma janela de 7 dias, seguindo uma abordagem semelhante à da média móvel simples, mas atribuindo ponderações exponenciais para dar mais peso aos dados mais recentes. O conjunto de treino abrangeu os dados

de vendas desde o início até o dia d_{1906} , e a partir desse conjunto, foi calculada a média móvel exponencial. As previsões foram geradas para o período subsequente, cobrindo os dias d_{1907} a d_{1913} , que constituíram o conjunto de teste. Essas previsões foram então comparadas com os valores reais desse período para avaliar o desempenho do modelo. Um *backtesting* foi realizado para validar a eficácia das previsões, utilizando métricas de erro como MAE, RMSE, MAPE e MSE, que medem a precisão do modelo. A abordagem foi aplicada tanto globalmente quanto por estado e loja, garantindo previsões adequadas para diferentes níveis de agregação.

Para prever as vendas dos próximos 7 dias utilizando o modelo de média móvel exponencial (MME), o processo seguiu os seguintes passos: primeiro, foi calculada a MME com base nos dados históricos de vendas até o último dia disponível no conjunto de treino. Em seguida, a partir desse valor e da tendência observada nos últimos dias, foram geradas previsões para os próximos 7 dias. Cada previsão diária foi ajustada adicionando-se a tendência à última média móvel exponencial calculada. O modelo também simulou um novo valor para cada dia, incluindo um pequeno ruído aleatório, e recalculou a MME conforme os novos dados eram incorporados. Essa metodologia permitiu que as previsões fossem dinâmicas, refletindo tanto a média móvel exponencial quanto a tendência recente observada nos dados.

3.4.1.3 ARIMA

No desenvolvimento do modelo ARIMA, também foram adotadas três abordagens distintas para prever as vendas: um modelo global, um modelo por estado e um modelo por loja. Nos modelos, todos os dados disponíveis até o dia d_{1906} foram utilizados para o treino e os dados do período de d_{1907} a d_{1913} para o teste. O *backtesting* foi realizado com uma janela deslizante de 7 dias, onde o modelo foi treinado até d_{1906} e testado nos 7 dias seguintes, permitindo prever o futuro imediato. Diversas combinações de parâmetros foram avaliadas, incluindo $(1,0,1)$, $(2,1,1)$, $(5,1,0)$, $(3,1,3)$ e $(0,1,1)$, onde p representa a ordem da parte autoregressiva (AR), d é o número de diferenciações necessárias para tornar a série estacionária, e q é a ordem da parte de média móvel (MA). A configuração com o menor Erro Absoluto Médio (MAE) foi selecionada como a melhor, e com ela ajustamos o modelo final para prever as vendas dos próximos 7 dias.

Além do modelo global, foram desenvolvidos modelos ARIMA específicos para cada estado e para cada loja. Esses modelos regionais e locais permitiram uma análise mais detalhada das vendas, capturando variações e tendências específicas. Usando uma janela deslizante de 7 dias para *backtesting*, os modelos foram treinados com dados até d_{1906} e testados nos 7 dias subsequentes. Essa abordagem multifacetada proporcionou uma visão abrangente e precisa das vendas, ajustando as previsões às diferentes escalas geográficas e locais.

3.4.1.4 SARIMA

O modelo SARIMA foi ajustado para prever vendas totais a partir de dados históricos, utilizando um conjunto de treino que vai desde o início da série temporal (d_1) até ao dia d_{1906} , e um conjunto de teste que compreende os dias d_{1907} até d_{1913} . A previsão é realizada com uma janela de 7 dias, o que significa que o modelo faz previsões para os sete dias do período de teste e

avalia a sua precisão comparando com os valores reais. Esse processo de treino e teste é essencial para avaliar a capacidade preditiva do modelo, assegurando que as previsões são adequadas antes de serem aplicadas a novos dados. Ao longo desse processo, o modelo testou diferentes configurações, ajustando-se aos padrões e à variabilidade dos dados de vendas.

No que diz respeito aos parâmetros testados, várias combinações de ordem foram avaliadas para encontrar a melhor configuração do modelo SARIMA. Para o componente ARIMA do modelo, os parâmetros testados foram p (ordem autoregressiva), d (ordem de diferenciação) e q (ordem do termo de média móvel). Foram testadas combinações como $(p=1, d=0, q=1)$, $(p=2, d=1, q=1)$, $(p=5, d=1, q=0)$, $(p=3, d=1, q=3)$ e $(p=0, d=1, q=1)$. Além disso, o modelo SARIMA inclui componentes sazonais com os parâmetros P , D , Q e S , que correspondem aos padrões sazonais. O parâmetro sazonal S foi configurado como 7, refletindo a suposição de sazonalidade semanal, enquanto diferentes combinações de P , D e Q foram testadas para capturar o comportamento periódico das vendas, como $(P=1, D=1, Q=1)$ e outras variações.

Para além do modelo global, que abrange todos os estados e lojas agregadas, as mesmas configurações de modelo SARIMA foram aplicadas a nível individual para cada estado e loja. Isto permite obter previsões mais personalizadas e ajustadas às características específicas de cada região ou loja, mantendo os parâmetros e o processo de treino e teste consistentes com o modelo global. A divisão dos dados em treino e teste, o uso da janela de 7 dias e a avaliação de múltiplas combinações de parâmetros garantem que o modelo seja robusto e adaptável a diferentes granularidades dos dados, ao mesmo tempo que oferece previsões adequadas ao comportamento individual de cada série temporal de vendas.

3.4.1.5 Holt-Winters

O modelo *Holt-Winters* foi configurado para prever as vendas com uma abordagem detalhada que envolve a divisão dos dados em conjuntos de treino e teste, a definição de uma janela deslizante e a experimentação com diferentes parâmetros sazonais. O conjunto de treino abrange as vendas desde d_1 até d_{1906} , enquanto o conjunto de teste vai de d_{1907} a d_{1913} . Para avaliar a eficácia do modelo, utilizou-se uma janela deslizante de 7 dias, o que significa que o modelo ajusta-se continuamente aos dados de treino e prevê as vendas para uma janela de 7 dias do conjunto de teste em cada iteração. Esta abordagem permite uma avaliação robusta da performance do modelo ao longo do tempo e a capacidade de prever períodos futuros com base em dados históricos.

No processo de configuração do modelo, foram testadas várias configurações de sazonalidade para determinar qual delas oferece a melhor performance. Especificamente, foram considerados períodos sazonais de 7 dias e 30 dias, correspondendo a ciclos semanais e mensais, respetivamente. O modelo *Holt-Winters* foi ajustado com componentes aditivos para nível, tendência e sazonalidade, adequados para capturar padrões sazonais e tendências nas séries temporais. A escolha da melhor configuração foi feita com base na métrica de erro absoluto médio (MAE), com o objetivo de minimizar o erro nas previsões feitas pelo modelo. Após encontrar a configuração ideal, o modelo foi ajustado e utilizado para prever o período de teste e os próximos 7 dias. Essas mesmas configurações e metodologias foram então aplicadas aos modelos individuais para cada

estado e loja, permitindo uma análise detalhada e comparativa das previsões em diferentes níveis de granularidade.

3.4.1.6 ETS

No modelo ETS global, a série temporal de vendas foi dividida em conjuntos de treino e teste para avaliação da previsão. O conjunto de treino incluiu dados até o dia d_1906, enquanto o conjunto de teste abrangeu o período de d_1907 a d_1913. Para realizar a avaliação da previsão, foi utilizada uma janela deslizante de 7 dias, permitindo que cada previsão fosse baseada em dados anteriores dentro desse intervalo de tempo. Este procedimento de *backtesting* foi repetido para diferentes configurações de sazonalidade, especificamente com períodos de sazonalidade de 7 e 30 dias.

O modelo ETS foi configurado para testar essas diferentes sazonalidades, mantendo componentes aditivos para erro, tendência e sazonalidade. A configuração ideal foi selecionada com base em métricas de erro, como o Erro Absoluto Médio (MAE), o Erro Quadrático Médio (MSE), a Raiz do Erro Quadrático Médio (RMSE) e o Erro Percentual Absoluto Médio (MAPE). As mesmas configurações de parâmetros foram então aplicadas individualmente para cada estado e loja, seguindo o mesmo processo de divisão dos dados em treino e teste, além da utilização da janela de 7 dias para avaliar o desempenho do modelo em diferentes contextos regionais e de lojas.

3.4.2. Métodos de previsão utilizando *Machine Learning*

No presente trabalho, aplicaram-se várias metodologias de aprendizagem automática para modelar e prever as vendas para os próximos 7 dias, considerando diversos fatores como feriados e fins de semana. Em todos os modelos, os dados das vendas diárias foram integrados com informações do calendário, incluindo indicadores de fins de semana e feriados. As variáveis categóricas associadas a eventos foram convertidas em variáveis binárias através de técnicas de codificação (Label Binarizer), o que possibilitou a sua inclusão nos modelos preditivos.

Para capturar as nuances regionais e operacionais, foram desenvolvidas três abordagens distintas de modelação. A primeira abordagem envolveu o desenvolvimento de modelos globais, treinados com os dados agregados de todos os estados e lojas. O objetivo desta estratégia foi identificar padrões mais amplos de vendas, considerando tendências que se pudessem aplicar de forma transversal a todas as regiões. Ao treinar um modelo único com os dados combinados de todos os estados e lojas, procurou-se captar fatores comuns globais que influenciam as vendas em diferentes áreas, oferecendo assim uma visão macro das dinâmicas de mercado.

Paralelamente, foram criados modelos que treinavam com os dados de vendas de cada estado, de forma individual. Esta abordagem teve como foco ajustar os modelos às características específicas de cada região, como sazonalidade, variações regionais nos padrões de consumo, feriados estaduais e eventos locais que pudessem impactar as vendas. Ao separar os dados por estado, os modelos foram capazes de capturar essas particularidades com maior precisão, resultando em previsões mais alinhadas com o comportamento de vendas de cada região. Adicionalmente, aplicou-se a mesma lógica a nível de loja. Foram desenvolvidos modelos para

prever as vendas de cada loja de forma individual, permitindo a criação de modelos altamente personalizados para cada unidade. Após a criação de modelos a nível de estado e por loja, foi calculada a média das métricas de cada um, permitindo assim a obtenção de um agregado global por estado e por loja, o que possibilitou a comparação destes modelos com o modelo global.

Para enriquecer a análise e explorar diferentes horizontes temporais de previsão, foram estabelecidos quatro cenários distintos de treino dos modelos. No primeiro cenário, os modelos foram treinados utilizando todo o histórico de dados disponível, proporcionando uma visão completa das vendas ao longo de todo o período analisado. Esta abordagem foi útil para capturar tendências de longo prazo e sazonalidades anuais que poderiam influenciar as previsões de vendas.

O segundo cenário limitou os dados de treino ao último ano de vendas, com o objetivo de capturar padrões mais recentes e relevantes, que poderiam ter maior influência sobre as vendas futuras, especialmente num ambiente económico ou de mercado em constante mudança. No terceiro cenário, o foco foi nos dados dos últimos três meses, visando identificar tendências de curto prazo que pudessem estar mais intimamente ligadas às condições atuais do mercado, como mudanças sazonais de curto prazo ou reações a eventos recentes. Por fim, o quarto cenário utilizou apenas os dados do último mês para treinar os modelos. Este cenário foi projetado para capturar as variações de curtíssimo prazo nas vendas, permitindo que o modelo fosse altamente sensível a tendências emergentes e flutuações rápidas no comportamento dos consumidores.

Para todos os modelos, as variáveis independentes, como fim de semana e eventos, e a variável dependente, que são as vendas diárias, foram normalizadas e divididas em conjuntos de treino e teste (com configuração 80% para treino e 20% para teste), a fim de garantir a robustez e a generalização dos modelos. Todos os modelos de *Machine Learning* também foram treinados com um recurso adicional, que foi a variável “*sales_lag*”, que representa as vendas do dia anterior. Esta variável foi incluída para captar as dependências de curto prazo nas vendas, permitindo que o modelo considere o impacto das vendas do dia anterior nas vendas atuais.

A inclusão de “*sales_lag*” é relevante porque muitas séries temporais apresentam autocorrelação, ou seja, os valores atuais tendem a estar correlacionados com os valores recentes passados. Ao adicionar esta variável como uma das independentes, o modelo pode aprender a importância desse fator e ajustar as suas previsões com base nesta informação histórica. Isto é útil para identificar padrões de sazonalidade diária e tendências de curto prazo que não seriam evidentes apenas com variáveis como eventos e fins de semana. A combinação de “*sales_lag*” com outras variáveis, como eventos e indicadores de fins de semana, permite que o modelo ajuste as suas previsões considerando múltiplos fatores que influenciam as vendas diárias. Isto não só melhora a precisão das previsões, como também oferece uma compreensão mais aprofundada de como diferentes variáveis afetam as vendas em contextos temporais variados e dias da semana.

Os modelos treinados foram então avaliados através de uma série de métricas, que permitiram uma análise detalhada da precisão dos modelos na predição das vendas e da qualidade do ajuste das previsões. A comparação das métricas entre os diferentes cenários de treino forneceu percepções valiosas sobre como a escolha do período de dados de treino influencia a acurácia das previsões. As previsões foram comparadas com as vendas reais em gráficos de dispersão, facilitando a visualização da relação entre as vendas reais e as vendas previstas. Esta análise gráfica destacou a eficácia dos modelos propostos em capturar as variações nas vendas diárias, ao mesmo

tempo que forneceu uma base visual para identificar possíveis áreas de melhoria no ajuste dos modelos.

3.4.2.1 Regressão Linear Simples

O primeiro modelo aplicado foi a regressão linear simples. A escolha pela regressão linear simples se deve à sua simplicidade e capacidade de fornecer uma interpretação direta das relações entre as variáveis preditoras e a variável dependente, neste caso, as vendas diárias. Além disso, o modelo é eficiente em termos computacionais, o que facilita sua aplicação em grandes volumes de dados, como os utilizados neste trabalho. O modelo de regressão linear simples é configurado para prever as vendas diárias com base em variáveis explicativas que incluem eventos do calendário, status de fim de semana e as vendas do dia anterior, conhecida como "*sales_lag*". Os dados são primeiramente divididos em conjuntos de treino e teste, alocando 80% dos dados para treinamento e 20% para teste. A normalização dos dados é realizada com o *StandardScaler*, garantindo que todas as variáveis preditoras estejam na mesma escala. A variável "*sales_lag*" é especialmente importante, pois captura a dependência temporal, permitindo que o modelo leve em consideração as vendas do dia anterior ao prever as vendas atuais, melhorando a acurácia das previsões.

Após o treinamento do modelo, sua performance é avaliada com métricas como erro médio quadrático (MSE), coeficiente de determinação (R^2) e erro absoluto médio (MAE). O modelo faz previsões para os próximos sete dias, onde a previsão de um dia é usada como entrada para prever o dia seguinte, formando uma cadeia de previsões. Esse processo possibilita a geração de uma série de previsões que consideram tanto o comportamento recente das vendas quanto variáveis como fins de semana e eventos específicos, capturando assim a dinâmica temporal e sazonal das vendas.

3.4.2.2 Regressão Polinomial

No modelo de regressão polinomial, as configurações da regressão linear simples foram aplicadas com uma transformação adicional para capturar relações não-lineares entre as variáveis. Os dados de vendas diárias foram organizados e combinados com informações de calendário, como fins de semana e feriados, que foram convertidos em formato binário. A transformação polinomial de grau 2 permitiu incluir interações e termos quadráticos, tornando o modelo mais eficaz em detectar padrões complexos entre as variáveis independentes e as vendas diárias. As previsões para os próximos 7 dias foram avaliadas com base em métricas de desempenho, como erro médio quadrático (MSE), erro absoluto médio (MAE), coeficiente de determinação (R^2) e erro médio percentual absoluto (MAPE), e o resultado inclui gráficos comparativos entre as vendas reais e previstas para cada estado, além de uma estimativa total das vendas previstas para o próximo dia.

3.4.2.3 Decision Tree

O modelo de árvore de decisão, conhecido como *Decision Tree*, opera segmentando os dados em subconjuntos com base em critérios específicos dos atributos, formando uma estrutura hierárquica. Cada nó da árvore representa uma decisão sobre um atributo, enquanto cada folha

corresponde a uma previsão ou classe final. A configuração das variáveis independentes e dependentes, bem como a divisão dos dados em conjuntos de treino e teste, foi mantida consistente com os demais modelos utilizados no projecto. O modelo inicial, construído com os parâmetros padrão (sem ajuste de hiperparâmetros), mostrou-se capaz de gerar previsões razoáveis. No entanto, havia margem para melhorias em termos de precisão e robustez.

Com o intuito de aprimorar o desempenho, aplicou-se a técnica de otimização de hiperparâmetros chamada *GridSearchCV*. Enquanto o modelo básico utilizava configurações padrão, o processo de otimização explorou uma variedade de parâmetros para ajustar melhor o comportamento da árvore de decisão. A grelha de hiperparâmetros testada incluiu variações na profundidade máxima da árvore (*max_depth*), com valores de 3, 5, 7 e sem limite; no número mínimo de amostras para dividir um nó (*min_samples_split*), com valores de 2, 5 e 10; e no número mínimo de amostras numa folha (*min_samples_leaf*), com valores de 1, 2 e 4.

O *GridSearchCV* avaliou todas as combinações possíveis destes hiperparâmetros usando validação cruzada de 3 dobras, o que garantiu uma avaliação consistente do modelo e minimizou o risco de *overfitting*. Como resultado, foi seleccionada a combinação de hiperparâmetros que proporcionou o melhor desempenho preditivo, tornando o modelo mais preciso e adaptado às características específicas dos dados. Este processo de otimização não apenas preservou a estrutura fundamental do modelo original, mas também refinou os parâmetros que controlam o comportamento da árvore de decisão. O modelo aprimorado resultante foi significativamente mais fiável e preciso, oferecendo uma ferramenta mais robusta para a previsão de vendas.

3.4.2.4 XGBoost

O modelo *XGBoost* foi inicialmente aplicado sem ajustes específicos de hiperparâmetros, utilizando as mesmas configurações dos outros modelos para prever as vendas por estado e por loja. Neste primeiro modelo, a configuração básica incluía um número fixo de estimadores (*n_estimators*), uma profundidade máxima padrão das árvores de decisão (*max_depth*), e uma taxa de aprendizagem padrão (*learning_rate*). Esta abordagem visava estabelecer uma linha de base para o desempenho do modelo, permitindo uma análise inicial das previsões geradas. O modelo básico já demonstrou ser eficaz, mas a ausência de ajustes nos parâmetros limitou o seu potencial máximo de precisão.

Para aprimorar o modelo, foi realizada uma otimização de hiperparâmetros por meio do *GridSearchCV*, testando várias combinações. O *n_estimators* foi ajustado para 50, 100 e 200, com mais árvores potencialmente melhorando a precisão, mas aumentando o risco de *overfitting*. O *max_depth* foi testado com valores de 3, 5 e 7, onde maior profundidade permite capturar interações mais complexas, mas também pode causar *overfitting*. A *learning_rate* variou entre 0,01, 0,1 e 0,2, com taxas menores promovendo uma aprendizagem mais gradual e robusta.

Outros parâmetros incluíram *subsample* (0,6, 0,8 e 1,0), ajustando a proporção de amostras usadas em cada árvore para reduzir a variabilidade; *colsample_bytree* (0,6, 0,8 e 1,0), que controla a diversidade das árvores; e *gamma* (0, 0,1 e 0,2), que regula o ganho mínimo necessário para dividir um nó, impactando a complexidade do modelo. O *GridSearchCV* identificou a combinação ideal que maximiza o *R² Score*, proporcionando previsões mais precisas.

Tentativas adicionais de otimização incluíram a criação de novas características, como *sales_diff* (diferença entre vendas de dias consecutivos), *sales_rolling_mean_7* (média móvel de sete dias) e *sales_rolling_std_7* (desvio padrão de sete dias), com o objetivo de capturar padrões temporais e variabilidade. Também foi implementado um *pipeline* avançado com pré-processamento usando *ColumnTransformer* e validação cruzada com KFold. No entanto, essas melhorias não resultaram em ganhos significativos, sugerindo que as limitações já identificadas não foram superadas pelos ajustes adicionais.

3.4.2.5 Support Vector Machines

O modelo SVM (*Support Vector Machines*) foi aplicado para prever as vendas de diferentes estados e lojas. Um SVM com *kernel* radial básico foi então aplicado, mas sem ajustes de hiperparâmetros, resultando em previsões razoáveis, porém limitadas pela ausência de variáveis temporais. Para aprimorar o modelo, foram adicionadas variáveis de vendas passadas e médias móveis (*sales_lag_1*, *sales_lag_7*, *sales_lag_14*, *sales_ma_7*, *sales_ma_14*) para capturar padrões temporais mais complexos. Utilizou-se *GridSearchCV* para testar várias combinações de hiperparâmetros, incluindo diferentes tipos de *kernel* (rbf, linear, poly), valores de C (0.1, 1, 10, 100), *gamma* (scale, auto), e *epsilon* (0.01, 0.1, 1), com validação cruzada de 5 dobras. O melhor modelo minimizou o erro quadrático médio (MSE) e maximizou o coeficiente de determinação (R^2), melhorando significativamente a precisão das previsões.

3.5. Evaluation

Na etapa de *Evaluation*, todos os modelos estatísticos e de *machine learning* desenvolvidos foram rigorosamente avaliados utilizando um conjunto abrangente de métricas de desempenho, nomeadamente R^2 , MAE, RMSE, MAPE e MSE. O coeficiente de determinação (R^2) foi utilizado para medir a proporção da variabilidade total dos dados que o modelo é capaz de explicar, sendo um indicador crucial da precisão geral do modelo. O erro absoluto médio (MAE) foi calculado para fornecer uma medida direta da média das diferenças absolutas entre os valores previstos e os valores reais, oferecendo uma visão clara sobre a magnitude dos erros de previsão. O erro quadrático médio (MSE) e a sua raiz quadrada (RMSE) foram empregados para penalizar mais severamente grandes desvios nas previsões, com o RMSE sendo particularmente útil para interpretar a precisão do modelo na mesma escala das variáveis originais. Finalmente, o erro percentual absoluto médio (MAPE) foi incluído para avaliar o desempenho do modelo em termos relativos, permitindo uma compreensão do erro médio em percentagem. Ao aplicar estas métricas de forma conjunta, foi possível obter uma avaliação detalhada e equilibrada do desempenho dos modelos, garantindo que não apenas a precisão, mas também a robustez e a consistência das previsões, fossem rigorosamente analisadas.

4. RESULTADOS E DISCUSSÃO

A previsão de vendas é um desafio crucial para o setor do retalho, onde a precisão nas estimativas pode influenciar diretamente a eficiência operacional e o sucesso comercial. No presente capítulo, realiza-se a análise e avaliação dos resultados obtidos através da aplicação de diferentes abordagens para prever as vendas dos próximos 7 dias de forma global, e para cada estado e loja individualmente.

O objetivo principal deste trabalho é identificar a estratégia mais eficaz para prever com maior precisão as vendas futuras, considerando tanto os dados gerais como os específicos de cada estado e loja. Para tal, foram utilizados diferentes modelos, abrangendo desde métodos tradicionais de previsão de séries temporais, como a média móvel simples, a média móvel exponencial, ARIMA, SARIMA, *Holt-Winters* e ETS, até técnicas mais avançadas de *machine learning*, incluindo Regressão Linear Simples, Regressão Polinomial, *Decision Tree*, *XGBoost*, e SVR.

Um aspecto fundamental a salientar é que todos os modelos estatísticos foram treinados e avaliados sob as mesmas condições, utilizando o mesmo conjunto de treino e teste, assim como a mesma janela deslizante. Isto assegura uma comparação justa entre as diferentes abordagens. Os modelos de *machine learning* também foram treinados e avaliados com características gerais semelhantes, o que permite uma comparação equitativa. Foram considerados vários períodos de treino para os modelos de aprendizagem automática, abrangendo todos os dados disponíveis, bem como intervalos mais recentes de um ano, três meses e um mês, a fim de avaliar o impacto do período de treino na performance dos modelos ao longo de diferentes cenários temporais.

Ao longo deste capítulo, são apresentados e discutidos os resultados detalhados de cada abordagem, destacando os modelos que demonstraram maior potencial em cada cenário de previsão. Esta análise permite uma compreensão aprofundada da eficácia de cada método, fornecendo perceções valiosas para a aplicação prática em cenários reais de previsão de vendas.

