**RELATÓRIO DE INICIAÇÃO CIENTÍFICA**

(adaptado da NBR10719/2011)

**RELATÓRIO DE INICIAÇÃO CIENTÍFICA**

**Título do Projeto:** Modelos Preditivos para Análise de Dados de Vendas

**Nome do aluno:** Jesua Isaque Calefi da Silva

**Curso:** Análise e Desenvolvimento de Sistemas

**Ano/semestre:** 2025/2

**Situação: [ ] Relatório Final**

**[ X ] Relatório Parcial**

**Período deste Relatório:** *(exemplo: agosto/2020 até dezembro/2020)*

**Carga Horária deste relatório:**

**Período do Projeto:** 01 de julho/2025 até 01 de dezembro/2025

**Carga Horária Total:** 264 horas

**Responsável/Coordenador:** Professor Mestre Marcio Rodrigues Sabino

**e-mail:** marcio.sabino@fatec.sp.gov.br **cel.:** (19) 98804-2681

**Categoria: [ ] BOLSISTA [ X ] NÃO BOLSISTA**

**CARGA HORÁRIA SEMANAL:  
Dias e Horários:** De segunda-feira a sábado das 16h00 às 18h00.

**PROPRIEDADE INTELECTUAL: [....] sim (aplica-se a): [ X ] não**

**[ ] inovação [ ] criação Potencial para Patente:**

**[ ] Invenção [ ] Modelo de Utilidade1.**

**Jesua Isaque Calefi da Silva**

**Modelo Preditivo para Análise de Vendas utilizando Random Forest**

Iniciação Científica apresentado ao Curso de Tecnologia em Análise e Desenvolvimento de Sistemas da Faculdade de Tecnologia de Mogi Mirim como pré-requisito para a obtenção do Título de Tecnólogo em Análise e Desenvolvimento de Sistemas.

**Orientador: Prof. Me. Marcio Rodrigues Sabino**

**MOGI MIRIM**

**2025**

**RESUMO**

A grande e crescente relevância da Inteligência Artificial está transformando a tomada de decisão em ambientes corporativos, nos quais modelos preditivos se tornaram ferramentas estratégicas para a análise de vendas. O presente trabalho tem como objetivo principal o desenvolvimento de um modelo preditivo para previsão de vendas, com foco na aplicação do algoritmo de *machine learning* Random Forest. A metodologia consiste em uma pesquisa aplicada e quantitativa, utilizando a linguagem de programação Python e suas bibliotecas para realizar a coleta, o pré-processamento e a análise exploratória de um conjunto de dados de vendas. O modelo será treinado sob a abordagem do aprendizado supervisionado, utilizando o algoritmo da Árvore de Decisão como base conceitual para o método de *ensemble* Random Forest, e avaliado por meio de métricas estatísticas de desempenho, como R² e RMSE. Espera-se, ao final, obter um modelo com alta capacidade preditiva e identificar as variáveis de maior impacto nas vendas, consolidando a aplicação prática dos conceitos de ciência de dados.

**Palavras