4.1. Modelação baseada em Métodos Tradicionais

Nesta primeira análise, foram aplicados ao conjunto de dados os modelos baseados em métodos tradicionais de previsão de séries temporais mencionados anteriormente. O modelo que apresentou o melhor desempenho foi o SARIMA, que será detalhado a seguir. Os resultados dos outros modelos estatísticos – Média Móvel Simples, Média Móvel Exponencial, ARIMA, *Holt-Winters* e ETS – estão disponíveis no Anexo A.

4.1.1. Resultados do Modelo SARIMA

A aplicação do modelo SARIMA aos dados de vendas apresentou resultados que demonstram tanto a eficácia quanto os desafios envolvidos na previsão de séries temporais com sazonalidade

complexa. O modelo SARIMA foi escolhido pela sua capacidade de capturar padrões sazonais e tendências, essenciais para a análise preditiva em contextos como o de vendas diárias.

O modelo global SARIMA, configurado com a melhor combinação de parâmetros ((2, 1, 2), (1, 1, 1, 7)), apresentou um desempenho notável nas métricas de avaliação. O Erro Absoluto Médio (MAE) foi de aproximadamente 1062, indicando uma precisão razoável nas previsões. O Erro Quadrático Médio (MSE) alcançou cerca de 1.930.808, enquanto a Raiz do Erro Quadrático Médio (RMSE) foi de aproximadamente 1389, sugerindo que, apesar da precisão geral, existiram algumas previsões com discrepâncias significativas. O Erro Percentual Absoluto Médio (MAPE) foi de 2,80%, refletindo um bom nível de precisão nas previsões, com erros percentuais relativamente baixos. Para os próximos 7 dias, o modelo SARIMA prevê vendas variando entre aproximadamente 36.889 e 49.587, proporcionando uma visão clara das expectativas futuras para o período analisado.

A análise por estado revelou diferenças significativas nas previsões de vendas. Para a Califórnia, a melhor configuração de parâmetros foi ((2, 1, 1), (2, 1, 0, 7)), com um MAE de aproximadamente 402 e um MAPE de 2,39%, indicando previsões relativamente precisas. No Texas (TX), a combinação de parâmetros ((1, 1, 2), (0, 1, 2, 7)) resultou em um MAE de 716 e um MAPE de 6,62%, mostrando previsões com maior margem de erro percentual. Para Wisconsin (WI), o modelo resultou em um MAE de 700 e um MAPE de 5,35%, com previsões variando de 9.525 a 14.625. Esses resultados sublinham a importância de ajustes específicos para cada região, oferecendo previsões ajustadas às características locais. Os resultados das métricas do modelo SARIMA podem ser visualizados na Tabela 2, e os resultados das previsões estão disponíveis na Tabela 3.

Tabela 2: Resultados das métricas com SARIMA

Estado/loja	Métricas de Avaliação				Parâmetro
	MAE	RMSE	MAPE	MSE	
Modelo por estado					
Califórnia	402	471	2,39	222.173	(2, 1, 1), (2, 1, 0, 7)
Texas	716	778	6,62	606.171	(1, 1, 2), (0, 1, 2, 7)
Wisconsin	700	853	5,35	727.875	(2, 1, 1), (2, 1, 0, 7)
Modelo por Loja					
CA_1	98	117	2,20	13.922	(2, 1, 1), (2, 1, 0, 7)
CA_2	102	133	2,69	17.880	(2, 1, 1), (2, 1, 0, 7)
CA_3	279	357	4,56	127.906	(2, 1, 1), (2, 1, 0, 7)
CA_4	78	115	2,69	13.377	(1, 1, 0), (0, 1, 1, 7)
TX_1	155	170	4,86	29.108	(2, 1, 2), (1, 1, 1, 7)
TX_2	265	327	6,98	107.295	(1, 1, 0), (0, 1, 1, 7)
TX_3	330	386	9,21	149.304	(1, 1, 0), (0, 1, 1, 7)
WI_1	118	217	2,76	47.093	(1, 1, 1), (1, 1, 1, 7)
WI_2	573	682	11,77	465.625	(2, 1, 1), (2, 1, 0, 7)
WI_3	242	290	6,23	84.544	(2, 1, 1), (2, 1, 0, 7)
Modelo Global	1062	1389	2,80	1.930.807	(2, 1, 2), (1, 1, 1, 7)

A análise das previsões de vendas para as lojas revelou uma diversidade de resultados entre as diferentes localizações. Na loja CA_1, a melhor configuração do modelo SARIMA foi ((2, 1, 1), (2, 1, 0, 7)), apresentando métricas de erro bastante favoráveis: um MAE de 98 e um MAPE de apenas 2,20%. As previsões para os próximos sete dias indicam vendas entre 3.799 e 6.028. A CA_2 utilizou

a mesma configuração, resultando em um MAE maior de 102 e MAPE de 2,69%, com vendas projetadas oscilando entre 3.657 e 6.041. Em contraste, a CA_3 teve um desempenho inferior, com um MAE de 279 e MAPE de 4,56%, prevendo vendas entre 5.317 e 8.142.

Nos resultados das lojas no Texas, as variações também foram notáveis. A TX_1 obteve um MAE de 155 e um MAPE de 4,86%, enquanto a TX_2 apresentou um MAE de 265 e MAPE de 6,98%, com previsões de vendas entre 3.024 e 4.397. Em Wisconsin, as variações continuaram: a WI_1 teve um MAE de 118 e MAPE de 2,76%, enquanto a WI_2 apresentou um MAE mais elevado, de 573, e um MAPE de 11,77%. Essas diferenças ressaltam a importância de ajustar os parâmetros do modelo SARIMA para capturar as particularidades de cada loja.

Tabela 3: Resultados das previsões com SARIMA

Estado/loja	Previsão das Vendas						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Califórnia	17333	15701	15165	14952	17062	21933	23441
CA_1	4515	3798	3712	3676	4351	5687	6027
CA_2	3882	3656	3663	3484	4461	5837	6040
CA_3	6146	5661	5317	5280	5699	7395	8141
CA_4	2894	2752	2680	2692	2835	3142	3207
Texas	10525	9550	9541	9545	10370	12332	12925
TX_1	3200	2983	2966	2972	3192	3843	4056
TX_2	3475	3023	3064	3096	3478	4275	4397
TX_3	3474	3109	3080	3079	3364	3933	4036
Wisconsin	10027	10109	9524	9649	11933	14624	13041
WI_1	3270	3080	3168	3241	4009	5265	4817
WI_2	3700	4043	3876	3712	4715	5207	4358
WI_3	3041	3146	2802	2790	3535	4478	4002
Total	39201	36888	36256	36527	40327	49374	49586

Após a análise das métricas de desempenho do modelo SARIMA, foi possível calcular modelos agregados por loja e por estado, conforme visualizado na Tabela 4. O modelo agregado por loja foi construído a partir da média das métricas de desempenho dos modelos individuais de cada loja, enquanto o modelo agregado por estado foi baseado na média das métricas dos modelos individuais de cada estado. Esta abordagem permite captar as nuances e as variabilidades específicas que podem afetar as vendas em diferentes contextos geográficos e operacionais.

O modelo agregado por loja destacou-se como o mais eficaz, apresentando um MAE de 224, RMSE de 279 e MSE de 105.605. Estes resultados refletem a menor discrepância média entre as previsões e os valores reais, além de demonstrar uma alta precisão na maioria dos casos. Esta precisão superior em termos absolutos sugere que o modelo é capaz de captar variáveis locais, como sazonalidade e comportamento do consumidor, que impactam diretamente as vendas. Em comparação, o modelo global, embora tenha exibido um MAPE de 2,8%, apresentava erros absolutos significativamente maiores, com um MAE de 1062 e um RMSE de 1389. Esta diferença evidencia que, enquanto o modelo global é eficaz para análises de desempenho gerais, ele pode obscurecer as particularidades que afetam as vendas de lojas individuais.

Tabela 4: Comparação entre os modelos agregados com SARIMA

Métrica	Agregado Global	Agregado por Estado	Agregado por Loja
MAE	1062	606	224
RMSE	1389	700	279
MAPE	2,8%	4,78%	5,39%
MSE	1.930.807	518.739	105.605
Previsão	Agregado Global	Agregado por Estado	Agregado por Loja
Dia 1	39201	37885	37597
Dia 2	36888	35360	35251
Dia 3	36256	34230	34328
Dia 4	36527	34146	34022
Dia 5	40327	39365	39639
Dia 6	49374	48889	49062
Dia 7	49586	49407	49081

A importância desta análise reside na capacidade de informar decisões estratégicas mais precisas e direcionadas. O modelo agregado por loja não apenas proporciona previsões mais exatas, mas também permite uma compreensão mais profunda das necessidades específicas de cada loja, possibilitando ajustes em estratégias de *marketing* e gestão de estoque. A escolha do modelo adequado para previsões é crucial, especialmente em sectores onde a competição é intensa e a capacidade de responder rapidamente às flutuações do mercado pode determinar o sucesso ou o fracasso de uma operação. Portanto, a análise das métricas de desempenho não apenas reflete a eficácia do modelo, mas também serve como um guia para a formulação de estratégias mais eficazes e fundamentadas, otimizando o desempenho das vendas em diferentes contextos.

É possível observar na Figura 20 o gráfico da comparação entre as vendas reais e as previsões para os 7 dias utilizando o modelo SARIMA. Observa-se que a linha das previsões fica bem próxima a linhas das vendas totais reais.

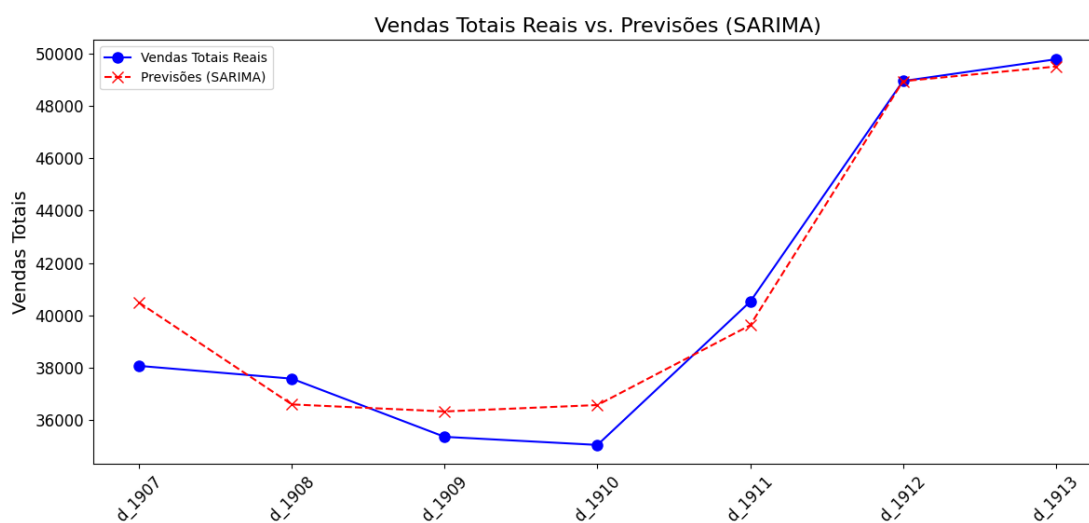


Figura 20: Gráfico de comparação entre as vendas reais e as previsões utilizando SARIMA

4.1.2. Comparação entre as Abordagens Estatísticas

A análise dos resultados dos modelos estatísticos aplicados com diferentes níveis de agregação – global, por estado e por loja –, com os resultados disponíveis da Tabela 5, revela tendências importantes sobre a eficácia de cada abordagem em termos de precisão e capacidade explicativa. Entre todos os modelos, o SARIMA surge como o mais eficaz, destacando-se em praticamente todos os agregados. No agregado por loja, o SARIMA atinge o menor erro absoluto (MAE de 224), o menor erro quadrático médio (RMSE de 279), e o menor MAPE de 5,39%, além de um valor baixo de erro quadrático médio total (MSE de 105.605). Estes resultados indicam que o modelo SARIMA, ao ser aplicado individualmente por loja, oferece as previsões mais precisas, minimizando tanto os desvios absolutos quanto os erros quadráticos, e mantendo um erro percentual relativamente baixo. Isso torna o agregado por loja com o modelo SARIMA o melhor conjunto de resultados, sugerindo que a abordagem por loja permite capturar melhor as particularidades de cada local individual, reduzindo os erros de previsão.

Tabela 5: Comparação entre os modelos estatísticos agregados

Modelo	Métricas de Avaliação			
	MAE	RMSE	MAPE	MSE
Agregado Global				
MMS	3618	5228	7,91	27.337.569
MME	3184	4395	7,08	19.316.030
ARIMA	2054	2930	5,28	8.589.153
SARIMA	1062	1389	2,80	1.930.807
Holt-Winters	1351	1618	3,47	2.619.261
ETS	1186	1304	2,90	1.702.732
Agregado por Estado				
MMS	1261	1776	7,98	3.789.220
MME	1136	1516	7,31	2.724.666
ARIMA	891	1120	6,67	1.314.569
SARIMA	606	700	4,78	518.739
Holt-Winters	699	804	5,68	681.105
ETS	653	726	5,30	574.064
Agregado por Loja				
MMS	387	546	8,31	360.729
MME	355	477	7,84	267.517
ARIMA	344	418	8,44	205.366
SARIMA	224	279	5,39	105.605
Holt-Winters	243	292	6,12	101.116
ETS	232	284	5,82	102.764

O segundo melhor modelo é o ETS, que também mostra um forte desempenho no agregado por loja, com um MAE de 232, RMSE de 284 e MAPE de 5,82%, ficando logo atrás do SARIMA em termos de precisão. No entanto, o ETS no agregado por estado oferece um equilíbrio interessante entre erro absoluto e percentual (MAE de 653, RMSE de 726 e MAPE de 5,30%). Este modelo

destaca-se pela sua robustez, particularmente em cenários onde se deseja uma explicação sólida da variabilidade dos dados a nível estadual.

Por outro lado, os modelos MMS e MME, embora apresentem um desempenho aceitável, são os piores em todos os agregados, com os maiores valores de MAE, RMSE e MAPE. No agregado global, por exemplo, o MMS apresenta um MAE de 3618 e um RMSE de 5228, o que indica grandes erros de previsão. A situação melhora nos agregados por estado e por loja, mas mesmo assim, esses modelos não conseguem competir com os melhores, como o SARIMA e o ETS.

Quando se comparam os diferentes níveis de agregação, fica claro que o agregado por loja proporciona os melhores resultados em termos de minimização dos erros de previsão. Isso deve-se ao fato de que modelos individualizados para cada loja conseguem capturar as nuances locais de venda, permitindo uma previsão mais precisa. O agregado por loja exhibe consistentemente os menores valores de MAE e RMSE em praticamente todos os modelos testados, sugerindo que a abordagem local é superior quando o foco é minimizar erros de previsão.

Por outro lado, o agregado por estado oferece uma explicação mais robusta da variabilidade dos dados numa escala mais ampla, como evidenciado pelo R^2 superior em vários modelos. Essa abordagem é mais eficaz quando se deseja entender as tendências gerais de um estado, mesmo que à custa de maior erro absoluto. O agregado por estado é uma opção interessante quando o foco é a análise regional, e mesmo assim, apresenta resultados de erro bastante competitivos em comparação com o agregado global.

Já o agregado global, embora tenha a vantagem de ser mais simples e menos complexo em termos de implementação e manutenção, tende a apresentar os piores resultados de erro, pois mistura as características individuais de cada local ou estado, o que pode levar a previsões menos precisas. O modelo SARIMA, por exemplo, que tem um MAE de 1062 no agregado global, apresenta uma redução significativa de erros quando desagregado por estado (MAE de 606) e por loja (MAE de 224), evidenciando que uma abordagem global perde a capacidade de capturar variações locais.

Os modelos globais têm a vantagem de serem simples e eficientes quando se lida com um grande volume de dados. No entanto, como evidenciado pelos resultados, essa abordagem tende a diluir as variações locais, resultando em previsões menos precisas. Em contrapartida, os modelos individualizados, seja por estado ou por loja, conseguem capturar melhor as especificidades de cada região ou loja, resultando numa menor margem de erro, como é o caso do SARIMA e do ETS no agregado por loja. A desvantagem desses modelos desagregados está na complexidade adicional de implementação e manutenção, especialmente quando há muitas lojas ou estados a serem modelados individualmente.

Em conclusão, o melhor modelo e abordagem dependem do objetivo final da análise. Se o foco for a precisão absoluta, o agregado por loja com o modelo SARIMA é claramente superior. No entanto, se a prioridade for compreender tendências regionais ou estaduais, o agregado por estado com o modelo ETS ou SARIMA pode ser apropriado, ainda que com um leve aumento no erro absoluto. A abordagem global, apesar de prática, mostra-se menos eficiente e recomendada apenas em cenários onde a simplicidade é mais importante do que a precisão detalhada.

4.2. Modelação baseada em *Machine Learning*

Nesta segunda abordagem, foram aplicados ao conjunto de dados modelos baseados em aprendizagem automática. O modelo que demonstrou o melhor desempenho foi o *Support Vector Machines* (SVM), que será descrito em detalhe a seguir. Os resultados dos outros modelos de aprendizagem automática — Regressão Linear Simples, Regressão Polinomial, *Decision Tree* e *XGBoost* — encontram-se disponíveis em detalhes no Anexo A.

4.2.1. *Support Vector Machines* (SVM)

A Máquina de Vetores de Suporte (SVM) é um algoritmo poderoso de aprendizagem supervisionada, que visa encontrar a melhor linha (ou hiperplano, em múltiplas dimensões) capaz de minimizar os erros de previsão, assegurando que a maioria dos pontos de dados se encontre dentro de uma margem de tolerância pré-definida.

Ao comparar os resultados obtidos com o modelo SVM antes e depois da aplicação de ajustes nos hiperparâmetros, observa-se uma melhoria significativa na precisão das previsões. Inicialmente, os valores do R^2 variavam entre 0,216 e 0,441, indicando um ajuste limitado, com explicação de apenas 44% da variabilidade nas vendas. Após a inclusão dos hiperparâmetros através do GridSearch, os valores do R^2 aumentaram substancialmente, alcançando entre 0,709 e 0,785, o que demonstra uma explicação muito mais robusta das variações observadas nas vendas. Além disso, a redução nos erros absolutos médios (MAE) e nos erros quadráticos médios (MSE) reforça a eficácia dos modelos ajustados, que agora apresentam uma capacidade preditiva significativamente superior.

O cenário 3 (dados dos últimos 3 meses) destacou-se como o melhor desempenho global para o modelo SVM, apresentando um R^2 de 0,7909, o que indica uma explicação significativa da variabilidade dos dados. Com um MAE de 1015 e um RMSE de 1235, este cenário mostra uma precisão relativamente alta e erros moderados em comparação com os restantes cenários. O MAPE de 7,95% é também baixo, sugerindo que as previsões são consistentes e próximas dos valores reais. As previsões diárias para o cenário 3 são estáveis e bem ajustadas, variando de 49.795 no primeiro dia a 51.784 no sétimo, demonstrando que o modelo oferece previsões relativamente confiáveis e coesas ao longo da semana. Os resultados das métricas de avaliação do modelo SVM global estão disponíveis na Tabela 6, enquanto as previsões para os 7 dias podem ser visualizadas na Tabela 7.

Tabela 6: Resultados das métricas do modelo global com SVM

Modelo Global	Métricas de Avaliação				
	R^2	MAE	RMSE	MAPE	MSE
Cenário 1	0,2989	1657	2251	12,67	5.294.870
Cenário 2	0,6488	936	1189	7,31	1.511.068
Cenário 3	0,7909	1015	1235	7,95	1.682.651
Cenário 4	0,7654	965	1144	6,24	1.344.072

Tabela 7: Resultados das previsões do modelo global com SVM

Modelo Global	Previsões						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Cenário 1	29525	29302	28873	27940	27940	34076	38121
Cenário 2	49795	39493	36527	35717	35717	44892	48187
Cenário 3	49795	39800	38202	37946	37946	49147	51784
Cenário 4	49795	39296	38487	38393	38393	48758	50114

Entre os cenários analisados para o modelo SVM por estado, o cenário 3 destaca-se como o mais eficiente. Com um R^2 de 0,9126 na Califórnia, o modelo demonstra uma forte capacidade de explicar a variabilidade dos dados nesta região, o que é crucial para a formulação de estratégias de negócio informadas. Os valores de MAE e RMSE são relativamente baixos, situando-se em 753 e 986, respectivamente, indicando uma boa precisão nas previsões. O MAPE de 4,24% sugere que as previsões estão bastante próximas dos valores reais, o que é um indicador de fiabilidade.

No Texas, o cenário 3 também se revela superior, apresentando um R^2 de 0,8418 e um MAE de 711. Esses resultados demonstram uma performance consistente em termos de precisão e minimização de erros. A capacidade de previsão precisa neste estado pode ajudar os gestores a otimizar o estoque e a planejar campanhas de marketing de forma mais eficaz.

Em contraste, Wisconsin apresenta um desempenho um pouco inferior, com um R^2 de 0,6657 e um MAE de 1449. Embora ainda seja o melhor entre os cenários oferecidos para este estado, a maior margem de erro indica que existem fatores específicos que podem estar a influenciar as vendas de forma menos previsível, sendo viável ajustar o modelo para capturar melhor as dinâmicas locais. Os resultados das métricas de avaliação do modelo SVM por estado estão disponíveis na Tabela 8, enquanto as previsões para os 7 dias podem ser visualizadas na Tabela 9.

Tabela 8: Resultados das métricas do modelo por estado com SVM

Modelo por Estado	Métricas de Avaliação				
	R^2	MAE	RMSE	MAPE	MSE
Cenário 1					
Califórnia	0,5449	1595	2274	1277,52	5.172.938
Texas	0,4978	1026	1456	821,44	2.118.521
Wisconsin	0,3521	1585	2109	1724,57	4.446.138
Cenário 2					
Califórnia	0,8139	952	1186	5,64	1.406.072
Texas	0,6374	631	808	5,89	652.368
Wisconsin	0,5009	1220	1568	10,32	2.457.666
Cenário 3					
Califórnia	0,9126	753	986	4,24	971.729
Texas	0,8418	711	775	6,44	600.773
Wisconsin	0,6657	1449	1715	12,47	2.940.079
Cenário 4					
Califórnia	0,8874	1124	1213	5,78	1.473.027
Texas	0,7284	633	890	4,89	791.670
Wisconsin	0,6804	1139	1329	8,07	1.767.519

Tabela 9: Resultados das previsões do modelo por estado com SVM

Modelo por Estado	Previsões						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Cenário 1							
Califórnia	13032	13185	13235	13252	13252	15277	16108
Texas	8545	9123	9198	9210	9210	10701	11182
Wisconsin	7948	7926	7919	7917	7917	8999	9401
Cenário 2							
Califórnia	23187	16919	15368	15024	15024	18634	19752
Texas	12282	10263	9804	9706	9706	11676	12252
Wisconsin	14326	10995	9998	9716	9716	11236	11736
Cenário 3							
Califórnia	23187	16650	15470	15280	15280	19172	20146
Texas	12282	10869	10698	10677	10677	12696	13074
Wisconsin	14326	12373	12191	12174	12174	14204	14485
Cenário 4							
Califórnia	23187	16653	15836	15749	15749	21491	22690
Texas	12282	10349	10437	10433	10433	12465	12444
Wisconsin	14326	12294	12214	12211	12211	14819	15015

Analisando o modelo por loja, o cenário 3 destaca-se como o mais eficaz entre os cenários avaliados. Este cenário apresenta um desempenho notável em diversas lojas, especialmente na Califórnia, onde as lojas CA_1, CA_2 e CA_4 obtêm elevados valores de R^2 (0,918, 0,8818 e 0,7417, respectivamente). Esses resultados refletem uma forte capacidade de explicação da variabilidade dos dados, e também indicam uma redução significativa nos erros de previsão, evidenciada pelos baixos valores de MAE e RMSE. Por exemplo, as previsões em CA_1 têm um MAPE de 3,99% e um MSE de 74.202, que denotam uma precisão considerável e uma boa capacidade de modelagem.

No Texas, o cenário 3 mantém a sua superioridade, com TX_1 e TX_2 a apresentarem R^2 de 0,8373 e 0,8415. Esses valores, aliados a MAEs de 208 e 257, sublinham a robustez do modelo, permitindo previsões precisas e com uma margem de erro reduzida. Essa consistência nas previsões é essencial para as empresas que operam em mercados competitivos, onde uma previsão precisa pode influenciar as decisões estratégicas sobre a gestão de estoques e campanhas de marketing.

Em Wisconsin, embora o cenário 3 mostre um desempenho um pouco inferior em comparação com as outras regiões, ainda assim supera os cenários 1 e 2, onde os erros de previsão eram significativamente maiores. Isso sugere que apesar de haver variáveis locais que possam estar a afetar as vendas de forma menos previsível, o cenário 3 oferece uma abordagem eficaz.

A análise dos resultados, como a Tabela 10 e 11 apresentada, mostra que o cenário 3 não só proporciona a melhor combinação entre a explicação da variabilidade dos dados e a precisão das previsões, mas também se destaca como o modelo mais equilibrado entre os cenários analisados. Isso é especialmente relevante para empresas que buscam maximizar a eficiência operacional e a satisfação do cliente. A capacidade de prever com precisão as vendas em diferentes locais é essencial para otimizar as operações e adaptar estratégias às especificidades regionais, o que pode, por sua vez, levar a um melhor desempenho financeiro e a uma maior competitividade no mercado.

Tabela 10: Resultados das métricas do modelo por loja com SVM

Modelo por Loja	Métricas de Avaliação				
	R ²	MAE	RMSE	MAPE	MSE
Cenário 1					
CA_1	0,5855	499	680	∞	462.667
CA_2	0,5351	439	619	713,62	383.438
CA_3	0,4838	613	892	799,81	795.210
CA_4	0,3268	252	361	∞	130.601
TX_1	0,5405	313	440	1427,95	193.217
TX_2	0,4909	424	605	∞	366.109
TX_3	0,4382	377	525	1030,97	276.123
WI_1	0,4410	550	754	1040,63	568.369
WI_2	0,3776	768	990	∞	980.250
WI_3	0,4240	522	690	1570,49	476.725
Cenário 2					
CA_1	0,7448	300	417	6,7	174.122
CA_2	0,7517	401	531	11,77	281.810
CA_3	0,7101	353	488	5,52	238.514
CA_4	0,4913	143	177	5,86	31.449
TX_1	0,6687	205	270	6,42	73.160
TX_2	0,5623	277	353	7,27	124.533
TX_3	0,4797	236	300	6,32	89.826
WI_1	0,5980	371	509	9,92	258.919
WI_2	0,4605	542	698	12,18	486.594
WI_3	0,5039	400	513	11,21	262.837
Cenário 3					
CA_1	0,9180	192	272	3,99	74.202
CA_2	0,8818	301	388	7,12	150.352
CA_3	0,8499	326	403	5,1	162.025
CA_4	0,7417	128	154	5,21	23.746
TX_1	0,8373	208	242	6,57	58.608
TX_2	0,8415	257	306	6,4	93.934
TX_3	0,5879	261	386	6,96	148.887
WI_1	0,8493	331	426	10,26	181.375
WI_2	0,5765	528	645	10,18	416.554
WI_3	0,6612	486	573	13,45	327.778
Cenário 4					
CA_1	0,8669	316	365	6,7	133.307
CA_2	0,8679	360	420	7,4	176.715
CA_3	0,8104	430	536	5,76	287.404
CA_4	0,5846	200	219	7,59	47.952
TX_1	0,8983	136	170	3,79	28.773
TX_2	0,5195	451	503	10,95	252.587
TX_3	0,4047	335	409	8,68	167.623
WI_1	0,8112	259	377	6,06	142.021
WI_2	0,3404	584	694	10,72	481.855
WI_3	0,5127	444	508	10,07	257.957

Tabela 11: Resultados das previsões do modelo por loja com SVM

Modelo por Loja	Previsões						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Cenário 1							
CA_1	3363	3360	3360	3360	3360	4202	4468
CA_2	2446	2399	2387	2384	2384	3164	3405
CA_3	5270	5293	5299	5300	5300	6185	6471
CA_4	1953	1937	1932	1930	1930	2161	2250
TX_1	2408	2647	2670	2673	2673	3260	3576
TX_2	3428	3352	3345	3345	3345	4030	4269
TX_3	2709	2818	2840	2845	2845	3281	3455
WI_1	2383	2249	2204	2190	2190	2242	2259
WI_2	3216	2838	2700	2651	2651	2847	2927
WI_3	2349	2776	2825	2834	2834	3373	3540
Cenário 2							
CA_1	6113	4227	4007	3977	3977	4565	4678
CA_2	6082	4036	3606	3518	3518	4076	4145
CA_3	7721	6001	5505	5370	5370	6451	6826
CA_4	3271	2594	2379	2316	2316	2471	2523
TX_1	4033	3229	2911	2811	2811	3195	3317
TX_2	4292	3579	3425	3394	3394	4184	4414
TX_3	3957	3560	3454	3428	3428	3994	4176
WI_1	4874	3190	3030	3011	3011	3870	4091
WI_2	5127	4229	3933	3832	3832	4030	4103
WI_3	4325	3441	3146	3056	3056	3520	3693
Cenário 3							
CA_1	6113	4200	4010	3992	3992	4630	4718
CA_2	6082	3745	3597	3590	3590	4869	5070
CA_3	7721	6160	5696	5570	5570	6850	7292
CA_4	3271	2630	2435	2383	2383	2425	2440
TX_1	4033	3168	2989	2954	2954	3599	3777
TX_2	4292	3872	3814	3806	3806	4580	4757
TX_3	3957	3572	3298	3140	3140	3203	3203
WI_1	4874	3147	3181	3181	3181	4385	4387
WI_2	5127	4999	4944	4920	4920	5325	5518
WI_3	4325	4140	4043	3991	3991	3317	2822
Cenário 4							
CA_1	6113	4427	3917	3792	3792	5435	6016
CA_2	6082	3650	3722	3720	3720	5848	5969
CA_3	7721	5801	5539	5511	5511	7382	7723
CA_4	3271	2544	2480	2476	2476	3036	3070
TX_1	4033	3086	3082	3082	3082	4006	4050
TX_2	4292	3822	3752	3746	3746	4464	4475
TX_3	3957	3725	3701	3700	3700	4084	4117
WI_1	4874	3182	3238	3236	3236	4858	4875
WI_2	5127	5330	5334	5335	5335	5434	5445
WI_3	4325	3393	3462	3457	3457	4701	4656

Na análise do modelo SVM, o agregado por loja destaca-se como o melhor em termos de erros absolutos. Apresenta o menor MAE (301) e RMSE (379), o que indica que as previsões individuais por loja são mais precisas em termos de desvios absolutos e quadráticos. Este desempenho reflete-se também no MSE (163.746), que é significativamente inferior aos outros agregados, mostrando que o modelo por loja é mais eficaz na minimização dos erros de previsão. Em termos de MAPE, o agregado por loja também é competitivo, com 7,52%, sugerindo que as previsões feitas para cada loja individualmente apresentam um erro percentual relativamente baixo e consistente.

No entanto, o agregado por estado oferece a melhor capacidade de explicar a variabilidade dos dados, com um R^2 de 0,8067, superior ao do agregado por loja (0,7745). Isto significa que, apesar de o modelo por loja ter erros mais baixos, o modelo por estado é mais eficaz na captura das tendências regionais e das variações nas vendas entre os diferentes estados. Em suma, a escolha entre o modelo por loja ou por estado depende do objetivo da análise. Se o foco for a minimização dos erros absolutos e quadráticos, o agregado por loja é preferível. No entanto, se o objetivo for explicar melhor a variabilidade dos dados em termos regionais e obter uma visão mais ampla das dinâmicas entre estados, o agregado por estado oferece uma explicação mais robusta e equilibrada. Os resultados dos modelos agregados estão disponíveis na Tabela 12, enquanto a Figura 21 representa o gráfico de dispersão entre as vendas reais e as vendas previstas para o SVM.

Tabela 12: Comparação entre os modelos agregados com SVM (cenário 3)

Métrica	Agregado Global	Agregado por Estado	Agregado por Loja
R^2	0,7909	0,8067	0,7745
MAE	1015	971	301
RMSE	1235	1158	379
MAPE	7,95%	7,71%	7,52%
MSE	1.682.651	1.504.194	163.746
Previsão	Agregado Global	Agregado por Estado	Agregado por Loja
Dia 1	49795	49795	49795
Dia 2	39800	39892	39633
Dia 3	38202	38359	38007
Dia 4	37946	38131	37527
Dia 5	37946	38131	37527
Dia 6	49147	46072	43183
Dia 7	51784	47705	43984

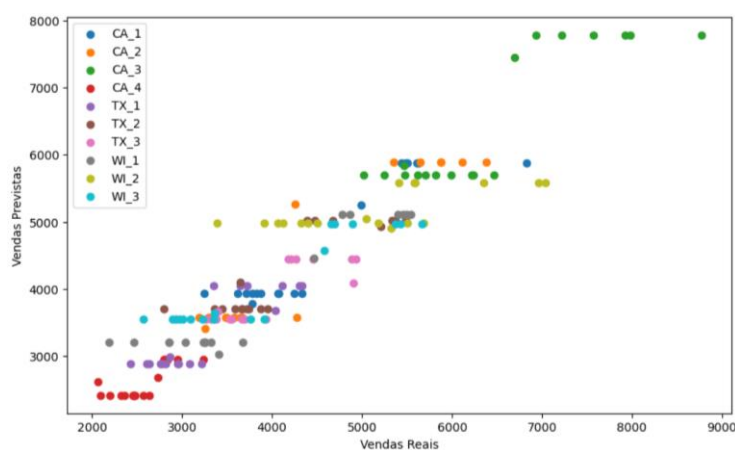


Figura 21: Gráfico de dispersão entre as vendas reais e previstas com SVM

4.2.2. Comparação entre as Abordagens de Aprendizagem Automática

A análise dos modelos de aprendizagem de máquina foi conduzida através de quatro cenários temporais distintos para cada modelo, permitindo uma compreensão abrangente da eficácia de cada abordagem. O cenário 1 consistiu no treino com todos os dados disponíveis, enquanto o cenário 2 focou nos dados do último ano, o cenário 3 nos últimos três meses e o cenário 4 apenas no último mês. Em quase todos os casos, exceto para a regressão polinomial, o cenário 3 destacou-se como o mais eficaz, apresentando as melhores métricas de desempenho. Esta constatação permitiu que esse cenário fosse escolhido como base para comparação entre os modelos de aprendizagem de máquina, fornecendo um *benchmark* claro para a avaliação da precisão preditiva. A Tabela 13 permite observar os resultados de cada modelo para o cenário 3.

Tabela 13: Comparação entre os modelos agregados de aprendizagem automática (cenário 3)

Modelo	Métricas de Avaliação				
	R ²	MAE	RMSE	MAPE	MSE
Agregado Global					
Regressão Linear Simples	0,3394	1771	2233	12,96%	5.489.701
Regressão Polinomial	0,4876	1590	1985	11,63%	4.367.694
<i>Decision Tree</i>	0,5100	1198	1588	8,80%	2.725.272
<i>XGBoost</i>	0,5699	1139	1497	8,41%	2.380.610
<i>Support Vector Machines</i>	0,7909	1015	1235	7,95%	1.682.651
Agregado por Estado					
Regressão Linear Simples	0,8162	950	1182	7,26%	1.524.447
Regressão Polinomial	0,5579	1085	1608	9,20%	2.667.587
<i>Decision Tree</i>	0,7782	1034	1292	7,92%	1.787.642
<i>XGBoost</i>	0,7675	1090	1300	8,19%	1.781.653
<i>Support Vector Machines</i>	0,8067	971	1158	7,71%	1.504.194
Agregado por Loja					
Regressão Linear Simples	0,7667	315	399	7,84%	182.559
Regressão Polinomial	0,0095	412	691	10,97%	774.531
<i>Decision Tree</i>	0,7130	354	445	8,68%	230.892
<i>XGBoost</i>	0,7096	365	452	8,85%	233.918
<i>Support Vector Machines</i>	0,7745	301	379	7,52%	163.746

Ao examinar os resultados, o modelo de *Support Vector Machines* (SVM) emergiu como o mais eficaz em termos de capacidade explicativa e precisão. No agregado global, o SVM alcançou um R² de 0,7909, um MAE de 1015 e um RMSE de 1235, o que demonstra uma forte capacidade de explicar a variabilidade dos dados com um erro relativamente baixo. No entanto, quando analisado o agregado por estado, o SVM ainda manteve uma performance robusta, apresentando um R² de 0,8067 e um MAE de 971, reafirmando a sua posição como o melhor modelo em termos de previsão. Já no agregado por loja, o SVM obteve um desempenho respeitável com um R² de 0,7745, mas aqui a Regressão Linear Simples destacou-se com um MAE de apenas 315, mostrando que, a nível local, a simplicidade pode ser uma vantagem, oferecendo previsões mais precisas em alguns contextos.

O segundo melhor modelo foi o *XGBoost*, que apresentou um desempenho consistente em todos os agregados, especialmente no agregado global, onde obteve um R^2 de 0,5699, MAE de 1139 e RMSE de 1497. Embora o *XGBoost* não tenha superado o SVM, a sua versatilidade e robustez tornam-no uma escolha valiosa, especialmente quando se busca uma abordagem que possa lidar com dados complexos e não lineares. No agregado por estado, o *XGBoost* ficou atrás do SVM, mas ainda apresentou resultados competitivos com um R^2 de 0,7675 e MAE de 1090. Já no agregado por loja, o *XGBoost* teve um desempenho ligeiramente inferior ao da Regressão Linear Simples e da Decision Tree, mas ainda assim mostrou-se eficaz com um MAE de 365.

Por outro lado, a Regressão Polinomial foi o modelo com o pior desempenho em quase todos os cenários, especialmente no agregado por loja, onde apresentou um R^2 de apenas 0,0095. Essa performance indica que a complexidade adicional da regressão polinomial não trouxe benefícios em termos de previsão, e, na verdade, pode ter contribuído para o sobreajuste aos dados. Além disso, a Decision Tree apresentou resultados medianos em comparação com os outros modelos, não se destacando em nenhum dos agregados, mas ainda oferecendo previsões aceitáveis.

Ao considerar os diferentes níveis de agregação, fica claro que o agregado por estado proporciona a melhor combinação de precisão e capacidade de explicação em comparação com o agregado global e o agregado por loja. No agregado global, a Regressão Linear Simples apresentou resultados relativamente fracos, enquanto o SVM no agregado por estado demonstrou um erro absoluto e percentual inferiores, indicando que essa abordagem regional permite capturar melhor as variações nos dados. A vantagem dos modelos globais reside na simplicidade e eficiência, mas eles tendem a diluir as características locais, resultando em previsões menos precisas. Em contraste, os modelos individuais, como o SVM, são mais eficazes para capturar as nuances de cada estado ou loja, minimizando os erros de previsão.

Em conclusão, o modelo SVM é o mais adequado para previsões a nível global e estadual, devido à sua robustez e precisão. Para análises a nível local, a Regressão Linear Simples demonstrou ser uma opção eficaz, sugerindo que, enquanto a complexidade pode trazer benefícios em alguns contextos, a simplicidade pode ser vantajosa em outros. O *XGBoost*, embora não tenha superado o SVM, é uma escolha sólida para cenários que exigem uma abordagem mais flexível e robusta. A escolha entre modelos globais e individuais deve ser baseada nos objetivos da análise, considerando as vantagens de precisão e interpretação versus a simplicidade e eficiência na implementação.

5. CONCLUSÃO

A previsão de vendas desempenha um papel crucial na gestão eficiente de negócios, influenciando diretamente a tomada de decisões estratégicas, como a gestão de stocks e a alocação de recursos. A capacidade de prever a procura futura é especialmente essencial em mercados onde a sazonalidade e as flutuações são frequentes, exigindo uma abordagem robusta para capturar as variações no comportamento dos consumidores. No entanto, prever vendas é um desafio complexo devido à interação de múltiplos fatores que podem influenciar a procura. Neste contexto, a aplicação de métodos estatísticos e de aprendizagem de máquina torna-se indispensável para melhorar a precisão das previsões e mitigar os riscos associados à gestão de stocks.

Entre os métodos estatísticos avaliados, o modelo agregado por loja com *SARIMA* destacou-se como a opção mais eficaz, apresentando um MAE de 224, um RMSE de 279 e um MAPE de 5,39%. Esses resultados evidenciam a sua capacidade de capturar as dinâmicas sazonais e tendências nos dados de vendas, tornando-o particularmente adequado para cenários onde a estrutura temporal dos dados desempenha um papel crítico. Por outro lado, no campo da aprendizagem automática, o modelo agregado por estado e loja com *Support Vector Machines (SVM)* revelaram-se os modelos mais eficazes, com um R^2 0,8067 e 0,7745, um MAE de 971 e 301, e um RMSE de 1158 e 379, respetivamente, demonstrando a versatilidade do SVM em modelar relações complexas e não lineares.

No geral, os modelos agregados por loja demonstraram um desempenho superior ao capturarem melhor as particularidades locais e oferecerem previsões mais precisas em comparação com os modelos agregados globais e por estado. Modelos individuais por estado e loja proporcionam uma personalização que permite ajustar as previsões às características específicas de cada unidade de venda, refletindo nuances regionais e locais que modelos globais frequentemente não capturam. Embora modelos globais sejam úteis para identificar tendências gerais, a granularidade dos modelos por loja melhora significativamente a precisão das previsões. Um modelo global pode captar tendências gerais, mas pode não refletir as particularidades regionais que influenciam as vendas locais. Além disso, a variação no desempenho dos modelos e a importância da inclusão de hiperparâmetros destacam a necessidade de adaptar os modelos às especificidades dos dados e ao contexto da aplicação para otimizar a eficácia das previsões.

Este estudo também demonstra a importância de testar diferentes períodos e modelos para otimizar a previsão de vendas. Em muitos casos, a utilização de dados mais recentes e em menor quantidade proporcionou melhores resultados, enfatizando a necessidade de uma abordagem adaptativa na modelação preditiva. No contexto deste trabalho, os melhores resultados foram alcançados com modelos treinados utilizando dados dos últimos três meses ou do último mês, o que ressalta a relevância de ajustar os modelos à temporalidade dos dados e ao comportamento recente do mercado.

Em conclusão, tanto os modelos estatísticos como os de aprendizagem automática têm o seu lugar na previsão de vendas, mas a escolha do modelo ideal deve ser orientada pelas características específicas dos dados e pelo contexto de aplicação. O *SARIMA* e o *SVM* destacam-se como ferramentas poderosas, cada uma com suas particularidades e adequações, assegurando uma

gestão mais eficaz e informada das operações de venda. A combinação de métodos e a consideração das dinâmicas locais são essenciais para garantir uma previsão precisa e uma tomada de decisão estratégica bem fundamentada. A preferência por modelos individuais por loja e estado, quando comparados com modelos globais, oferece uma vantagem significativa ao capturar as nuances específicas e ao ajustar-se às variações regionais, resultando em previsões mais precisas e úteis para a gestão de vendas.

5.1. Limitações e investigação futura

Para trabalhos futuros, uma área de grande potencial é o desenvolvimento e a aplicação de modelos híbridos, que combinam métodos estatísticos e de aprendizagem de máquina para aproveitar as forças de cada abordagem. Modelos híbridos podem integrar a robustez dos métodos estatísticos, como o SARIMA, na captura de padrões sazonais e tendências de longo prazo, com a flexibilidade e capacidade preditiva dos algoritmos de machine learning, como Random Forest e XGBoost. Essa integração pode resultar em uma previsão mais precisa e adaptativa, combinando a interpretação explícita dos componentes temporais com a capacidade de aprender padrões complexos dos dados.

Outra abordagem a ser investigada é a aplicação de técnicas avançadas de modelagem, como redes neurais profundas e métodos de ensemble sofisticados. Essas técnicas têm o potencial de identificar padrões complexos e não lineares que podem ser desafiantes para métodos tradicionais. A comparação de modelos de aprendizagem profunda com abordagens como *XGBoost* pode fornecer novas percepções e oportunidades para aprimorar as previsões de vendas.

Além disso, futuros estudos podem focar na otimização de hiperparâmetros e no desenvolvimento de modelos dinâmicos adaptativos. Técnicas de otimização avançada, como otimização bayesiana, podem melhorar a eficiência da configuração dos modelos, enquanto modelos adaptativos podem oferecer previsões mais ágeis em contextos de alta volatilidade. A combinação dessas abordagens pode resultar em previsões mais precisas e robustas, ajustadas às rápidas mudanças no mercado.

REFERÊNCIAS BIBLIOGRÁFICAS

Abbasi, Raza Abid, et al. "Short term load forecasting using XGBoost." *Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 33rd International Conference on Advanced Information Networking and Applications (WAINA-2019) 33*. Springer International Publishing, 2019.

Aju Mathew, Prof.E.M.Somasekaran Nair, Asst Prof. Jenson Joseph E (2013). "Demand Forecasting For Economic Order Quantity in Inventory Management". In url: [ijsrp-p2223-libre.pdf](https://www.ijsrp-p2223-libre.pdf) (d1wqtxts1xzle7.cloudfront.net)

AlmaBetterBytes. "Decision Tree in Machine Learning". (2024)

Anirudh, VK. "How The CRISP-DM Method Can Help Manage Your Next Data Science Project". Analyticsindiamag (2020)

Anusha, S. Lakshmi, Swati Alok, and Ashiff Shaik. "Demand forecasting for the Indian pharmaceutical retail: A case study." *Journal of Supply Chain Management Systems* 3.2 (2014): 1-8.

Ávila, Rafael. "Como calcular o estoque de segurança de um produto no seu controle de estoque". Blog luz. (2015)

Bayer, Débora Missio, Nilza Maria dos Reis Castro, and Fernando Mariano Bayer. "Modelagem e previsão de vazões médias mensais do rio Potiribu utilizando modelos de séries temporais." *Rbrh: revista brasileira de recursos hídricos. Porto Alegre, RS. Vol. 17, n. 2 (abr./jun. 2012), p. 229-239* (2012).

Bianco, Vincenzo, Oronzio Manca, and Sergio Nardini. "Electricity consumption forecasting in Italy using linear regression models." *Energy* 34.9 (2009): 1413-1421.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bowley, A. L. (1920). On the Methods of Measuring the Relation between Variables. *Journal of the Royal Statistical Society*, 83(1), 18-34.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.

Breiman, L., Friedman, J., Olshen, R.A., & Stone, C.J. (1984). *Classification and Regression Trees* (1st ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9781315139470>

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

Caixeta, André Fernandes, et al. "Previsão de demanda em um restaurante de pequeno porte." *Brazilian Journal of Production Engineering* 9.5 (2023): 115-124.

Canela, Miguel Ángel, et al. "Holt-winters forecasting." *Quantitative Methods for Management: A Practical Approach* (2019): 121-128.

Castanho, Leonor. "O Método ABC para a Gestão de Inventário. Outvio (2023)

- Chatfield, Chris. "The Holt-winters forecasting procedure." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 27.3 (1978): 264-279.
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).
- Chicco, Davide, Matthijs J. Warrens, and Giuseppe Jurman. "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation." *Peerj computer science* 7 (2021): e623.
- Christiansen, Ray. "Inventory Management: 20 Challenges, 20 Solutions" Solatatech (2024)
- Colin D Lewis, *Demand Forecasting and Inventory Control*, Butterworth-Heinemann, 1997, Page v, ISBN 9781855732414,
- Danielsson, Jenny. "Demand forecasting challenges – how to deal with fluctuating demand". Easystock (2023).
- Dhoka, Dinesh, and Y. Lokeswara Choudary. "ABC classification for inventory optimization." *IOSR Journal of Business and Management* 15.1 (2013): 38-41.
- Divisekara, Roshani W., G. J. M. S. R. Jayasinghe, and K. W. S. N. Kumari. "Forecasting the red lentils commodity market price using SARIMA models." *SN Business & Economics* 1.1 (2020): 20.
- dos Reis, João Gilerto Mendes, et al. "Previsão de demanda no mercado de varejo: um estudo de caso em um comércio de baterias automotivas." *South American Development Society Journal* 1.1 (2015): 1-16.
- Dudek, Grzegorz. "Short-term load forecasting using random forests." *Intelligent Systems' 2014: Proceedings of the 7th IEEE International Conference Intelligent Systems IS'2014, September 24-26, 2014, Warsaw, Poland, Volume 2: Tools, Architectures, Systems, Applications*. Springer International Publishing, 2015.
- Ediger, Volkan Ş., and Sertac Akar. "ARIMA forecasting of primary energy demand by fuel in Turkey." *Energy policy* 35.3 (2007): 1701-1708.
- Edward A. Silver, (1981) *Operations Research in Inventory Management: A Review and Critique*. *Operations Research* 29(4):628-645.
- Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246-263.
- Giannoccaro, Ilaria, and Pierpaolo Pontrandolfo. "Inventory management in supply chains: a reinforcement learning approach." *International Journal of Production Economics* 78.2 (2002): 153-161.
- Granger, C.W.J. (1993). 'Strategies for Modelling Nonlinear Time-Series Relationships'. In: *Economic Record* 69.3, pp. 233–238
- Guan, Y., & Zhao, W. (2015). A review of forecasting and its applications in supply chain management. *Journal of Industrial Engineering and Management*, 8(4), 1109-1132.

- Hamza, Muhammad, Mahdi Abolghasemi, and Abraham Oshni Alvandi. "Forecasting sales with Bayesian networks: a case study of a supermarket product in the presence of promotions." arXiv preprint arXiv:2112.08706 (2021).
- Hanggara, Fuad Dwi. "Forecasting car demand in Indonesia with moving average method." *Journal of Engineering Science and Technology Management (JES-TM)* 1.1 (2021): 1-6.
- Hansun, Seng. "A new approach of moving average method in time series analysis." *2013 conference on new media studies (CoNMedia)*. IEEE, 2013.
- Harris, F. W. (1913). *How many parts to make at once*. *Operations Research*, 38(6), 947-950
- Henning Baars, Hans-Georg Kemper, Heiner Lasi, Marc Siegel (2008). "Combining RFID Technology and Business Intelligence for Supply Chain Optimization Scenarios for Retail Logistics" In url: <https://ieeexplore.ieee.org/abstract/document/4438776>
- Hoffman, Sydney. "Demand Forecasting: Types, Techniques, and Examples". *SoftwareConnect* (2023)
- Holzinger, Andreas, et al. "Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable AI." *Machine Learning and Knowledge Extraction: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27–30, 2018, Proceedings 2*. Springer International Publishing, 2018.
- Huang, Wei, Yoshiteru Nakamori, and Shou-Yang Wang. "Forecasting stock market movement direction with support vector machine." *Computers & operations research* 32.10 (2005): 2513-2522.
- Hung, Lon-Chen, Kuo-En Fu, Wei-Zhen Chen, ShihSen Peng (2012) "An ABC Analysis Model for the Multiple Products Inventory Control". In url: [p1-libre.pdf \(d1wqtxts1xzle7.cloudfront.net\)](http://p1-libre.pdf(d1wqtxts1xzle7.cloudfront.net))
- Hyndman, R.J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. [Online] Available: <https://otexts.com/fpp3/>
- Hyun, Jin. "Pros & Cons of Qualitative vs. Quantitative Demand Forecasting Methods". *ZIPforecasting* (2020)
- Jenkins, Abby. "Demand Forecasting: Everything You Need to Know" Oracle Net suite (2022).
- Johnston, F. R., et al. "Some properties of a simple moving average when applied to forecasting a time series." *Journal of the Operational Research Society* 50.12 (1999): 1267-1271.
- Karim, SA Abdul, and Saiful Azli Alwi. "Electricity load forecasting in UTP using moving averages and exponential smoothing techniques." *Appl. Math. Sci* 7.77-80 (2013): 4003-4014.
- Khan, Shakir, and Hela Alghulaiakh. "ARIMA model for accurate time series stocks forecasting." *International Journal of Advanced Computer Science and Applications* 11.7 (2020).
- Kumar, Rajesh. "Decision tree for the weather forecasting." *International Journal of Computer Applications* 76.2 (2013): 31-34.
- Kumar, V., & Reinartz, W.J. (2016). An empirical comparison of global and local forecasting methods in retail. *Journal of Retailing*, 92(2), 209-222. doi:10.1016/j.jretai.2015.12.002

- Kurani, Akshit, et al. "A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting." *Annals of Data Science* 10.1 (2023): 183-208.
- Liu, Chang, et al. "Forecasting copper prices by decision tree learning." *Resources Policy* 52 (2017): 427-434.
- LÜDKE, Menga; André, Marli D. A. A Pesquisa em educação: abordagens qualitativas. São Paulo: EPU, 1999.
- Martinello, Larissa Maria, et al. "Estudo comparativo entre os modelos de previsão ARIMA e ETS para dados temporais da produção de leite no Brasil." *Revista do Instituto de Laticínios Cândido Tostes* 76.1 (2021): 12-27.
- Maybach, Victoria. "La historia de Walmart y el diseño de su logotipo". Turbologo (2021).
- Mei, Jie, et al. "A random forest method for real-time price forecasting in New York electricity market." *2014 IEEE PES general meeting | conference & exposition*. IEEE, 2014.
- Minhaz, Muntasir. "Quantidade de pedido econômico". iEduNote (2024).
- Mohanty, A. (2019). 'Multilayer Percetptron (MLP) Models on Real World Banking Data'. In: url: <https://becominghuman.ai/multi-layer-perceptron-mlp-models-onreal-world-banking-data-f6dd3d7e998f>.
- Mondal, Prapanna, Labani Shit, and Saptarsi Goswami. "Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices." *International Journal of Computer Science, Engineering and Applications* 4.2 (2014): 13.
- Nau, Robert. "Forecasting with moving averages". Fuqua School of Business, Duke University (2014)
- Nalini. "Demand Forecasting: Definition, Factors, and Techniques!" Deskera (2024).
- Nenni, Maria Elena, Luca Giustiniano, and Luca Pirolo. "Demand forecasting in the fashion industry: a review." *International Journal of Engineering Business Management* 5 (2013): 37.
- N. Li, F. T. S. Chan, S. H. Chung and B. Niu, "The impact of non-stationary demand and forecasting on a failure-prone manufacturing system," 2015 International Conference on Industrial Engineering and Operations Management (IEOM), Dubai, United Arab Emirates, 2015, pp. 1-7, doi: 10.1109/IEOM.2015.7093707.
- Olah, C. (2015). 'Understanding LSTM networks'. In: url: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>.
- Oliveira, Bruno. "Características das séries temporais". Statplace (2019)
- Olson, David L., and Meiping Xie. "A comparison of coordinated supply chain inventory management systems." *International Journal of Services and Operations Management* 6.1 (2010): 73-88.
- Panda, Mrutyunjaya. "Application of ARIMA and Holt-Winters forecasting model to predict the spreading of COVID-19 for India and its states." *medRxiv* (2020): 2020-07.
- Panigrahi, Sibarama, et al. "Forecasting of sunspot time series using a hybridization of ARIMA, ETS and SVM methods." *Solar Physics* 296.1 (2021): 6.

- Pawan Kumar Singh, adunath Gupta, Nilpa Jha, and Aruna Rajan. 2019. Fashion Retail: Forecasting Demand for New Items. In KDD 2019 Workshop: AI for fashion The fourth international workshop on fashion and KDD, August 2019, Anchorage, Alaska - USA., 10 pages.
- P. Paama, R. Berrettaa, M. Heydara, R. García-Flores (2019). "The impact of inventory management on economic and environmental sustainability in the apple industry". In url: <https://www.sciencedirect.com/science/article/pii/S0168169918315060>
- Prabhakaran, Selva (2022). "ARIMA Model – Complete Guide to Time Series Forecasting in Python". Machinelearningplus.
- Priyadarshi, Rahul, et al. "Demand forecasting at retail stage for selected vegetables: a performance analysis." *Journal of Modelling in Management* 14.4 (2019): 1042-1063.
- Rahman, Md Siddikur, Arman Hossain Chowdhury, and Miftahuzzannat Amrin. "Accuracy comparison of ARIMA and XGBoost forecasting models in predicting the incidence of COVID-19 in Bangladesh." *PLOS Global Public Health* 2.5 (2022): e0000495.
- Rangel Logistics Solutions. "Gestão de inventário: as 7 melhores práticas a implementar" (2023).
- Rosa, H., Mayerle, S. F., & Gonçalves, M. B.. (2010). Controlo de stock por revisão contínua e revisão periódica: uma análise comparativa utilizando simulação. *Production*, 20(4), 626–638.
- Sarac, Aysegul, Nabil Absi, and Stéphane Dauzère-Pérès. "A literature review on the impact of RFID technologies on supply chain management." *International journal of production economics* 128.1 (2010): 77-95.
- Segal, Troy. "Make to Stock (MTS): Definition, Example, and How it works". Investopedia (2022)
- Silva, Aryelle Monteiro Santos, Bruna Eduarda Nascimento, and Janicleide Alves Ferreira. "Importância de controle de estoque: estudo realizado em uma microempresa de moda praia da cidade de Maceió-AL." (2023).
- Silva, Adilane Ribeiro (2021). "Introdução às Séries Temporais" Statplace.
- SILVA, G. L. F. KNECHTEL, Maria do Rosário. Metodologia da pesquisa em educação: uma abordagem teórico-prática dialogada. Curitiba: Intersaberes, 2014. *Práxis Educativa*, [S. l.], v. 11, n. 2, p. 531–534, 2017. DOI: 10.5212/PraxEduc.v.11i2.0013. Disponível em: <https://revistas.uepg.br/index.php/praxiseducativa/article/view/8846>. Acesso em: 2 fev. 2024.
- Silva, Adilane Ribeiro. "Introdução às Séries Temporais". Statplace (2021)
- Sinaga, H., and N. Irawati. "A medical disposable supply demand forecasting by moving average and exponential smoothing method." *Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia*. 2020.
- Singh, Deepesh, and Ajay Verma. "Inventory management in supply chain." *Materials Today: Proceedings* 5.2 (2018): 3867-3872.
- Singh, Pawan Kumar, et al. "Fashion retail: Forecasting demand for new items." *arXiv preprint arXiv:1907.01960* (2019).
- Sharda, R. (1994). *Neural networks for the MS/OR analyst: An application bibliography*

- Shukla, Manish, and Sanjay Jharkharia. "ARIMA models to forecast demand in fresh supply chains." *International Journal of Operational Research* 11.1 (2011): 1-18.
- Souza, Daiane. "Stock mínimo e máximo: o que é, diferenças, como calcular e exemplos" (ca. 2020)
- Soylev, Beytullah. "What is SVM Classification?". Medium, 2023.
- Stergiou, K. I., 1991. Short-term fisheries forecasting: comparison of smoothing, ARIMA and regression techniques. *Journal of Applied Ichthyology* 7(4), 193-204.
- Stoll, Franz, "A Comparison of Machine Learning and Traditional Demand Forecasting Methods" (2020). All Theses. 3367.
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.
- Swami, Devendra, Alay Dilipbhai Shah, and Subhrajeet KB Ray. "Predicting future sales of retail products using machine learning." *arXiv preprint arXiv:2008.07779* (2020).
- Teunter, Ruud H., M. Zied Babai, and Aris A. Syntetos. "ABC classification: service levels and inventory costs." *Production and Operations Management* 19.3 (2010): 343-352.
- Timonen, Jonas. "Demand Forecasting: Importance, Benefits and Driving Factors". Sniffie (2023).
- Thomas, Guilherme Massotii. "Previsão de demanda aplicada a uma indústria de rolhas." (2019).
- Thomassey, Sébastien, and Antonio Fiordaliso. "A hybrid sales forecasting system based on clustering and decision trees." *Decision Support Systems* 42.1 (2006): 408-421.
- Trapero, J. R., Pedregal, D. J., Fildes, R., Kourentzes, N., 2013. Analysis of judgmental adjustments in the presence of promotions. *International Journal of Forecasting* 29(2), 234-243.
- Twin, Alexandra. "Delphi Method Forecasting: Definition and How It's Used". Investopedia (2023)
- Tyralis, Hristos, and Georgia Papacharalampous. "Variable selection in time series forecasting using random forests." *Algorithms* 10.4 (2017): 114.
- Vashishtha, Rajesh Kumar, et al. "Product age based demand forecast model for fashion retail." *arXiv preprint arXiv:2007.05278* (2020).
- Vásquez, Jeison, Piero Ojeda, and Lenis Wong. "Model to Predict Incoming Tech Support Demand in a Banking Company Applying CRISP-DM and Machine Learning." 2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2024.
- Wacker, J. G., Lummus, R. R., 2002. Sales forecasting for strategic resource planning. *International Journal of Operations & Production Management* 22(9), 1014-1031.
- Walliman, N. (2011). *Research Methods - The Basics*.
- Walmart. "About Walmart." Walmart Corporate, Walmart, www.corporate.walmart.com/about. Accessed 14 Aug. 2024.
- Walter, Olga Maria Formigoni Carvalho, et al. "Aplicação de um modelo SARIMA na previsão de vendas de motocicletas." *Exacta* 11.1 (2013): 77-88.
- Wang, C.-C.; Kuo, P.-H.; Chen, G.-Y. Machine Learning Prediction of Turning Precision Using Optimized XGBoost Model. *Appl. Sci.* 2022, 12, 7739. <https://doi.org/10.3390/app12157739>

Wei, W. W. S. (2006). *Time Series Analysis: Univariate and Multivariate Methods* (2^a ed.). Pearson Education.

Wi, Young-Min, Sung-Kwan Joo, and Kyung-Bin Song. "Holiday load forecasting using fuzzy polynomial regression with weather feature selection and adjustment." *IEEE Transactions on Power Systems* 27.2 (2011): 596-603.

Wibisana, Bayu, et al. "Forecasting of Rice Production Using a Linear Regression and Polynomial Regression." *Journal of Informatics and Communication Technology (JICT)* 6.1 (2024): 33-41.

Wu, Don Chi Wai, et al. "Forecasting tourist daily arrivals with a hybrid Sarima–Lstm approach." *Journal of hospitality & tourism research* 45.1 (2021): 52-67.

Zöller, Marc-André, Marius Lindauer, and Marco F. Huber. "auto-sktime: Automated Time Series Forecasting." *arXiv preprint arXiv:2312.08528* (2023).

Zwaida, Tarek Abu (2021). "Optimization of Inventory Management to Prevent Drug Shortages in the Hospital Supply Chain" In url: <https://www.mdpi.com/2076-3417/11/6/2726>

ANEXO A

Complemento aos Resultados e Discussão

Neste anexo, é apresentada uma expansão do capítulo 4, onde os resultados e discussões são descritos. Com o objetivo de oferecer uma visão mais completa e transparente das análises realizadas, este suplemento fornece tabelas e gráficos adicionais que abrangem os resultados específicos dos modelos estatísticos e dos modelos de aprendizagem automática que não foram explicados em detalhes no corpo do documento. As seções A.1 a A.5 apresentam os resultados dos modelos estatísticos utilizados neste trabalho, enquanto as seções A.6 a A.10 fornecem os resultados dos modelos de *Machine Learning*.

A.1 Média Móvel Simples

No âmbito da previsão de vendas para os próximos 7 dias, foi aplicada a técnica de Média Móvel Simples (MMS), uma abordagem clássica em séries temporais que se baseia na média das vendas dos últimos dias para fazer previsões. Esta técnica é particularmente apreciada pela sua simplicidade e facilidade de interpretação.

As previsões geradas pela média móvel simples (MMS) para os próximos dias apresentaram valores variados, conforme o modelo global e os modelos segmentados por estado e loja. O modelo global fez a previsão das vendas dos próximos 7 dias utilizando os dados de vendas disponíveis das lojas e dos estados de forma agregada, resultando em previsões que variaram de 40.754 a 43.368 unidades. Os resultados do backtesting mostraram que o modelo teve um Erro Absoluto Médio (MAE) de 3.618 unidades, uma Raiz do Erro Quadrático Médio (RMSE) de 5.228 unidades, e um Erro Percentual Absoluto Médio (MAPE) de 7,91%, indicando uma margem de erro relativamente alta. Estas métricas evidenciam que, embora o modelo global ofereça uma visão geral das tendências de vendas, há uma necessidade de ajustes para melhorar a precisão das previsões.

Os resultados das previsões por estado utilizando a média móvel simples mostram variações na precisão do modelo. No estado da Califórnia (CA), as previsões apresentaram um Erro Absoluto Médio (MAE) de 2016 e um Erro Percentual Absoluto Médio (MAPE) de 10,02%, indicando uma maior divergência entre as previsões e os valores reais. No Texas (TX), o modelo foi mais preciso, com um MAE de 754 e um MAPE de 6,61%. Já no Wisconsin (WI), o desempenho foi intermediário, com um MAE de 1014 e um MAPE de 7,31%. As previsões para os próximos 7 dias indicam uma tendência de crescimento gradual nas vendas em todos os estados analisados, sendo o desempenho no Texas o mais próximo dos valores reais.

Os resultados das previsões por loja utilizando a média móvel simples mostram variações de precisão entre as diferentes lojas. Na loja CA_1, o modelo apresentou um Erro Absoluto Médio (MAE) alto de 571 e um Erro Percentual Absoluto Médio (MAPE) alto de 11,02%, enquanto a loja CA_4 teve o melhor desempenho, com um MAE de 170 e um MAPE de 5,85%. As lojas do Texas (TX_1 e TX_2) mostraram desempenho moderado, com MAEs de 283 e 277, respetivamente. As lojas do Wisconsin também apresentaram variações, com a loja WI_1 registrando um alto erro

percentual (MAPE de 10%), enquanto WI_2 teve um MAPE de 5,49%. As previsões para os próximos 7 dias indicam crescimento estável em várias lojas, com tendência de aumento gradual nas vendas. Os resultados das métricas do modelo de Média Móvel Simples estão disponíveis na Tabela A.1, enquanto os valores das previsões para os próximos dias podem ser visualizados na Tabela A.2.

Tabela A. 1: Resultados das métricas com Média Móvel Simples

Estado/loja	Métricas de Avaliação			
	MAE	RMSE	MAPE	MSE
Modelo por estado				
Califórnia	2016	2857	10,02	8.167.547
Texas	754	966	6,61	934.548
Wisconsin	1014	1505	7,31	2.265.566
Modelo por Loja				
CA_1	571	810	11,02	657.315
CA_2	647	957	12,17	917.310
CA_3	627	890	9,29	792.712
CA_4	170	238	5,85	56.967
TX_1	283	377	8,20	142.577
TX_2	277	382	6,66	146.014
TX_3	230	283	5,96	80.118
WI_1	445	648	10,00	420.957
WI_2	278	389	5,49	151.838
WI_3	350	491	8,45	241.491
Modelo Global	3618	5228	7,91	2.733.7569

Tabela A. 2: Resultados das previsões com Média Móvel Simples

Estado/loja	Previsão das Vendas						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Califórnia	17798	17904	18207	18645	19238	19545	19218
CA_1	4521	4545	4653	4786	4940	5019	4940
CA_2	4411	4494	4617	4749	4956	5027	4905
CA_3	6110	6115	6177	6311	6496	6621	6507
CA_4	2757	2749	2759	2796	2845	2876	2864
Texas	10736	10860	10851	11030	11234	11323	11188
TX_1	3300	3284	3287	3343	3424	3481	3450
TX_2	3792	3860	3880	3950	4022	4040	3977
TX_3	3644	3715	3683	3736	3787	3801	3760
Wisconsin	12219	12373	12589	12874	13150	13199	12961
WI_1	3787	3865	3979	4084	4193	4222	4143
WI_2	4702	4753	4794	4880	4949	4944	4878
WI_3	3730	3755	3815	3909	4007	4032	3939
Total	40754	41139	41649	42549	43623	44067	43368

A análise das previsões revelou que, apesar das limitações inerentes ao modelo de média móvel simples, ele ainda oferece uma base útil para estimar tendências gerais. A capacidade da MMS para suavizar flutuações e identificar padrões subjacentes foi observada nas previsões, especialmente

quando aplicada a dados agregados por estado e loja. No entanto, para melhorar a precisão das previsões e captar variações mais subtis, pode ser necessário considerar métodos mais sofisticados ou combinar a MMS com técnicas adicionais de análise de séries temporais. Assim, a Média Móvel Simples serve como um ponto de partida valioso, mas o refinamento contínuo dos modelos de previsão é crucial para otimizar a precisão das estimativas e melhorar a tomada de decisão estratégica. A Figura A.1 compara as previsões para o período de teste com as vendas totais reais utilizando o modelo MMS.

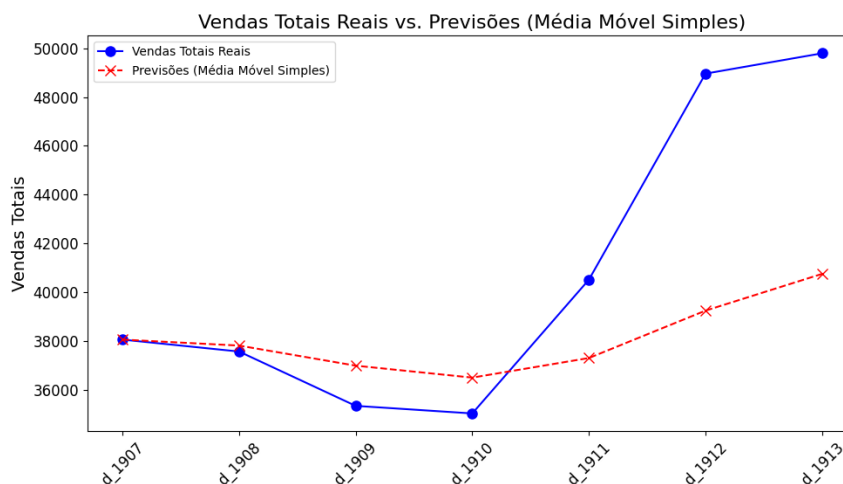


Figura A. 1: Gráfico de comparação entre as vendas reais e as previsões utilizando MMS

No modelo de previsão utilizando a média móvel simples, foram realizados três tipos de agregados: global, por estado e por loja, cada um com uma metodologia diferente de cálculo das métricas. No agregado global, as previsões e métricas foram obtidas utilizando um único modelo agregado para toda a série temporal, o que resultou num MAE de 3618, RMSE de 5228, e MAPE de 7,91%. No agregado por estado, foi treinado um modelo separado para cada estado, e para os valores de MAE, RMSE, MSE e MAPE foi calculada a média dos erros dos três estados, resultando num MAE de 1261. Por fim, no agregado por loja, cada loja individual também teve o seu próprio modelo, e, novamente, foi feita a média das métricas de erro, atingindo um MAE de 387.

Tabela A. 3: Comparação entre os modelos agregados com MMS

Métrica	Agregado Global	Agregado por Estado	Agregado por Loja
MAE	3618	1261	387
RMSE	5228	1776	546
MAPE	7,91%	7,98%	8,31%
MSE	27.337.569	3.789.220	360.729
Previsão	Agregado Global	Agregado por Estado	Agregado por Loja
Dia 1	40754	40753	40754
Dia 2	41139	41137	41135
Dia 3	41649	41647	41644
Dia 4	42549	42549	42544
Dia 5	43623	43622	43619
Dia 6	44067	44067	44063
Dia 7	43368	43367	43363

Ao analisar os resultados, verifica-se que o agregado por loja apresentou os melhores valores de erro, com o MAE mais baixo (387) e o menor MSE (360.729), indicando ser o mais eficiente. Embora o agregado global e o por estado tenham mostrado erros absolutos maiores, o MAPE foi ligeiramente inferior. A previsão para os próximos 7 dias foi consistente entre todos os agregados, mas o agregado por loja destacou-se pelo melhor desempenho geral em termos de erro. Isto deve-se à capacidade desta abordagem de adaptar-se às características específicas de cada loja, ajustando os parâmetros do modelo para capturar melhor os padrões e variações locais. Esta personalização permite fornecer previsões mais precisas e alinhadas com a realidade observada em cada loja, resultando numa redução significativa dos erros de previsão. Os resultados dos modelos podem ser consultados na Tabela A.3.

A.2 Média Móvel Exponencial

A aplicação da Média Móvel Exponencial (MME) no contexto deste estudo revelou-se uma ferramenta poderosa para a previsão de vendas, permitindo uma análise detalhada das tendências recentes. A MME foi utilizada com uma janela de sete dias, o que permitiu captar com maior precisão as variações diárias, ao atribuir maior peso aos dados mais recentes.

Utilizando o modelo global de média móvel exponencial, as previsões para o período de teste ficaram um pouco distantes dos valores reais de vendas, com o MAE em 3184 e o RMSE em 4395. O MAPE foi de 7,09%, indicando uma precisão moderada das previsões. Para os próximos 7 dias, o modelo projetou um aumento contínuo nas vendas, com valores previstos entre 44775 e 47709.

Tabela A. 4: Resultados das métricas com MME

Estado/loja	Métricas de Avaliação			
	MAE	RMSE	MAPE	MSE
Modelo por estado				
Califórnia	1839	2410	9,47	5.808.322
Texas	700	878	6,14	770.889
Wisconsin	869	1262	6,33	1.594.787
Modelo por Loja				
CA_1	530	681	10,75	463.960
CA_2	569	803	11,02	646.077
CA_3	578	768	8,86	591.116
CA_4	164	211	5,81	44.912
TX_1	277	340	8,27	116.048
TX_2	260	355	6,29	126.269
TX_3	237	322	6,06	104.293
WI_1	380	542	8,61	294.781
WI_2	246	338	4,88	114.473
WI_3	317	416	7,88	173.248
Modelo Global	3184	4395	7,08	19316030

As previsões utilizando o modelo de média móvel exponencial para os estados revelaram diferentes níveis de desempenho. Na Califórnia (CA), o modelo teve um MAE de 1839 e um RMSE

de 2410, com previsões futuras variando de 20014 a 21548. No Texas (TX), o modelo apresentou um MAE mais baixo de 700 e um RMSE de 878, com previsões para os próximos 7 dias entre 11441 e 12045. Em Wisconsin (WI), o modelo teve um MAE de 869 e um RMSE de 1262, com previsões futuras de 13319 a 14116. Em geral, o modelo mostrou capacidade de ajuste às vendas regionais e de previsão eficaz para cada estado. A Tabela A.4 apresenta os resultados das métricas encontrados com a aplicação do modelo de MME, enquanto a Tabela A.5 apresenta os resultados das previsões para os próximos 7 dias.

Tabela A. 5: Resultados das previsões com MME

Estado/loja	Previsão das Vendas						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Califórnia	20014	20270	20525	20781	21037	21292	21548
CA_1	5169	5243	5316	5390	5464	5537	5611
CA_2	5163	5257	5350	5444	5538	5632	5726
CA_3	6748	6816	6885	6954	7023	7092	7161
CA_4	2933	2953	2972	2991	3010	3030	3049
Texas	11441	11542	11642	11743	11844	11944	12045
TX_1	3553	3579	3605	3631	3656	3682	3708
TX_2	4067	4108	4148	4189	4230	4271	4312
TX_3	3821	3855	3888	3922	3956	3990	4023
Wisconsin	13319	13452	13585	13717	13850	13983	14116
WI_1	4312	4380	4449	4517	4585	4653	4721
WI_2	4960	4992	5025	5057	5090	5122	5155
WI_3	4046	4079	4111	4143	4175	4207	4239
Total	44775	45265	45753	46243	46731	47220	47709

As previsões de vendas por loja, utilizando a média móvel exponencial, mostraram uma variação significativa nos resultados. Entre os desempenhos mais destacados, a loja CA_4 sobressaiu com o menor erro absoluto médio (MAE) de 164 e o menor erro quadrático médio (MSE) de 44.912, evidenciando uma alta precisão nas previsões. Outra performance notável foi a da loja TX_3, com um MAE de 237 e um RMSE de 322, refletindo previsões bastante precisas. Em contrapartida, os piores resultados foram observados na loja CA_2, que apresentou um MAE elevado de 569 e o MSE mais alto de 646.077, sugerindo previsões menos precisas. A loja CA_3 também teve um desempenho abaixo da média, com um MAE de 578 e um MSE de 591.116, indicando dificuldades em captar com precisão as variações nas vendas.

Em termos de desempenho global do modelo de média móvel exponencial, o agregado por loja revelou-se o mais eficaz, com um MAE de 355, RMSE de 477 e MSE de 267.517, superando o agregado por estado (MAE de 1136 e RMSE de 1516) e o agregado global (MAE de 3184 e RMSE de 4395). Apesar das previsões para os próximos 7 dias terem sido praticamente idênticas entre os três agregados, o modelo por loja apresentou o menor erro, destacando-se como a abordagem mais eficiente para prever vendas com maior precisão. Os resultados do modelo agregado global, por estado e por loja para a média móvel exponencial podem ser consultados na Tabela A.6.

Tabela A. 6: Comparação entre os modelos agregados com MME

Métrica	Agregado Global	Agregado por Estado	Agregado por Loja
MAE	3184	1136	355
RMSE	4395	1516	477
MAPE	7,08%	7,31%	7,84%
MSE	19.316.030	2.724.666	267.517
Previsão	Agregado Global	Agregado por Estado	Agregado por Loja
Dia 1	44775	44774	44772
Dia 2	45265	45264	45262
Dia 3	45753	45752	45749
Dia 4	46243	46241	46238
Dia 5	46731	46731	46727
Dia 6	47220	47219	47216
Dia 7	47709	47709	47705

Em resumo, a Média Móvel Exponencial (MME) revelou-se uma ferramenta superior em comparação com a Média Móvel Simples (MMS) para a previsão de vendas. Enquanto a MMS forneceu previsões menos estáveis e mais vulneráveis a flutuações abruptas, a MME conseguiu capturar de forma mais eficaz as tendências e variações sazonais, ajustando-se melhor às mudanças nos padrões de vendas. A MME demonstrou menores erros médios absolutos (MAE) e raízes dos erros quadráticos médios (RMSE), refletindo uma maior precisão nas previsões. Em contraste, a MMS, com a sua abordagem mais simplista, enfrentou dificuldades para lidar com a complexidade e volatilidade dos dados, resultando em previsões menos fiáveis e com maior margem de erro. O gráfico das vendas totais e das previsões obtidas com a MME para o período de teste pode ser visualizado na Figura A.2.

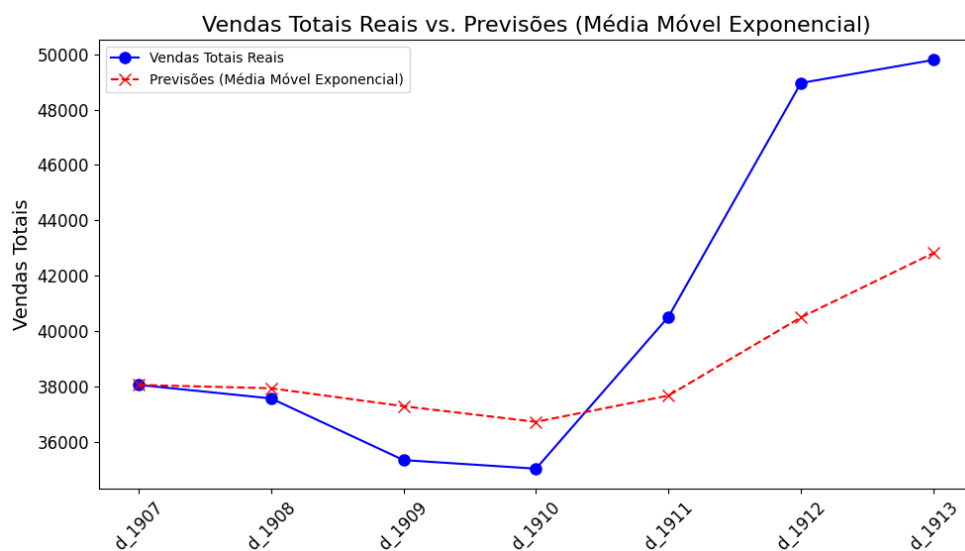


Figura A. 2: Gráfico de comparação entre as vendas reais e as previsões utilizando MME

A.3 ARIMA

No estudo, o modelo ARIMA foi crucial para a análise e previsão de vendas, utilizando uma janela móvel de 7 dias para ajustar as previsões de acordo com os padrões recentes. Foram testadas várias combinações de parâmetros para identificar a configuração ideal, garantindo previsões precisas e adaptadas às variações observadas. Os resultados das métricas para cada modelo estão disponíveis na Tabela A.7, e os resultados das previsões podem ser visualizados na Tabela A.8.

O modelo global ARIMA, otimizado com a combinação de parâmetros (3, 1, 3), demonstrou um bom desempenho nas previsões. O Erro Absoluto Médio (MAE) foi de aproximadamente 2054, o Erro Quadrático Médio (MSE) foi cerca de 8,59 milhões, e a Raiz do Erro Quadrático Médio (RMSE) foi de aproximadamente 2930. O Erro Percentual Absoluto Médio (MAPE) foi de 5,28%, refletindo uma margem de erro relativamente baixa nas previsões. Para os próximos 7 dias, o modelo previu vendas variando de 44.183 a 48.446 unidades, evidenciando a capacidade do modelo em fornecer estimativas precisas dos valores reais esperados.

Tabela A. 7: Resultados das métricas com ARIMA

Estado/loja	Métricas de Avaliação				Parâmetro
	MAE	RMSE	MAPE	MSE	
Modelo por estado					
Califórnia	1217	1463	6,67	2.141.115	(5, 1, 0)
Texas	719	925	6,88	856.288	(3, 1, 3)
Wisconsin	737	972	6,45	946.304	(3, 1, 3)
Modelo por Loja					
CA_1	258	367	5,24	135.351	(5, 1, 0)
CA_2	451	508	10,32	258.909	(3, 1, 3)
CA_3	553	633	9,45	400.713	(3, 1, 3)
CA_4	142	187	4,91	35.139	(3, 1, 3)
TX_1	155	178	4,83	32.007	(3, 1, 3)
TX_2	284	413	7,83	170.605	(3, 1, 3)
TX_3	334	423	9,91	179.367	(0, 1, 1)
WI_1	359	432	10,12	187.166	(5, 1, 0)
WI_2	643	752	14,47	566.749	(1, 0, 1)
WI_3	265	296	7,35	87.660	(3, 1, 3)
Modelo Global	2054	2930	5,28	8.589.153	(3, 1, 3)

A análise dos modelos ARIMA ajustados por estado mostra que cada um teve desempenho distinto. Para a Califórnia (CA), o modelo com parâmetros (5, 1, 0) apresentou um MAE menor em comparação aos outros parâmetros de 1217 e um MAPE de 6,67%. As previsões indicam uma tendência de crescimento nas vendas, variando de aproximadamente 17.198 a 21.257. Embora o modelo capture a dinâmica das vendas, o erro foi maior em comparação com os outros estados.

No Texas (TX) e Wisconsin (WI), ambos os modelos utilizaram a combinação (3, 1, 3), com o Texas apresentando um MAE de 719 e um MAPE de 6,88%, e Wisconsin mostrando um MAE de 737 e um MAPE de 6,45%. As previsões para o Texas foram mais estáveis e precisas, variando de 10.259 a 12.590, enquanto Wisconsin apresentou uma tendência de aumento nas vendas, com valores

projetados entre 11.207 e 14.437. Em geral, os modelos demonstraram boa capacidade de adaptação às variações regionais, com o modelo do Texas destacando-se pela maior precisão.

Tabela A. 8: Resultados das previsões com ARIMA

Estado/loja	Previsão das Vendas						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Califórnia	19797	17190	15511	15549	18269	21331	21257
CA_1	5101	4281	3849	3917	4621	5540	5567
CA_2	4634	3709	3340	3812	4768	5488	5431
CA_3	6756	5783	5237	5481	6321	7122	7284
CA_4	2832	2542	2388	2455	2690	2914	2960
Texas	11590	10258	9530	9981	11274	12434	12590
TX_1	3543	3092	2834	2950	3352	3736	3814
TX_2	4086	3544	3286	3507	4040	4484	4504
TX_3	3822	3822	3822	3822	3822	3822	3822
Wisconsin	12747	11207	10910	12096	13880	14920	14436
WI_1	3957	3512	3418	3624	4128	4612	4377
WI_2	4966	4898	4834	4773	4714	4658	4604
WI_3	3703	3217	3236	3745	4360	4619	4329
Total	44183	38005	35170	37620	43458	48277	48445

Os modelos ARIMA ajustados para lojas individuais mostram uma variabilidade significativa na precisão das previsões e na configuração dos parâmetros. Nas lojas da Califórnia, as previsões tendem a ser menos precisas do que em outros estados. A loja CA_4 destacou-se com um MAE de 142 e um MAPE de 4,91%, com previsões variando entre 2.388 e 2.960 unidades. Em contraste, as lojas CA_2 e CA_3 apresentaram MAEs e MAPEs mais elevados, indicando menor precisão nas previsões. As vendas projetadas para CA_2 variam de 3.709 a 5.489 unidades, enquanto CA_3 tem uma faixa de 5.238 a 7.284 unidades, refletindo uma complexidade maior na adaptação às variações de vendas.

Nos estados do TX e WI, o modelo ARIMA também demonstrou variações na performance. TX_1 obteve o menor MAE de 155 e um MAPE de 4,83%, com previsões entre 3.092 e 3.814 unidades. Em contraste, TX_2 e TX_3 apresentaram erros maiores, com MAEs de 284 e 334 e MAPEs de 7,83% e 9,92%, respectivamente. Em Wisconsin, WI_1 teve um MAE de 359 e um MAPE de 10,12%, com previsões variando entre 3.512 e 4.612 unidades. As previsões para WI_2 foram menos precisas, com um MAE de 643 e um MAPE de 14,47%, enquanto WI_3 apresentou um MAE de 265 e um MAPE de 7,35%, evidenciando diferenças notáveis na capacidade de previsão entre as lojas.

A análise dos modelos ARIMA revelou uma capacidade robusta de previsão tanto a nível global como regional. O modelo global, configurado com os parâmetros (3, 1, 3), mostrou um desempenho sólido, com um MAE de aproximadamente 2054 e um MAPE de 5,28%, indicando previsões precisas para as vendas futuras. Regionalmente, os modelos ajustados para estados e lojas individuais apresentaram variações significativas em termos de precisão e desempenho. Enquanto o Texas se destacou com um MAE mais baixo e previsões mais estáveis, a Califórnia e Wisconsin exibiram uma maior variabilidade nos erros. A análise detalhada por loja revelou ainda mais discrepâncias, com algumas lojas apresentando alta precisão e outras exibindo maiores erros, sublinhando a importância de ajustes finos para capturar as particularidades regionais e locais. Estes resultados

demonstram que, apesar da robustez do modelo global, a personalização para contextos regionais e locais é essencial para alcançar previsões mais precisas e relevantes.

Tabela A. 9: Comparação entre os modelos agregados com ARIMA

Métrica	Agregado Global	Agregado por Estado	Agregado por Loja
MAE	2054	891	344
RMSE	2930	1120	418
MAPE	5,28%	6,67%	8,44%
MSE	8.589.153	1.314.569	205.366
Previsão	Agregado Global	Agregado por Estado	Agregado por Loja
Dia 1	44183	44134	43400
Dia 2	38005	38655	38400
Dia 3	35170	35951	36244
Dia 4	37620	37626	38086
Dia 5	43458	43423	42816
Dia 6	48277	48685	46995
Dia 7	48445	48283	46692

Nos resultados do modelo ARIMA, o agregado por loja obteve o melhor desempenho com MAE de 344, RMSE de 418 e MSE de 205.366, superando o agregado por estado, que apresentou um MAE de 891 e RMSE de 1120, e o agregado global, com MAE de 2054 e RMSE de 2930. As previsões para os próximos 7 dias mostraram algumas variações entre os três agregados, conforme representado na Tabela A.9. A Figura A.3 representa as vendas totais reais e as previsões do modelo ARIMA para os 7 dias do período de teste.

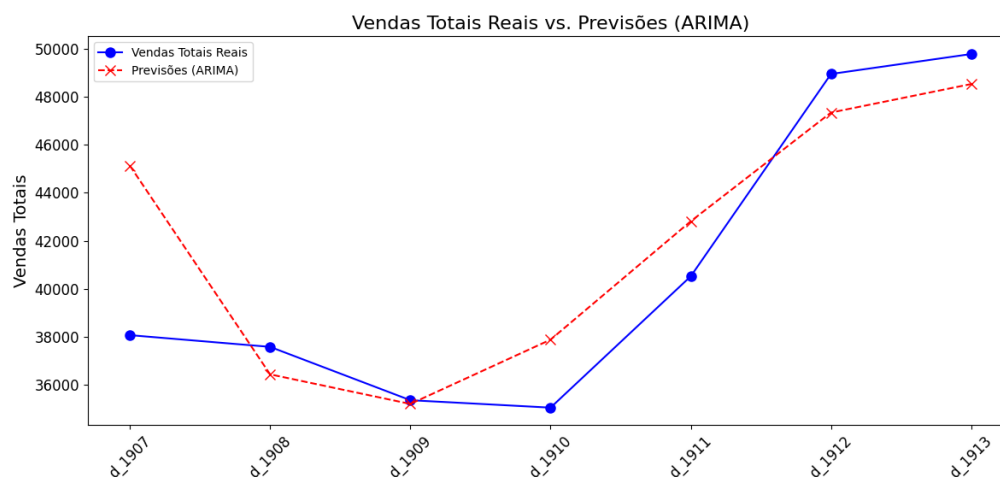


Figura A. 3: Gráfico de comparação entre as vendas reais e as previsões utilizando ARIMA

A.4 Holt-Winters

No âmbito da análise de vendas, foi aplicado o modelo *Holt-Winters*, uma técnica de suavização exponencial que integra componentes aditivos de tendência e sazonalidade. Este modelo é especialmente adequado para séries temporais com padrões sazonais regulares e tendências

persistentes. No estudo, a metodologia *Holt-Winters* foi utilizada para prever as vendas diárias e avaliar a precisão das previsões através de backtesting.

O modelo global *Holt-Winters*, otimizado com uma sazonalidade de 7 dias, demonstrou bom desempenho, com um Erro Absoluto Médio (MAE) de 1351 e um Erro Percentual Absoluto Médio (MAPE) de 3,47%, indicando previsões relativamente precisas. As previsões para os próximos 7 dias sugerem um aumento nas vendas, variando entre 35.735 e 49.565 unidades. Este padrão sugere uma tendência crescente, possivelmente refletindo variações sazonais ou um aumento na procura.

Para o estado da Califórnia (CA), o modelo também apresentou um bom desempenho com uma sazonalidade de 7 dias, com um MAE de 436 e um MAPE de 2,56%, indicando previsões bastante precisas. As previsões para os próximos 7 dias mostram um aumento consistente nas vendas, começando em 17.512 e atingindo 22.514 unidades, sugerindo uma tendência crescente na procura ao longo da semana. No Texas (TX) e em Wisconsin (WI), o desempenho foi menos preciso. O Texas apresentou um MAE de 663 e um MAPE de 6,17%, com previsões variando entre 10.478 e 12.868 unidades, refletindo uma maior variabilidade nas vendas. Wisconsin teve o maior MAE de 999 e um MAPE de 8,32%, com previsões variando entre 10.862 e 13.692 unidades, indicando uma tendência de aumento nas vendas, mas com maior incerteza quanto à precisão das previsões. Os resultados das métricas do modelo *Holt-Winters* podem ser consultados na Tabela A.10, e os resultados das previsões estão disponíveis na Tabela A.11.

Tabela A. 10: Resultados das métricas com *Holt-Winters*

Estado/loja	Métricas de Avaliação			
	MAE	RMSE	MAPE	MSE
Modelo por estado				
Califórnia	436	608	2,56	369.687
Texas	663	759	6,17	576.567
Wisconsin	999	1047	8,32	1.097.062
Modelo por Loja				
CA_1	148	226	2,92	51.504
CA_2	147	198	3,55	39.585
CA_3	238	295	4,37	87.066
CA_4	127	156	4,48	24.354
TX_1	173	181	5,35	33.002
TX_2	272	338	7,34	114.344
TX_3	317	365	9,11	133.900
WI_1	132	247	3,04	61.313
WI_2	579	601	12,68	362.115
WI_3	305	322	8,38	103.984
Modelo Global	1351	1618	3,47	2.619.261

As lojas dos estados da Califórnia e do Texas demonstraram um desempenho geral sólido com o modelo *Holt-Winters* configurado com uma sazonalidade de 7 dias. Na Califórnia, as lojas CA_2 e CA_4 apresentaram os melhores resultados, com baixos valores de MAE e MAPE, indicando previsões precisas e consistentes. Em contraste, a loja CA_3 mostrou um MAE mais elevado e maior variação nas previsões, enquanto a loja CA_1 também obteve bons resultados, com erros relativamente baixos e uma tendência de aumento nas vendas. No Texas, a loja TX_1 apresentou

previsões confiáveis, com MAE e MAPE moderados, enquanto as lojas TX_2 e TX_3 exibiram maior variabilidade e erros mais elevados, especialmente a TX_3, que utilizou uma sazonalidade mais longa de 30 dias.

No estado de Wisconsin, a loja WI_1 destacou-se com o melhor desempenho geral, apresentando baixos valores de MAE e MAPE e previsões de vendas crescentes e precisas. A loja WI_2 apresentou a maior variabilidade nas previsões, com MAE e MAPE significativamente mais elevados, refletindo menor precisão. A loja WI_3 mostrou um desempenho intermédio, com uma precisão razoável, mas com erros mais elevados em comparação com a WI_1. Em todos os estados, as lojas mostraram uma tendência de aumento nas vendas nas previsões para os 7 dias, embora a precisão e a consistência das previsões variem consoante as características específicas de cada loja.

Tabela A. 11: Resultados das previsões com *Holt-Winters*

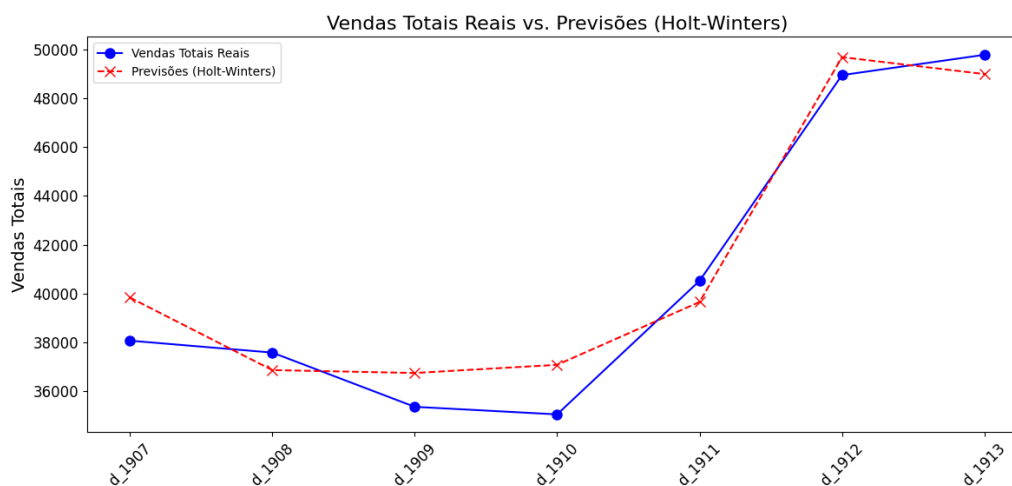
Estado/loja	Previsão das Vendas						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Califórnia	17512	16165	15870	15870	17347	22012	22514
CA_1	4352	3862	3788	3800	4289	5654	5715
CA_2	3888	3716	3669	3671	4387	5847	5782
CA_3	6193	5681	5487	5412	5667	7214	7629
CA_4	2762	2561	2509	2532	2587	3018	3081
Texas	10478	9575	9461	9633	10155	12270	12868
TX_1	3159	2955	2921	2932	3113	3825	4092
TX_2	3742	3338	3345	3449	3671	4455	4662
TX_3	3747	3829	3825	3963	3968	4018	4096
Wisconsin	10862	10674	10723	11097	12638	15208	13692
WI_1	3232	3059	3162	3243	4007	5297	4781
WI_2	4511	4433	4462	4533	4860	5359	4933
WI_3	3318	3287	3177	3348	3829	4710	4312
Total	38470	36210	35735	36480	39889	49565	49153

Os resultados do modelo *Holt-Winters* mostram que o agregado por loja teve o melhor desempenho, com um MAE de 243 e um RMSE de 292, enquanto o agregado por estado e o agregado global apresentaram erros mais elevados, especialmente o agregado global, com um RMSE de 1618. O MAPE foi menor no agregado global, com 3,47%, comparado a 5,68% no agregado por estado e 5,68% no agregado por loja. As previsões diárias foram semelhantes entre os três níveis, mas o modelo por loja ofereceu maior precisão e consistência, sendo superior no geral em termos de desempenho preditivo. Os resultados do modelo agregado global, por estado e por loja para o modelo *Holt-Winters* podem ser visualizados na Tabela A.12.

Em suma, o modelo *Holt-Winters* demonstrou uma capacidade robusta para prever as vendas, com uma precisão satisfatória a nível global e variações regionais que oferecem percepções úteis para ajustes específicos. A metodologia revelou-se eficaz na captura de padrões sazonais e tendências, com resultados que suportam a utilização do modelo em contextos reais de previsão de vendas. No entanto, a análise também destacou a necessidade de considerar as particularidades regionais para melhorar ainda mais a precisão das previsões. A Figura A.5 apresenta o gráfico das previsões do modelo *Holt-Winters* em comparação com as vendas reais durante o período de teste.

Tabela A. 12: Comparação entre os modelos agregados com *Holt-Winters*

Métrica	Agregado Global	Agregado por Estado	Agregado por Loja
MAE	1351	699	243
RMSE	1618	804	292
MAPE	3,47%	5,68%	6,12%
MSE	2.619.261	681.105	101.116
Previsão	Agregado Global	Agregado por Estado	Agregado por Loja
Dia 1	38470	38852	38904
Dia 2	36210	36414	36721
Dia 3	35735	36054	36345
Dia 4	36480	36600	36883
Dia 5	39889	40140	40378
Dia 6	49565	49490	49397
Dia 7	49153	49074	49083

Figura A. 4: Gráfico de comparação entre as vendas Reais e as previsões utilizando *Holt-Winters*

A.5 ETS

A aplicação do modelo de Suavização Exponencial Tripla (ETS) para a previsão de vendas forneceu insights significativos sobre a precisão das previsões em diferentes contextos. O modelo ETS global, configurado com sazonalidade de 7 dias, demonstrou uma boa precisão nas previsões de vendas. Com um MAE de 1.186 e um MAPE de 2,90%, o modelo revelou-se eficaz na captura da tendência geral das vendas, apesar de apresentar algum erro. As previsões para os próximos 7 dias indicam um crescimento constante, começando em 37.690 e subindo para 49.621, sugerindo um aumento significativo na procura ao longo do período.

O modelo ETS para o estado da Califórnia (CA), com sazonalidade de 7 dias, mostrou um desempenho sólido com métricas de erro relativamente baixas. O Erro Absoluto Médio (MAE) foi de 367 e o Erro Percentual Absoluto Médio (MAPE) foi de 2,15%, indicando previsões bastante precisas. As previsões para os próximos 7 dias sugerem um crescimento significativo nas vendas, começando em cerca de 16.989 e atingindo 22.804, o que aponta para uma tendência consistente de aumento na procura.

Em contraste, o modelo ETS para o Texas (TX) e Wisconsin (WI) apresentou erros mais elevados, refletidos em MAEs de 707 e 887, respetivamente, e MAPEs de 6,52% e 7,23%. Essas métricas indicam uma maior variabilidade nas previsões em comparação com a Califórnia. Para o Texas, as previsões para os próximos 7 dias variam de aproximadamente 10.220 a 12.983, sugerindo um crescimento moderado. Para Wisconsin, as previsões começam em 10.738 e vão até 15.523, prevendo um aumento mais acentuado na procura. Esses resultados destacam a necessidade de ajustes mais refinados nos modelos para melhorar a precisão das previsões nos estados com maior erro. Os resultados das métricas do modelo ETS podem ser consultados na Tabela A.13, e as previsões estão disponíveis na Tabela A.14.

Tabela A. 13: Resultados das métricas com ETS

Estado/loja	Métricas de Avaliação			
	MAE	RMSE	MAPE	MSE
Modelo por estado				
Califórnia	367	440	2,15	193.750
Texas	707	784	6,52	615.380
Wisconsin	887	955	7,23	913.063
Modelo por Loja				
CA_1	142	172	3,03	29.726
CA_2	125	173	3,00	29.932
CA_3	177	237	3,21	56.271
CA_4	147	176	5,27	31.318
TX_1	166	191	5,09	36.496
TX_2	280	361	7,43	130.422
TX_3	314	366	9,09	134.562
WI_1	134	248	3,09	61.626
WI_2	610	673	12,92	453.376
WI_3	230	252	6,10	63.918
Modelo Global	1186	1304	2,90	1.702.732

Tabela A. 14: Resultados das previsões com ETS

Estado/loja	Previsão das Vendas						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Califórnia	16989	15583	15096	15238	17195	21955	22804
CA_1	4310	3760	3675	3767	4408	5747	5908
CA_2	3855	3701	3642	3590	4477	5820	5885
CA_3	6051	5563	5338	5352	5633	7267	7730
CA_4	2741	2520	2463	2524	2580	3045	3118
Texas	10220	9734	9403	9721	10399	12352	12982
TX_1	3093	2954	2865	2882	3091	3791	4087
TX_2	3639	3391	3292	3511	3761	4499	4725
TX_3	3768	3837	3831	3955	3947	3992	4045
Wisconsin	10737	10567	10530	10992	12810	15523	14004
WI_1	3212	3037	3133	3218	3994	5281	4786
WI_2	4719	4405	4908	5048	5279	5207	6406
WI_3	3223	3195	2999	3199	3795	4722	4302
Total	37690	35678	34530	35460	40300	49668	49621

A análise dos modelos ETS por loja revela variações significativas nos erros de previsão entre diferentes localizações. Nas lojas da Califórnia, como a CA_1 e a CA_2, foram observadas métricas robustas, com baixos MAE de 142 e 125, respectivamente, e MAPEs em torno de 3%, indicando previsões bastante precisas. Para os próximos 7 dias, as vendas nessas lojas seguem uma tendência de crescimento constante, com a CA_1 variando de 4.310 a 5.908 unidades e a CA_2 de 3.855 a 5.885 unidades, sugerindo uma demanda relativamente estável e previsível. Em contraste, as lojas no Texas e em Wisconsin mostraram maior variabilidade e erros mais elevados. Por exemplo, a loja TX_3 e a WI_2 apresentaram MAEs de 314 e 610, respectivamente, e MAPEs de 9,10% e 12,93%, indicando uma dificuldade maior em prever com precisão as vendas. A TX_3 prevê um aumento gradual nas vendas, de 3.768 para 4.045 unidades, enquanto a WI_2, com uma sazonalidade de 30 dias, projeta variações mais amplas, com um salto significativo de 4.719 para 6.406 unidades. Estes resultados evidenciam que, em algumas lojas, os padrões de demanda são mais instáveis, exigindo ajustes mais refinados no modelo para melhorar a precisão das previsões.

A Figura A.5 apresenta o gráfico das previsões do modelo ETS em comparação com as vendas reais durante o período de teste.

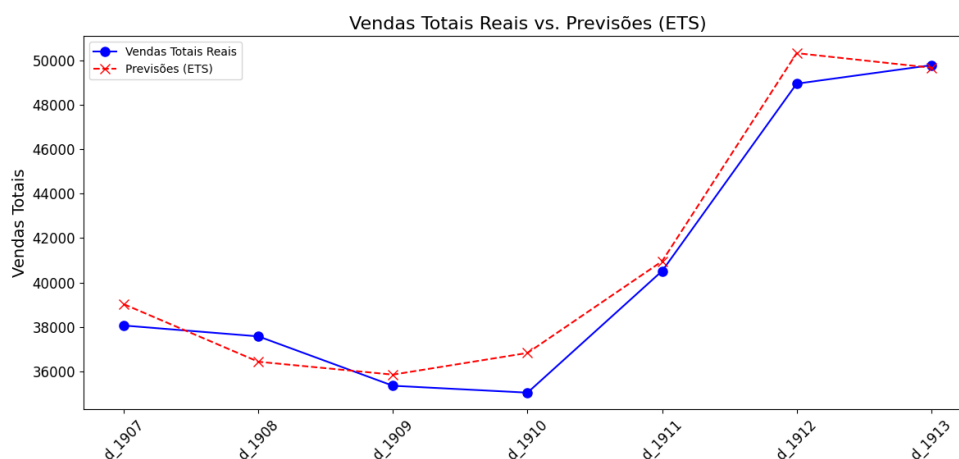


Figura A. 5: Gráfico de Comparação entre as Vendas Reais e as Previsões utilizando ETS

Tabela A. 15: Comparação entre os modelos agregados com ETS

Métrica	Agregado Global	Agregado por Estado	Agregado por Loja
MAE	1186	653	232
RMSE	1304	726	284
MAPE	2,9%	5,3%	5,82%
MSE	1.702.732	574.064	102.764
Previsão	Agregado Global	Agregado por Estado	Agregado por Loja
Dia 1	37690	37946	38611
Dia 2	35678	35884	36363
Dia 3	34530	35029	36146
Dia 4	35460	35951	37046
Dia 5	40300	40404	40965
Dia 6	49668	49830	49371
Dia 7	49621	49790	50992

Os resultados do modelo ETS indicam que o agregado por loja apresenta o melhor desempenho em relação aos erros, com um MAE de 232 e um RMSE de 284, enquanto o agregado por estado e o agregado global tiveram erros mais altos, com o agregado global apresentando o maior RMSE de 1304. As previsões diárias foram próximas entre os três agregados, mas o desempenho geral do modelo por loja é o mais eficaz, com menor variação nos erros e maior precisão nas previsões. Os resultados do modelo agregado global, por estado e por loja para o modelo ETS podem ser visualizados na Tabela A.15.

A.5 Regressão Linear Simples

A regressão linear simples foi utilizada para prever vendas diárias em diferentes estados e lojas do Walmart, assumindo uma relação linear entre variáveis preditoras e vendas. A simplicidade e interpretabilidade da regressão linear tornam-na uma escolha clássica e eficaz para análises de séries temporais, especialmente quando se procura entender a influência de fatores específicos, como fins de semana e feriados, sobre o volume de vendas. Para todos os cenários e modelos desenvolvidos, não foram previstas as vendas para dias que coincidiam com feriados, o que permitiu uma análise mais controlada das variáveis que influenciam as vendas.

O modelo global de regressão linear apresentou resultados variados em diferentes cenários. O melhor desempenho em termos de R^2 foi no Cenário 3 (dados dos últimos 3 meses), com 0,3394, indicando uma explicação de cerca de 33,94% da variabilidade das vendas, enquanto o pior resultado foi no cenário 4 (dados do último mês), com um R^2 negativo (-0,145), indicando que o modelo teve um desempenho inferior ao de uma simples média das vendas. As métricas de erro absoluto médio (MAE) e raiz do erro quadrático médio (RMSE) variaram, com o cenário 2 (dados do último ano) apresentando o menor erro (MAE de 1456 e RMSE de 1966), sugerindo que esse cenário teve previsões mais precisas em termos absolutos. As previsões diárias também variaram significativamente entre os cenários, com as vendas diárias previstas a oscilarem de 29.236 a 56.062 unidades nos diferentes cenários e dias. Os resultados das métricas de avaliação do modelo global estão disponíveis na Tabela A.16, enquanto as previsões podem ser visualizadas na Tabela A.17.

Tabela A. 16: Resultados das métricas do modelo global com Regressão Linear Simples

Modelo Global	Métricas de Avaliação				
	R^2	MAE	RMSE	MAPE	MSE
Cenário 1	0,2733	1722	2286	1140	5.484.316
Cenário 2	0,0633	1456	1966	11,09	4.068.701
Cenário 3	0,3394	1771	2233	12,96	5.489.701
Cenário 4	-0,145	1997	2354	14,32	5.568.125

Tabela A. 17: Resultados das previsões do modelo global com Regressão Linear Simples

Modelo Global	Previsões						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Cenário 1	29525	29384	29295	29236	29198	36718	41584
Cenário 2	49795	44827	41370	38965	37292	45338	50937
Cenário 3	49795	44664	41609	39788	38704	47701	53062
Cenário 4	49795	43310	39879	38063	37102	49501	56062

Os resultados do modelo por estado mostram um desempenho variável entre os cenários, com destaque para o cenário 3 (últimos 3 meses), que apresentou um bom desempenho geral em termos de R^2 , MAE e RMSE. Em particular, a Califórnia teve um R^2 de 0,8417 e um MAE de 851, demonstrando a alta precisão do modelo para esse estado. O Texas teve o melhor resultado entre todos os estados, com um R^2 de 0,8716 e o menor MAE de 621, enquanto o Wisconsin também apresentou um desempenho sólido, com um R^2 de 0,7352. Estes resultados sugerem que o cenário 3 é eficaz para prever vendas, com boa precisão e erros absolutos e quadráticos baixos.

Tabela A. 18: Resultados das métricas do modelo por estado com Regressão Linear Simples

Modelo por Estado	Métricas de Avaliação				
	R^2	MAE	RMSE	MAPE	MSE
Cenário 1					
Califórnia	0,5515	1642	2257	1179	5.096.889
Texas	0,5136	1027	1432	773	2.051.923
Wisconsin	0,3490	1605	2113	1610	4.466.981
Cenário 2					
Califórnia	0,7830	970	1280	5,74	1.638.770
Texas	0,6089	659	838	6,19	703.631
Wisconsin	0,4681	1278	1618	10,98	2.618.465
Cenário 3					
Califórnia	0,8417	851	1326	4,76	1.758.472
Texas	0,8716	621	697	5,41	487.018
Wisconsin	0,7352	1380	1525	11,62	2.327.852
Cenário 4					
Califórnia	0,8068	1491	1589	7,64	2.527.769
Texas	0,8645	542	628	4,42	394.982
Wisconsin	0,8082	985	1029	7,22	1.060.290

Tabela A. 19: Resultados das previsões do modelo por estado com Regressão Linear Simples

Modelo por Estado	Previsões						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Cenário 1							
Califórnia	13032	12166	11904	11824	11800	15881	17109
Texas	8545	8501	8490	8487	8486	10967	11585
Wisconsin	7948	7237	6960	6853	6811	8611	9296
Cenário 2							
Califórnia	23187	17516	15888	15420	15420	20329	21739
Texas	12282	10527	10052	9923	9923	12411	13085
Wisconsin	14326	12174	11335	11008	11008	13016	13799
Cenário 3							
Califórnia	23187	16966	15951	15785	15785	21609	22559
Texas	12282	10651	10332	10270	10270	12835	13337
Wisconsin	14326	12461	12154	12103	12103	15375	15914
Cenário 4							
Califórnia	23187	15865	15869	15869	15869	22555	22552
Texas	12282	10175	10179	10179	10179	12966	12960
Wisconsin	14326	12104	11954	11944	11944	16105	16386

Por outro lado, o cenário 2 (último ano) do modelo por estado também teve um bom desempenho, especialmente para a Califórnia, que atingiu um R^2 de 0.7830 e um MAE de 970. O Cenário 4 (último mês), também teve bons resultados principalmente no Texas (R^2 de 0.8645 e o menor MAE de 542), e em Wisconsin (R^2 de 0.8082 e o menor MAE de 985), se destacando como o melhor cenário em termos de média das métricas para o modelo individual por estado, quando comparado com os demais cenários analisados. Os resultados das métricas e previsões do modelo por estado estão disponíveis na Tabela A.18 e Tabela A.19, respectivamente.

Os resultados do modelo por loja mostram uma variação significativa no desempenho das previsões entre os diferentes cenários. O Cenário 3 destacou-se como o melhor cenário geral, com altos valores de R^2 em várias lojas. Em particular, as lojas da Califórnia (CA_1 a CA_4) tiveram um bom desempenho, com R^2 variando de 0.8555 a 0.6578 e valores baixos de MAE e RMSE, indicando previsões mais precisas. No Texas, as lojas TX_1, TX_2 e TX_3 também obtiveram bons resultados, com R^2 superiores a 0.80 e erros absolutos e quadráticos relativamente baixos. As lojas de Wisconsin apresentaram uma performance mais mista, com WI_1 e WI_3 tendo bons resultados, enquanto WI_2 teve uma performance inferior, com o maior RMSE e um R^2 mais baixo.

O cenário 2 também apresentou bons resultados, especialmente nas lojas da CA e TX, com valores de MAE e RMSE um pouco mais altos, mas equilibrados. O cenário 4, embora tenha mostrado resultados decentes em algumas lojas, como TX_1, que teve o menor MAE de todos os cenários (140), não conseguiu superar o desempenho do Cenário 3 em termos de precisão geral, quando realizada a média das métricas. Portanto, o Cenário 3 pode ser considerado o mais eficaz para prever vendas por loja utilizando regressão linear simples. Os resultados das métricas e previsões do modelo por loja estão disponíveis na Tabela A.20 e Tabela A.21, respectivamente.

Como o melhor cenário identificado, em termos de média das métricas em comparação com os restantes cenários, foi o terceiro, utilizando os dados dos últimos três meses, foi realizada uma agregação do modelo por estado e por loja, com a média das métricas do cenário 3, para permitir uma comparação mais precisa com o modelo agregado global. Os resultados das agregações mostram que o modelo global, embora mais simples e direto, apresenta métricas significativamente menos favoráveis em comparação com os modelos individuais por estado e por loja. Por exemplo, o agregado global resultou num R^2 de 0,3394, enquanto o agregado por estado atingiu um R^2 de 0,8162, e o agregado por loja obteve 0,7667, o que indica que os modelos individualizados explicam melhor a variabilidade das vendas do que o modelo global.

Além disso, o MAE e o RMSE também foram menores nos modelos por estado e por loja. O MAE do modelo global foi de 1771, comparado a 950 no agregado por estado e a 315 no agregado por loja, mostrando que os modelos mais detalhados são mais precisos nas previsões. Da mesma forma, o RMSE foi muito menor nos agregados por estado (1182) e por loja (399) em comparação ao global (2233). Esses resultados indicam que a segmentação por estado ou loja oferece previsões mais precisas e detalhadas, sendo o modelo por loja o mais eficaz em termos de menor erro absoluto e quadrático, embora o agregado por estado também tenha mostrado um bom equilíbrio entre simplicidade e precisão. Portanto, o modelo individual por loja é o melhor para capturar nuances e variações locais nas vendas, enquanto o modelo por estado também pode ser uma boa opção dependendo do nível de detalhamento necessário. Os resultados do modelo agregado global, por estado e por loja para o modelo de regressão linear simples podem ser visualizados na Tabela A.22.

Tabela A. 20: Resultados das métricas do modelo por loja com Regressão Linear Simples

Modelo por Loja	Métricas de Avaliação				
	R ²	MAE	RMSE	MAPE	MSE
Cenário 1					
CA_1	0,5873	509	678	∞	460.610
CA_2	0,5351	464	619	666	383.380
CA_3	0,4906	645	885	716	784.567
CA_4	0,3265	263	361	∞	130.657
TX_1	0,5388	319	440	1383	193.926
TX_2	0,4581	435	624	∞	389.642
TX_3	0,4479	376	520	973	271.303
WI_1	0,3663	605	802	942	644.269
WI_2	0,3759	765	991	∞	982.785
WI_3	0,4356	532	683	1526	467.056
Cenário 2					
CA_1	0,7353	306	424	6,88	180.537
CA_2	0,6924	453	590	13,29	349.091
CA_3	0,6564	368	531	5,74	282.665
CA_4	0,3484	155	200	6,33	40.280
TX_1	0,6464	219	279	6,86	78.063
TX_2	0,5450	285	359	7,56	129.416
TX_3	0,4555	246	306	6,66	93.993
WI_1	0,6141	374	498	10,30	248.478
WI_2	0,4722	553	689	12,64	475.987
WI_3	0,5116	419	508	12,05	258.757
Cenário 3					
CA_1	0,8555	236	361	4,86	130.711
CA_2	0,8058	338	496	7,89	246.850
CA_3	0,7567	391	512	6,19	262.593
CA_4	0,6578	140	177	5,65	31.455
TX_1	0,8351	215	243	6,85	59.395
TX_2	0,8642	239	283	5,69	80.435
TX_3	0,8070	211	264	5,34	69.709
WI_1	0,7814	402	512	12,81	263.046
WI_2	0,5214	580	686	11,82	470.746
WI_3	0,7822	407	458	11,38	210.659
Cenário 4					
CA_1	0,7716	419	478	8,94	228.644
CA_2	0,8486	380	449	7,47	202.386
CA_3	0,7691	466	591	6,25	349.990
CA_4	0,4521	228	251	8,57	63.242
TX_1	0,8649	140	195	3,76	38.192
TX_2	0,8101	291	315	6,88	99.793
TX_3	0,6861	273	297	7,04	88.371
WI_1	0,8055	332	382	8,40	146.314
WI_2	0,4139	560	654	10,69	428.146
WI_3	0,7676	256	350	5,69	122.985

Tabela A. 21: Resultados das previsões do modelo por loja com Regressão Linear Simples

Modelo por Loja	Previsões						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Cenário 1							
CA_1	3363	3103	3029	3009	3003	4310	4677
CA_2	2446	2271	2207	2183	2174	3265	3663
CA_3	5270	4924	4807	4768	4755	6049	6481
CA_4	1953	1808	1758	1741	1735	2093	2216
TX_1	2408	2519	2547	2553	2555	3354	3548
TX_2	3428	3280	3243	3234	3232	4227	4473
TX_3	2709	2637	2611	2602	2598	3264	3504
WI_1	2383	1923	1716	1622	1580	2299	2605
WI_2	3216	2759	2513	2381	2309	2669	2824
WI_3	2349	2486	2541	2563	2572	3260	3541
Cenário 2							
CA_1	6113	4511	4079	3962	3962	5398	5785
CA_2	6082	4162	3555	3363	3363	4875	5353
CA_3	7721	6463	5955	5750	5750	7253	7861
CA_4	3271	2683	2456	2369	2369	2753	2902
TX_1	4033	3204	2955	2881	2881	3723	3976
TX_2	4292	3704	3566	3534	3534	4478	4698
TX_3	3957	3627	3534	3508	3508	4216	4416
WI_1	4874	3357	3251	3244	3244	4453	4537
WI_2	5127	4844	4672	4568	4568	4840	5005
WI_3	4325	3643	3386	3289	3289	4001	4270
Cenário 3							
CA_1	6113	4249	3959	3914	3914	5635	5903
CA_2	6082	4047	3727	3677	3677	5464	5745
CA_3	7721	6211	5842	5752	5752	7490	7914
CA_4	3271	2639	2464	2415	2415	2937	3082
TX_1	4033	3189	2992	2946	2946	3856	4069
TX_2	4292	3806	3726	3713	3713	4725	4893
TX_3	3957	3681	3621	3607	3607	4246	4387
WI_1	4874	3235	3386	3372	3372	5227	5056
WI_2	5127	5127	5128	5128	5128	5572	5795
WI_3	4325	3736	3637	3621	3621	4699	4880
Cenário 4							
CA_1	6113	4020	3955	3953	3953	5880	5941
CA_2	6082	3646	3815	3803	3803	6013	5859
CA_3	7721	5742	5657	5654	5654	7625	7710
CA_4	3271	2446	2454	2454	2454	3063	3057
TX_1	4033	3037	3003	3002	3002	3969	4002
TX_2	4292	3583	3599	3599	3599	4698	4673
TX_3	3957	3569	3573	3573	3573	4299	4291
WI_1	4874	2979	3331	3266	3266	5316	4935
WI_2	5127	5177	5194	5199	5199	6067	6343
WI_3	4325	3571	3563	3563	3563	4876	4890

Tabela A. 22: Comparação entre os modelos agregados com Regressão Linear Simples (cenário 3)

Métrica	Agregado Global	Agregado por Estado	Agregado por Loja
R ²	0,3394	0,8162	0,7667
MAE	1771	950	315
RMSE	2233	1182	399
MAPE	12,96%	7,26%	7,84%
MSE	5.489.701	1.524.447	182.559
Previsão	Agregado Global	Agregado por Estado	Agregado por Loja
Dia 1	49795	49795	49795
Dia 2	44664	40078	39920
Dia 3	41609	38437	38482
Dia 4	39788	38158	38145
Dia 5	38704	38158	38145
Dia 6	47701	49819	49851
Dia 7	53062	51810	51724

A regressão linear simples é uma técnica fundamental para modelar a relação entre uma variável dependente e uma independente, permitindo previsões baseadas em tendências lineares. No contexto das previsões de vendas, é crucial avaliar modelos tanto a nível global quanto segmentados por estado e loja, uma vez que um modelo global pode captar tendências gerais, mas pode não refletir as particularidades regionais que influenciam as vendas locais. Treinar modelos com diferentes períodos de dados, como três meses ou um ano, é igualmente importante, pois permite comparar o impacto de intervalos de tempo distintos nas previsões, ajudando a identificar o modelo mais eficiente e adaptado às necessidades específicas, garantindo maior precisão e relevância nas decisões estratégicas. O gráfico que mostra a dispersão entre as vendas reais e previstas para cada loja esta disponível na Figura A.6.

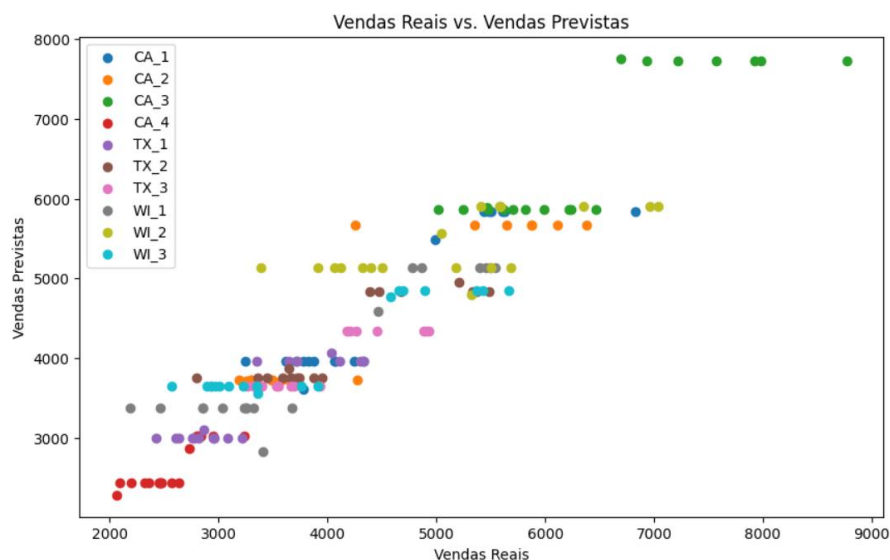


Figura A. 6: Gráfico de dispersão entre as vendas reais e previstas com Regressão Linear Simples

A.6 Regressão Polinomial

Neste estudo, a regressão polinomial foi empregada como uma abordagem adicional para modelar as vendas diárias em diversas localidades e filiais do Walmart nos Estados Unidos, com o objetivo de prever tendências futuras do mercado. Diferente da regressão linear simples, que assume uma relação direta e linear entre as variáveis preditoras e a variável dependente, a regressão polinomial permite a inclusão de termos polinomiais, oferecendo uma maneira de capturar padrões mais complexos e não lineares.

No modelo global com regressão polinomial, o cenário 3 foi o mais eficaz, com um R^2 de 0,4876, indicando uma explicação razoável da variabilidade das vendas e apresentando erros moderados, com um MAE de 1590 e um RMSE de 1985. Este cenário proporcionou um equilíbrio entre a explicação da variabilidade dos dados e a precisão das previsões. Em contraste, o cenário 2, apesar de apresentar o menor MAE (1385) e RMSE (1817), explicou menos a variabilidade dos dados (R^2 de 0,2241). O cenário 4, com o R^2 mais baixo e maiores erros absolutos e quadráticos, revelou-se o menos eficiente. Assim, o cenário 3 é recomendado de forma geral, por oferecer a melhor combinação entre explicação da variabilidade e precisão das previsões. Os resultados das métricas de avaliação do modelo global estão disponíveis na Tabela A.23, enquanto as previsões podem ser visualizadas na Tabela A.24.

Tabela A. 23: Resultados das métricas do modelo global com Regressão Polinomial

Modelo Global	Métricas de Avaliação				
	R^2	MAE	RMSE	MAPE	MSE
Cenário 1	0,2937	1713	2267	1128	5.425.607
Cenário 2	0,2241	1385	1817	10,52	3.544.917
Cenário 3	0,4876	1590	1985	11,63	4.367.694
Cenário 4	0,1740	1829	2131	13,19	4.640.027

Tabela A. 24: Resultados das previsões do modelo global com Regressão Polinomial

Modelo Global	Previsões						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Cenário 1	29525	29451	29327	29231	29171	36091	40962
Cenário 2	49795	44724	40852	38091	36246	44019	50312
Cenário 3	49795	44657	41684	40044	39166	44805	50527
Cenário 4	49795	43525	39158	36683	35612	45039	51363

Os resultados do modelo por estado evidenciam diferenças significativas no desempenho dos cenários de regressão polinomial. O cenário 4 destacou-se como o mais eficaz em termos de métricas de avaliação, apresentando o maior valor de R^2 para a Califórnia (0,9505) e o Texas (0,8287), o que demonstra uma excelente capacidade de explicação da variabilidade das vendas. Este cenário também registou os menores valores de MAE (662 para a Califórnia e 642 para o Texas) e RMSE (805 para a Califórnia e 707 para o Texas), refletindo previsões mais precisas. Em contraste, o cenário 3, embora tenha apresentado um bom desempenho na Califórnia, com um R^2 de 0,8677, revelou um desempenho inferior no Texas (R^2 de 0,2064) e no Wisconsin (R^2 de 0,5998), com erros de previsão superiores quando comparado ao cenário 4.

Os resultados dos cenários 2 e 1 revelam desempenhos intermédios, com o cenário 2 a demonstrar uma maior precisão nas previsões em relação ao Cenário 1. De forma geral, o Cenário 4 oferece a melhor combinação de precisão e capacidade explicativa das vendas em todos os estados, sendo a opção preferencial para a modelação dos dados. Em comparação, o cenário 3, apesar de apresentar resultados relativamente positivos em algumas métricas, não supera o cenário 4 em termos de precisão e capacidade explicativa global. Os resultados das métricas e previsões do modelo por estado estão disponíveis na Tabela A.25 e Tabela A.26, respetivamente

Tabela A. 25: Resultados das métricas do modelo por estado com Regressão Polinomial

Modelo por Estado	Métricas de Avaliação				
	R ²	MAE	RMSE	MAPE	MSE
Cenário 1					
Califórnia	0,5514	1654	2257	1161	5.098.076
Texas	0,5158	1031	1429	780	2.042.354
Wisconsin	0,3421	1611	2124	1657	4.514.560
Cenário 2					
Califórnia	0,7590	1016	1349	6,03	1.820.657
Texas	0,6133	647	834	6,06	695.619
Wisconsin	0,4708	1260	1614	10,81	2.605.519
Cenário 3					
Califórnia	0,8677	882	1212	5,14	1.470.004
Texas	0,2064	888	1736	9,11	3.013.315
Wisconsin	0,5998	1487	1876	13,37	3.519.443
Cenário 4					
Califórnia	0,9505	662	805	3,59	647.541
Texas	0,8287	642	707	5,58	499.361
Wisconsin	0,3894	1606	1838	12,61	3.376.508

Tabela A. 26: Resultados das previsões do modelo por estado com Regressão Polinomial

Modelo por Estado	Previsões						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Cenário 1							
Califórnia	13032	11940	11393	11127	11127	12900	13885
Texas	8545	8215	8077	8021	8021	9701	10469
Wisconsin	7948	7287	7009	6897	6897	7480	7734
Cenário 2							
Califórnia	23187	17052	15732	15411	15411	20295	21813
Texas	12282	10426	9989	9893	9893	12275	13171
Wisconsin	14326	12017	11207	10910	10910	13123	14007
Cenário 3							
Califórnia	23187	16834	15817	15717	15717	21379	23200
Texas	12282	10748	10510	10449	10449	12665	13552
Wisconsin	14326	12557	12367	12329	12329	15627	16316
Cenário 4							
Califórnia	23187	16930	15758	15716	15716	21555	23405
Texas	12282	10530	10169	10106	10106	12720	13054
Wisconsin	14326	13146	12582	12323	12323	15871	16291

A análise dos resultados do modelo por loja revela variações notáveis na precisão das previsões e no ajustamento dos modelos. O Cenário 4 demonstrou ser o mais eficiente, com a maioria das lojas a apresentar valores elevados de R^2 , e baixos erros. Em particular, as lojas CA_1, CA_2 e CA_3 destacaram-se, com R^2 de 0,8862, 0,9227 e 0,9303, respectivamente, e MAE de 293, 255 e 276. Estes valores indicam que o Cenário 4 oferece o melhor ajustamento e o menor erro médio para estas lojas. Por outro lado, o Cenário 3 apresentou um desempenho inferior em várias lojas, com alguns valores de R^2 negativos e MAE elevados, especialmente em lojas como CA_2 e TX_1, que registaram valores de MAE de 745 e 308, e R^2 de -0,481 e -0,096. Este cenário não conseguiu captar adequadamente as variações dos dados, resultando em previsões menos precisas. Os resultados das métricas e previsões do modelo por loja estão disponíveis nas tabelas A.27, A.28, A.29 e A.30.

Tabela A. 27: Resultados das métricas do modelo por loja com Regressão Polinomial – Parte 1

Modelo por Loja	Métricas de Avaliação				
	R^2	MAE	RMSE	MAPE	MSE
Cenário 1					
CA_1	0,5874	510	678	∞	460.514
CA_2	0,5362	463	618	683	382.515
CA_3	0,4959	648	881	674	776.364
CA_4	0,3246	261	361	∞	131.024
TX_1	0,5415	320	439	1360	192.765
TX_2	0,4780	430	612	∞	375.344
TX_3	0,4453	376	522	967	272.585
WI_1	0,3904	599	787	982	619.733
WI_2	0,3718	768	994	∞	989.283
WI_3	0,4332	532	684	1528	469.007
Cenário 2					
CA_1	0,7052	315	448	7,06	201.078
CA_2	0,7184	420	565	12,42	319.571
CA_3	0,6626	376	526	5,93	277.555
CA_4	0,3348	152	202	6,23	41.120
TX_1	0,6642	220	272	6,89	74.138
TX_2	0,5268	283	366	7,50	134.617
TX_3	0,4367	246	311	6,62	97.240
WI_1	0,6031	374	505	10,20	255.562
WI_2	0,4285	557	717	12,75	515.435
WI_3	0,5178	414	505	11,88	255.444
Cenário 3					
CA_1	0,8695	256	344	5,78	118.018
CA_2	-0,481	745	1373	19,6	1.884.593
CA_3	0,8549	337	396	5,33	156.644
CA_4	0,7044	142	165	5,73	27.174
TX_1	-0,096	307	629	11,12	395.136
TX_2	0,8868	225	259	5,25	67.081
TX_3	-1,341	410	920	12,13	846.189
WI_1	0,8634	308	405	9,77	164.407
WI_2	-2,946	994	1970	23,84	3.882.413
WI_3	0,7895	402	451	11,22	203.660

Tabela A. 28: Resultados das métricas do modelo por loja com Regressão Polinomial – Parte 2

Cenário 4					
CA_1	0,8862	293	337	6,32	113.920
CA_2	0,9227	255	321	5,29	103.354
CA_3	0,9303	276	324	3,75	105.581
CA_4	0,7530	140	168	5,54	28.510
TX_1	0,9331	104	137	3,05	18.897
TX_2	0,7056	360	393	8,92	154.722
TX_3	0,5519	312	355	8,34	126.153
WI_1	0,5010	504	612	11,55	375.331
WI_2	-3,04	1357	1719	24,97	2.955.894
WI_3	0,3094	474	604	10,04	365.574

Tabela A. 29: Resultados das previsões do modelo por loja com Regressão Polinomial – Parte 1

Modelo por Loja	Previsões						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Cenário 1							
CA_1	3363	3075	2938	2874	2874	3424	3708
CA_2	2446	2263	2186	2155	2155	2697	3022
CA_3	5270	4821	4570	4438	4438	4995	5308
CA_4	1953	1826	1786	1775	1775	1940	1994
TX_1	2408	2392	2386	2383	2383	3014	3326
TX_2	3428	3231	3156	3128	3128	3673	3943
TX_3	2709	2568	2504	2476	2476	2949	3137
WI_1	2383	2072	1922	1845	1845	1959	2050
WI_2	3216	2826	2590	2450	2450	2261	2150
WI_3	2349	2480	2528	2547	2547	3102	3337
Cenário 2							
CA_1	6113	4392	4056	3968	3968	5448	5830
CA_2	6082	3559	3443	3406	3406	4815	5506
CA_3	7721	6404	5925	5742	5742	7299	7850
CA_4	3271	2662	2443	2363	2363	2741	2883
TX_1	4033	3152	2904	2854	2854	3660	4021
TX_2	4292	3692	3560	3530	3530	4444	4737
TX_3	3957	3613	3526	3504	3504	4201	4439
WI_1	4874	3222	3233	3234	3234	4486	4630
WI_2	5127	4864	4715	4629	4629	4902	5051
WI_3	4325	3602	3339	3246	3246	3986	4291
Cenário 3							
CA_1	6113	4258	3971	3906	3906	5698	6063
CA_2	6082	3665	3838	3929	3929	4917	5774
CA_3	7721	6186	5792	5716	5716	7501	8139
CA_4	3271	2694	2465	2395	2395	2972	3183
TX_1	4033	3139	3057	3031	3031	3779	4136
TX_2	4292	3814	3734	3719	3719	4677	4957
TX_3	3957	3744	3701	3690	3690	4276	4462
WI_1	4874	3284	3384	3385	3385	5318	5103
WI_2	5127	5364	5453	5479	5479	5932	6272
WI_3	4325	3758	3664	3645	3645	4892	4867

Tabela A. 30: Resultados das previsões do modelo por loja com Regressão Polinomial – Parte 2

Cenário 4							
CA_1	6113	4551	4006	3915	3915	5578	6199
CA_2	6082	3704	3794	3780	3780	5812	6054
CA_3	7721	6063	5629	5573	5573	7316	7972
CA_4	3271	2715	2496	2435	2435	2985	3192
TX_1	4033	3078	2980	2980	2980	3842	4140
TX_2	4292	3724	3600	3576	3576	4638	4656
TX_3	3957	3656	3584	3568	3568	4237	4284
WI_1	4874	3469	3340	3316	3316	5537	4416
WI_2	5127	5189	5239	5280	5280	5386	5483
WI_3	4325	3683	3617	3605	3605	5216	4430

Os resultados das agregações para o quarto cenário, que foi o melhor avaliado em termos de média das métricas, mostram variações significativas entre os modelos global, por estado e por loja. No agregado global, o modelo de regressão polinomial obteve um R^2 de 0,1740, um MAE de 1829 e um RMSE de 2131. Estes valores indicam um ajustamento geral limitado, com um erro relativamente elevado em comparação com os modelos mais específicos.

Em contraste, o modelo agregado por estado apresentou uma melhoria significativa nas métricas, com um R^2 de 0,7228, MAE de 970 e RMSE de 1116. Estes resultados indicam um ajustamento mais preciso e um menor erro médio em relação ao modelo global. O modelo agregado por loja, embora forneça previsões mais detalhadas, registou um R^2 de 0,3453, MAE de 407 e RMSE de 497, sugerindo uma precisão intermédia entre os modelos global e por estado. A menor variabilidade e a maior precisão nas previsões do modelo por estado indicam que este é o mais eficiente para captar as nuances dos dados, em comparação com o modelo global e o modelo por loja. Em suma, o agregado por estado oferece a melhor combinação de precisão e ajustamento, superando tanto o modelo global quanto o modelo por loja em termos de qualidade das previsões e precisão das métricas. Os resultados do modelo agregado global, por estado e por loja para o modelo de regressão polinomial podem ser visualizados na Tabela A.31.

Tabela A. 31: Comparação entre os modelos agregados com Regressão Polinomial (cenário 4)

Métrica	Agregado Global	Agregado por Estado	Agregado por Loja
R^2	0,1740	0,7228	0,3453
MAE	1829	970	407
RMSE	2131	1116	497
MAPE	13,19	7,26	8,77
MSE	4.640.027	1.507.803	434.793
Previsão	Agregado Global	Agregado por Estado	Agregado por Loja
Dia 1	49795	49795	49795
Dia 2	43525	40606	39832
Dia 3	39158	38509	38285
Dia 4	36683	38145	38028
Dia 5	35612	38145	38028
Dia 6	45039	50146	50547
Dia 7	51363	52750	50826

Este estudo demonstrou que a regressão polinomial é uma ferramenta poderosa para modelar as vendas diárias em várias filiais do Walmart, captando com sucesso padrões não lineares que a regressão linear simples poderia não identificar. Além disso, o desempenho do modelo variou significativamente entre as lojas, sugerindo que características específicas de cada filial, como a localização geográfica, o perfil da clientela e factores externos, devem ser considerados na modelação preditiva. O cenário de curto prazo provou ser eficaz para ajustar-se rapidamente a mudanças no mercado, como promoções e eventos locais, que influenciam significativamente as vendas diárias. Assim, o quarto cenário revelou-se o mais robusto e adequado para previsões de vendas num ambiente dinâmico, sendo o mais indicado para apoiar decisões estratégicas imediatas no sector do retalho.

O gráfico que compara a diferença entre as vendas reais e as previstas para as lojas, utilizando o modelo de regressão polinomial, pode ser consultado na Figura A.7.

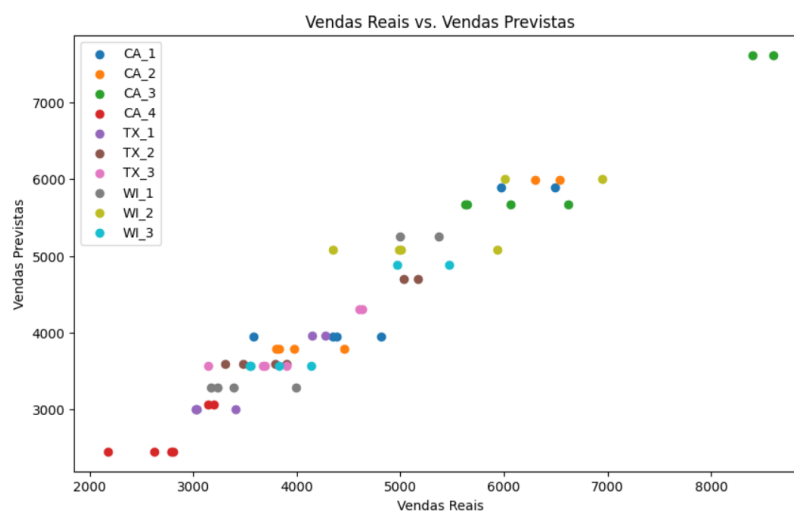


Figura A. 7: Gráfico de dispersão entre as vendas reais e previstas com Regressão Polinomial

A.7 Decision Tree

A *Decision Tree* é uma técnica eficaz para modelar relações não lineares de forma intuitiva e interpretável. Nas previsões de vendas, utiliza dados históricos para dividir o espaço de atributos em regiões homogêneas com base nas variáveis predictoras, identificando padrões complexos, como dias da semana e feriados. A eficácia do modelo depende da configuração adequada dos hiperparâmetros, como a profundidade máxima da árvore, o número mínimo de amostras para divisão e o número mínimo de amostras por folha. Para otimizar o modelo, os hiperparâmetros foram ajustados utilizando o *GridSearchCV*, explorando combinações de profundidade (3, 5, 7, Nenhum), amostras para divisão (2, 5, 10) e amostras por folha (1, 2, 4). A validação cruzada com 3 dobras foi utilizada para encontrar a melhor configuração entre 108 possíveis combinações, assegurando o melhor desempenho do modelo.

O modelo global de árvore de decisão apresentou variações consideráveis nas suas métricas de avaliação ao longo dos diferentes cenários. O cenário 3 destacou-se como o mais eficiente, com um R^2 de 0,51, MAE de 1198 e RMSE de 1588. Por outro lado, o Cenário 1, embora com um R^2 de 0,28, inferior em comparação com os outros cenários, apresentou um MAE de 1655 e RMSE de 2201,

demonstrando um desempenho menos eficiente em termos de precisão e ajustamento das previsões. As previsões variaram entre os cenários, sendo que o cenário 3 ofereceu as estimativas mais detalhadas e ajustadas para os dias subsequentes, destacando-se na capacidade de modelar as tendências de forma mais precisa. Os resultados das métricas de avaliação do modelo global estão disponíveis na Tabela A.32, enquanto as previsões podem ser visualizadas na Tabela A.33.

Tabela A. 32: Resultados das métricas do modelo global com *Decision Tree*

Modelo Global	Métricas de Avaliação				
	R ²	MAE	RMSE	MAPE	MSE
Cenário 1	0,28	1655	2201	633	5.047.134
Cenário 2	0,46	1246	1700	457	2.964.415
Cenário 3	0,51	1198	1588	8,80	2.725.272
Cenário 4	0,29	1420	2065	10,05	4.558.027

Tabela A. 33: Resultados das previsões do modelo global com *Decision Tree*

Modelo Global	Previsões						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Cenário 1	29525	28882	28081	28081	28081	33182	36972
Cenário 2	49795	40867	39456	39456	39456	46823	49943
Cenário 3	49795	44729	43889	43889	43889	54961	61021
Cenário 4	49795	45495	42982	42801	42801	42801	42801

A análise dos resultados do modelo de árvore de decisão por estado revela uma variação significativa no desempenho entre os diferentes cenários. No cenário 2, a Califórnia apresentou o melhor desempenho, com um R² de 0,7994 e um MAE de 975, demonstrando elevada precisão nas previsões. O Texas, por sua vez, obteve um desempenho mais modesto, com um R² de 0,3000 e um MAE de 796, enquanto o Wisconsin apresentou resultados intermédios, com um R² de 0,4874 e um MAE de 1213, indicando uma precisão moderada.

O cenário 3 destacou-se pelo desempenho global, especialmente para a Califórnia e o Texas, com R² de 0,8219 e 0,8199, respetivamente, e menores MAEs. O Wisconsin, embora tenha registado um R² de 0,693, apresentou um MAPE de 11,79%, refletindo a necessidade de melhorias. No cenário 4, o Texas foi o principal destaque, com um R² de 0,8876 e um MAE de 465, evidenciando a melhor precisão entre todos os cenários. Em suma, o cenário 3 revelou-se o mais equilibrado para o modelo por estado, oferecendo a melhor combinação entre precisão e erro em comparação com os outros cenários. Os resultados das métricas do modelo por estado estão disponíveis na Tabela A.34 e A.35 e os resultados das previsões estão na Tabela A.36.

Tabela A. 34: Resultados das métricas do modelo por estado com *Decision Tree* – Parte 1

Modelo por Estado	Métricas de Avaliação				
	R ²	MAE	RMSE	MAPE	MSE
Cenário 1					
Califórnia	0,5365	1651	2295	1277	5.268.245
Texas	0,5069	1032	1442	802	2.080.247
Wisconsin	0,3613	1567	2094	1890	4.383.293

Tabela A. 35: Resultados das métricas do modelo por estado com *Decision Tree* – Parte 2

Cenário 2					
Califórnia	0,7994	975	1231	5,79	1.515.468
Texas	0,3000	796	1122	7,55	1.259.486
Wisconsin	0,4874	1213	1589	10,23	2.523.944
Cenário 3					
Califórnia	0,8219	944	1407	5,37	1.979.464
Texas	0,8199	715	827	6,61	683.938
Wisconsin	0,6930	1444	1643	11,79	2.699.523
Cenário 4					
Califórnia	0,7640	1665	1758	8,59	3.088.845
Texas	0,8876	465	572	3,83	327.580
Wisconsin	0,5424	1385	1591	10,47	2.530.887

Tabela A. 36: Resultados das previsões do modelo por estado com *Decision Tree*

Modelo por Estado	Previsões						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Cenário 1							
Califórnia	13032	13262	13262	13262	13262	13687	13687
Texas	8545	8907	8907	8907	8907	9461	9461
Wisconsin	7948	8211	8211	8211	8211	8443	8443
Cenário 2							
Califórnia	23187	16821	15631	15631	15631	20732	22204
Texas	12282	10620	10941	10247	9856	12267	12382
Wisconsin	14326	12003	10894	10894	10894	13141	14488
Cenário 3							
Califórnia	23187	16820	15995	15995	15995	21804	23351
Texas	12282	11225	9825	10362	10362	12416	13629
Wisconsin	14326	12413	12413	12413	12413	16643	14104
Cenário 4							
Califórnia	23187	15938	15938	15938	15938	22398	22398
Texas	12282	10441	9854	10134	9854	12920	13106
Wisconsin	14326	13195	13195	13195	13195	15589	15589

A análise dos resultados do modelo por loja revela diferentes níveis de desempenho entre os cenários avaliados. No cenário 1, as lojas da Califórnia apresentam um desempenho variável, com a loja CA_1 a registrar o melhor R^2 de 0,5692, enquanto a CA_4 apresenta o menor R^2 de 0,287. No Texas, a loja TX_1 alcança um R^2 de 0,523, mas as lojas TX_2 e TX_3 mostram resultados menos satisfatórios. Em Wisconsin, as lojas WI_1 e WI_3 têm R^2 de aproximadamente 0,43 e 0,41, respectivamente, indicando uma precisão moderada.

No cenário 3, todas as lojas mostram uma melhoria geral, com CA_4 e TX_1 a destacarem-se com R^2 de 0,6691 e 0,8109, respectivamente. O cenário 4 revela-se promissor para a maioria das lojas, mas não supera a média das métricas se comparado com o cenário 3. Assim, o cenário 3 foi o mais eficaz para o modelo por loja, oferecendo os melhores resultados em termos de precisão e erro, refletindo a eficácia da árvore de decisão na previsão de vendas para as lojas. Os resultados das métricas e previsões do modelo por loja encontram-se nas tabelas A.37 e A.38.

Tabela A. 37: Resultados das métricas do modelo por loja com *Decision Tree*

Modelo por Loja	Métricas de Avaliação				
	R ²	MAE	RMSE	MAPE	MSE
Cenário 1					
CA_1	0,5692	512	693	∞	480.932
CA_2	0,5340	443	620	720,41	384.301
CA_3	0,4621	659	910	776,31	828.545
CA_4	0,2870	274	372	∞	138.326
TX_1	0,5230	324	448	1439,89	200.569
TX_2	0,4576	440	625	∞	390.030
TX_3	0,4447	382	522	991,58	272.902
WI_1	0,4347	557	758	1039,23	574.790
WI_2	0,3267	794	1030	∞	1.060.391
WI_3	0,4077	526	700	1704,99	490.142
Cenário 2					
CA_1	0,7021	337	451	7,5	203.196
CA_2	0,8246	329	446	9,05	199.057
CA_3	0,7180	375	482	5,95	232.010
CA_4	0,3429	161	202	6,58	40.623
TX_1	0,6015	233	297	7,37	88.006
TX_2	0,4449	308	397	8,2	157.929
TX_3	0,2802	267	353	7,18	124.267
WI_1	0,6282	365	489	10	239.452
WI_2	0,4400	560	711	12,66	505.093
WI_3	0,4741	412	528	11,48	278.654
Cenário 3					
CA_1	0,8251	313	398	6,75	158.207
CA_2	0,7322	422	584	9,76	340.527
CA_3	0,7392	399	531	6,11	281.543
CA_4	0,6691	136	174	5,49	30.417
TX_1	0,8109	225	261	7,23	68.139
TX_2	0,8329	250	315	6,03	99.027
TX_3	0,7254	250	315	6,44	99.211
WI_1	0,7637	394	533	12,47	284.394
WI_2	0,3125	716	822	14,78	676.291
WI_3	0,7197	437	521	11,78	271.168
Cenário 4					
CA_1	0,6568	530	586	11,12	343.703
CA_2	0,9315	261	303	5,35	91.582
CA_3	0,7585	524	605	7,31	366.076
CA_4	0,5840	211	219	7,88	48.023
TX_1	0,9098	128	160	3,82	25.496
TX_2	0,6033	427	457	10,87	208.541
TX_3	0,4932	334	378	8,34	142.706
WI_1	0,7346	384	447	9,03	199.641
WI_2	0,2075	679	761	12,98	578.957
WI_3	0,3792	467	573	10,16	328.642

Tabela A. 38: Resultados das previsões do modelo por loja com *Decision Tree*

Modelo por Loja	Previsões						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Cenário 1							
CA_1	3363	3632	3632	3632	3632	3633	3633
CA_2	2446	2555	2555	2555	2555	3608	3608
CA_3	5270	5024	5024	5024	5024	5361	5361
CA_4	1953	1711	1711	1711	1711	1732	1732
TX_1	2408	2449	2609	2737	2737	2938	2938
TX_2	3428	3353	3353	3353	3353	4205	4506
TX_3	2709	2699	2699	2699	2699	2929	2929
WI_1	2383	2297	2297	2297	2297	1877	2163
WI_2	3216	2495	1970	1970	1970	1970	1970
WI_3	2349	2663	2928	3236	3236	3548	3928
Cenário 2							
CA_1	6113	4329	3906	3906	3906	5693	5693
CA_2	6082	3593	3593	3593	3593	5288	5288
CA_3	7721	6384	5935	5935	5935	7172	7902
CA_4	3271	2666	2326	2326	2326	2696	2885
TX_1	4033	3071	2819	2819	2819	3508	3850
TX_2	4292	3559	3559	3559	3559	4570	4570
TX_3	3957	3451	3451	3451	3451	4256	4557
WI_1	4874	3314	3314	3314	3314	3763	4576
WI_2	5127	5079	5079	5079	5079	5079	5079
WI_3	4325	3618	3223	3223	3223	3719	4151
Cenário 3							
CA_1	6113	4006	4006	4006	4006	5699	6048
CA_2	6082	3739	3739	3739	3739	5194	6052
CA_3	7721	6330	5876	5876	5876	7450	7746
CA_4	3271	2687	2428	2428	2428	2967	3191
TX_1	4033	3088	2835	2835	2835	3794	4198
TX_2	4292	4170	3707	3707	3707	4793	5013
TX_3	3957	3668	3668	3668	3668	4174	4433
WI_1	4874	3384	3384	3384	3384	5421	4562
WI_2	5127	4978	4978	4978	4978	5275	5800
WI_3	4325	3507	3899	3899	3899	4857	4521
Cenário 4							
CA_1	6113	4246	4246	4246	4246	5816	5653
CA_2	6082	3614	3837	3614	3837	6026	6026
CA_3	7721	6096	5627	5627	5627	7620	7620
CA_4	3271	2572	2492	2492	2492	3046	2988
TX_1	4033	3088	3204	3204	3204	3879	4208
TX_2	4292	3777	3777	3777	3777	4933	4709
TX_3	3957	3524	3796	3524	3796	4264	4264
WI_1	4874	3459	3459	3459	3459	5492	4589
WI_2	5127	5860	5109	5860	5109	5109	5860
WI_3	4325	3483	4008	3483	4008	5102	4449

Ao comparar os resultados dos modelos global, por estado e por loja, o cenário 3 destaca-se como o mais eficaz, devido ao seu desempenho equilibrado em todos os níveis de agregação. Ao analisar os resultados do modelo de árvore de decisão no cenário 3, o agregado por estado sobressai em termos de R^2 , com um valor de 0,7782, superior ao do agregado por loja (0,7130) e ao do agregado global (0,51). Isto indica que o modelo por estado é o mais eficiente a explicar a variabilidade dos dados. Além disso, o agregado por estado apresenta um desempenho superior em termos de MAPE (7,92%), evidenciando um menor erro percentual médio absoluto em comparação com os agregados global e por loja.

No entanto, o modelo agregado por loja destaca-se pelas métricas de erro absoluto e quadrático mais baixas, como o MAE (354) e o RMSE (445), o que significa que as previsões ao nível das lojas são mais precisas em termos de valores absolutos. Ainda assim, considerando o equilíbrio entre a explicação da variabilidade dos dados (R^2) e a precisão relativa (MAPE), o modelo agregado por estado revela-se a melhor opção global, mantendo um bom balanço entre a capacidade explicativa e os erros relativos reduzidos. Os resultados do modelo agregado global, por estado e por loja para o modelo de regressão polinomial podem ser visualizados na Tabela A.39.

Tabela A. 39: Comparação entre os modelos agregados com Decision Tree (cenário 3)

Métrica	Agregado Global	Agregado por Estado	Agregado por Loja
R^2	0,51	0,7782	0,7130
MAE	1198	1034	354
RMSE	1588	1292	445
MAPE	8,8%	7,92%	8,68%
MSE	2.725.272	1.787.642	230.892
Previsão	Agregado Global	Agregado por Estado	Agregado por Loja
Dia 1	49795	49795	49795
Dia 2	44729	40458	39557
Dia 3	43889	38233	38520
Dia 4	43889	38770	38520
Dia 5	43889	38770	38520
Dia 6	54961	50863	49624
Dia 7	61021	51084	51564

Testar diversos cenários é crucial para identificar o modelo mais adequado às especificidades dos dados em análise, permitindo compreender melhor a interação das variáveis e a influência de diferentes períodos de treino na precisão das previsões. A árvore de decisão demonstrou ser uma ferramenta valiosa devido à sua capacidade de lidar com variáveis complexas e não lineares, essencial para captar as nuances do comportamento de vendas em diferentes contextos. A sua flexibilidade e facilidade de interpretação tornam-na fundamental para otimizar previsões e apoiar decisões estratégicas empresariais. O gráfico de dispersão das vendas reais e previstas é visualizado na Figura 26.

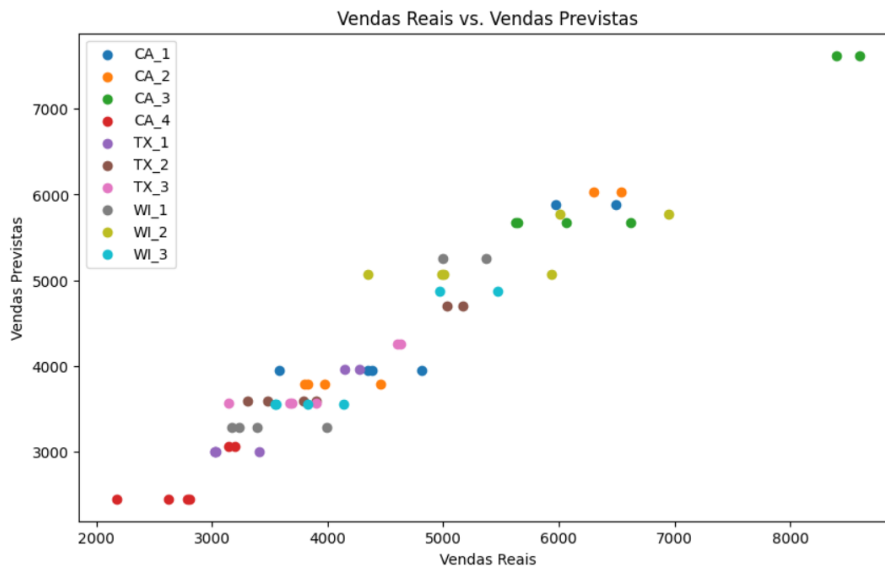


Figura A. 8: Gráfico de dispersão entre as vendas reais e previstas com *Decision Tree*

A.8 XGBoost

O *XGBoost* é um dos algoritmos mais avançados e eficazes para problemas de regressão e classificação, combinando as previsões de múltiplas árvores de decisão, o que ajuda a reduzir o viés e a variância dos modelos. Este algoritmo destaca-se pela sua eficiência computacional e flexibilidade, permitindo a configuração de uma vasta gama de hiperparâmetros que melhoram o desempenho dos modelos. Além disso, o *XGBoost* é capaz de lidar com dados em falta e inclui mecanismos de regularização incorporados, ajudando a evitar o sobreajuste e promovendo uma modelação mais generalizável.

A comparação entre os resultados dos modelos de previsão de vendas utilizando o *XGBoost*, antes e depois da afinação dos hiperparâmetros com o *GridSearch*, revela melhorias significativas no desempenho geral. Sem a otimização dos hiperparâmetros, os modelos apresentaram erros médios absolutos (MAE) a variar entre 306 e 1875, e erros quadráticos médios (MSE) entre 154952 e 5962578, com um R^2 a variar de 0,201 a 0,669. Após a aplicação do *GridSearch*, observou-se uma redução consistente nos erros: o MAE variou de 207 a 1313, e o MSE de 93848 a 3759054, com um R^2 que atingiu até 0,723, evidenciando uma melhoria notável na precisão das previsões. Estes resultados demonstram a eficácia da afinação dos hiperparâmetros, traduzindo-se numa melhoria geral no desempenho dos modelos, o que tem impacto direto numa previsão mais precisa e numa redução dos erros — fatores essenciais para a formulação de estratégias de vendas mais eficazes e fundamentadas, tanto por estado como por loja.

Ao analisar o desempenho do modelo global utilizando *XGBoost*, o cenário 3 destaca-se como o mais eficiente. Este cenário alcança o maior valor de R^2 com 0,5699, indicando que é o modelo que melhor explica a variabilidade dos dados em comparação com os outros cenários. Em termos de MAE e RMSE, o Cenário 3 também apresenta os valores mais baixos, com 1139 e 1497, respetivamente, o que sugere que o modelo tem um desempenho relativamente preciso na previsão dos valores. O MAPE de 8,41% para este cenário é o mais baixo, indicando que o modelo é eficaz na previsão com uma pequena margem percentual de erro. Assim, com um bom equilíbrio

entre a capacidade explicativa e a precisão das previsões, o cenário 3 é a melhor escolha entre os cenários avaliados. Os resultados das métricas de avaliação do modelo global para o *XGBoost* estão disponíveis na Tabela A.40, enquanto as previsões podem ser visualizadas na Tabela A.41.

Tabela A. 40: Resultados das métricas do modelo global com *XGBoost*

Modelo Global	Métricas de Avaliação				
	R ²	MAE	RMSE	MAPE	MSE
Cenário 1	0,4890	1461	2012	800	4.048.902
Cenário 2	0,4909	1221	1638	455,73	2.754.415
Cenário 3	0,5699	1139	1497	8,41	2.380.610
Cenário 4	0,5031	1354	1788	9,39	3.516.886

Tabela A. 41: Resultados das previsões do modelo global com *XGBoost*

Modelo Global	Previsões						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Cenário 1	43875	39877	38122	36987	36987	41798	46928
Cenário 2	49795	41468	38993	37232	37232	43621	47478
Cenário 3	49795	43302	42224	41718	41718	48611	54832
Cenário 4	49795	42230	39719	37678	37678	44682	50349

O cenário 3 é o mais robusto, destacando-se por sua alta precisão e capacidade explicativa. Com valores de R² de 0,8476 na Califórnia, 0,7714 no Texas e 0,6835 em Wisconsin, e baixos MAE e RMSE, especialmente no Texas, onde o MAE é 788 e o RMSE é 931. O MAPE também é favorável, com 5,56% na Califórnia e 7,05% no Texas. Assim, o cenário 3 é o melhor para análise por estado, combinando alta precisão e baixos erros de previsão. Os resultados das métricas do modelo por estado estão disponíveis na Tabela A.42, enquanto as previsões estão disponíveis na Tabela A.43.

Tabela A. 42: Resultados das métricas do modelo por estado com *XGBoost*

Modelo por Estado	Métricas de Avaliação				
	R ²	MAE	RMSE	MAPE	MSE
Cenário 1					
Califórnia	0,6641	1318	1954	1087,26	3.818.145
Texas	0,6421	905	1229	579,64	1.509.935
Wisconsin	0,5521	1251	1753	1226,5	3.073.570
Cenário 2					
Califórnia	0,7978	917	1236	5,48	1.527.643
Texas	0,7014	608	733	5,75	537.241
Wisconsin	0,5808	1061	1437	9,15	2.064.002
Cenário 3					
Califórnia	0,8476	987	1301	5,56	1.693.969
Texas	0,7714	788	931	7,05	867.952
Wisconsin	0,6835	1494	1668	11,98	2.783.037
Cenário 4					
Califórnia	0,8144	1471	1559	7,8	2.429.511
Texas	0,8376	570	688	5,02	473.388
Wisconsin	0,0319	1642	2314	10,99	5.353.504

Tabela A. 43: Resultados das previsões do modelo por estado com *XGBoost*

Modelo por Estado	Previsões						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Cenário 1							
Califórnia	13032	13669	13722	13696	13722	16958	18481
Texas	8545	8737	8737	8737	8737	10735	11320
Wisconsin	7948	8219	8219	8219	8219	9556	9933
Cenário 2							
Califórnia	23187	17361	17234	16960	16667	20057	20926
Texas	12282	10352	9722	9766	9778	11976	12359
Wisconsin	14326	11794	11064	11121	11121	12769	13458
Cenário 3							
Califórnia	23187	17030	17034	17034	17034	21989	22516
Texas	12282	10494	9989	9997	9997	12128	12431
Wisconsin	14326	12787	12455	12455	12455	15144	15248
Cenário 4							
Califórnia	23187	16422	16091	15834	15834	20635	21478
Texas	12282	10323	9681	9906	10173	11614	11663
Wisconsin	14326	11917	11483	11256	11256	14285	14129

Ao analisar o desempenho do modelo por loja, o cenário 3 destaca-se como o mais eficaz na previsão de vendas. Este cenário sobressai particularmente nas métricas de avaliação. O R^2 é elevado para várias lojas, incluindo CA_1 (0,8737) e TX_1 (0,7348), indicando uma boa capacidade do modelo para explicar a variabilidade dos dados. Os valores de MAE e RMSE são relativamente baixos em comparação com outros cenários, com MAE de 253 e RMSE de 338 para CA_1, sugerindo previsões mais precisas. O MAPE também é favorável, com valores de 5,47% para CA_1 e 8,55% para TX_1, evidenciando uma pequena margem de erro percentual nas previsões. Além disso, o cenário 3 proporciona previsões de vendas que estão mais alinhadas com as observações reais em vários dias, especialmente para CA_1 e TX_1, demonstrando uma consistência superior. Assim, o cenário 3 é o mais adequado para o modelo por loja, oferecendo um equilíbrio otimizado entre precisão e capacidade de previsão. Os resultados das métricas e previsões do modelo por loja encontram-se nas tabelas A.44 e A.45.

Como o cenário 3 demonstrou o melhor desempenho, foi realizado um agregado para comparar as métricas de avaliação e previsões a nível global, por estado e por loja. O agregado global do cenário 3 apresenta um R^2 de 0,5699, indicando uma capacidade razoável do modelo para explicar a variabilidade dos dados. No entanto, o agregado por estado revela um desempenho superior, com um R^2 de 0,7675, refletindo uma melhor adaptação do modelo às variações regionais. Por outro lado, o agregado por loja alcança um R^2 de 0,7095, que também é elevado, mas não tão robusto quanto o do nível estadual. Os resultados do modelo agregado global, por estado e por loja para o modelo *XGBoost* podem ser visualizados na Tabela A.46.

Tabela A. 44: Resultados das métricas do modelo por loja com *XGBoost*

Modelo por Loja	Métricas de Avaliação				
	R ²	MAE	RMSE	MAPE	MSE
Cenário 1					
CA_1	0,7001	414	578	∞	334.763
CA_2	0,7295	293	472	776,3	223.104
CA_3	0,6362	521	749	402,17	560.420
CA_4	0,5293	207	302	∞	91.320
TX_1	0,6724	268	371	958,7	137.744
TX_2	0,5953	377	539	∞	290.988
TX_3	0,6048	318	441	604,37	194.214
WI_1	0,6817	375	569	783,94	323.636
WI_2	0,5497	590	842	∞	709.206
WI_3	0,639	415	547	1200,88	298.722
Cenário 2					
CA_1	0,83	240	341	5,39	115.962
CA_2	0,8578	295	402	8,04	161.452
CA_3	0,7335	331	468	5,14	219.294
CA_4	0,4037	139	192	5,71	36.868
TX_1	0,7976	171	211	5,38	44.691
TX_2	0,6293	251	325	6,73	105.449
TX_3	0,5185	231	288	6,28	83.131
WI_1	0,6954	323	443	8,8	196.181
WI_2	0,6136	451	590	10,24	348.506
WI_3	0,5726	352	476	10,08	226.436
Cenário 3					
CA_1	0,8737	253	338	5,47	114.210
CA_2	0,7019	446	616	9,79	379.042
CA_3	0,7342	425	536	6,61	286.986
CA_4	0,8504	86	117	3,57	13.756
TX_1	0,7348	264	309	8,55	95.562
TX_2	0,695	331	425	8,13	180.734
TX_3	0,6226	304	369	7,85	136.348
WI_1	0,8181	355	468	10,48	218.936
WI_2	0,4375	662	744	13,56	553.266
WI_3	0,6275	521	600	14,53	360.343
Cenário 4					
CA_1	0,4637	660	733	13,04	537.008
CA_2	0,9421	230	278	4,32	77.456
CA_3	0,7433	571	624	8,16	389.150
CA_4	0,4504	249	252	9,23	63.438
TX_1	0,9146	142	155	4,12	24.141
TX_2	0,6139	413	451	10,05	203.007
TX_3	0,5515	287	355	7,64	126.274
WI_1	0,4938	474	617	10,91	380.788
WI_2	0,3004	562	715	10,61	511.091
WI_3	0,1098	431	686	8,75	471.261

Tabela A. 45: Resultados das previsões do modelo por loja com *XGBoost*

Modelo por Loja	Previsões						
	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7
Cenário 1							
CA_1	3363	3380	3380	3380	3380	4668	5232
CA_2	2446	2709	2729	2729	2729	3749	3703
CA_3	5270	5310	5278	5310	5278	6523	7224
CA_4	1953	2205	2207	2207	2207	2515	2572
TX_1	2408	2574	2574	2574	2574	3272	3565
TX_2	3428	3449	3449	3449	3449	4394	4515
TX_3	2709	2692	2692	2692	2692	3339	3501
WI_1	2383	2234	2133	2140	2140	2313	2413
WI_2	3216	3163	3163	3163	3163	3549	3635
WI_3	2349	2536	2536	2536	2536	3172	3196
Cenário 2							
CA_1	6113	4422	4161	4218	4218	5322	5569
CA_2	6082	4006	4088	4023	4088	5519	5664
CA_3	7721	6110	5827	5675	5634	6860	7413
CA_4	3271	2592	2707	2714	2701	3199	2927
TX_1	4033	3151	3002	2978	2978	3661	3841
TX_2	4292	3615	3528	3514	3514	4224	4307
TX_3	3957	3478	3330	3330	3330	3972	4067
WI_1	4874	3313	3218	3211	3211	4091	4190
WI_2	5127	5004	5004	5004	5004	5306	5383
WI_3	4325	3453	3341	3341	3341	4072	4225
Cenário 3							
CA_1	6113	4238	4215	4215	4215	5976	5989
CA_2	6082	3787	3789	3789	3789	5417	5714
CA_3	7721	5610	5528	5495	5495	6951	7382
CA_4	3271	2620	2634	2622	2634	3000	3073
TX_1	4033	3314	3353	3353	3353	4031	4086
TX_2	4292	3656	3196	3290	3429	4331	4451
TX_3	3957	3501	3483	3413	3449	3996	4270
WI_1	4874	3303	3212	3206	3212	4555	4587
WI_2	5127	5023	5016	5016	5016	5245	5459
WI_3	4325	3853	3826	3843	3840	4362	4350
Cenário 4							
CA_1	6113	3955	3668	3678	3678	4998	5269
CA_2	6082	3884	3628	3882	3628	5519	5837
CA_3	7721	5612	5181	6008	5350	6748	7223
CA_4	3271	2624	2580	2580	2580	3044	3075
TX_1	4033	3443	3357	3364	3364	3927	4023
TX_2	4292	3790	3726	3439	3452	4235	4344
TX_3	3957	3523	3541	3386	3787	3912	3857
WI_1	4874	3526	3397	3397	3397	4319	4280
WI_2	5127	5452	5420	5420	5420	5531	5530
WI_3	4325	3538	3657	3581	3581	4483	4594

Ao analisar os agregados, verifica-se que o agregado por loja apresenta as melhores métricas de erro, com MAE e RMSE mais baixos (365 e 452, respetivamente) em comparação com os agregados por estado e global. O MAPE e o MSE também são mais favoráveis para o agregado por loja, com valores de 8,854 e 233,918. Estes resultados indicam que o agregado por loja oferece previsões mais precisas e fiáveis. Porém, o agregado por estado também tem um desempenho robusto, especialmente na explicação da variabilidade dos dados.

Tabela A. 46: Comparação entre os modelos agregados com *XGBoost* (cenário 3)

Métrica	Agregado Global	Agregado por Estado	Agregado por Loja
R ²	0,5699	0,7675	0,7095
MAE	1139	1090	365
RMSE	1497	1300	452
MAPE	8,41%	8,19%	8,85%
MSE	2.380.610	1.781.653	233.918
Previsão	Agregado Global	Agregado por Estado	Agregado por Loja
Dia 1	49795	49795	49795
Dia 2	43302	40311	38905
Dia 3	42224	39478	38252
Dia 4	41718	39486	38242
Dia 5	41718	39486	38432
Dia 6	48611	49261	47864
Dia 7	54832	50195	49361

A experimentação com diferentes cenários e combinações de hiperparâmetros não só aprimora a precisão dos modelos, como também oferece uma compreensão mais profunda sobre a influência das variáveis e a estrutura temporal dos dados no desempenho das previsões. Cada cenário fornece percepções valiosas sobre como ajustar a abordagem de modelagem conforme a disponibilidade e o horizonte temporal dos dados, otimizando assim a eficácia das previsões. O gráfico de dispersão entre as vendas reais e previstas para o cenário 3 encontra-se disponível na Figura A.9.

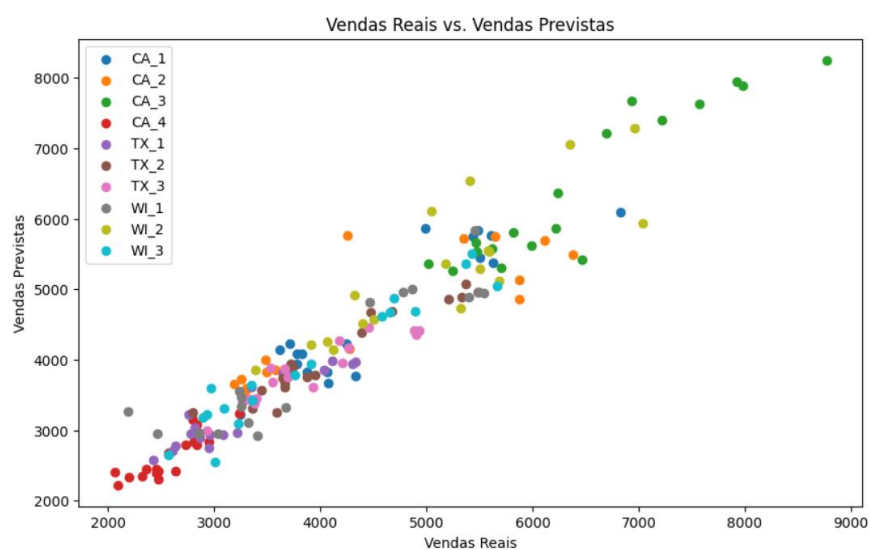


Figura A. 9: Gráfico de dispersão entre as vendas reais e previstas com *XGBoost*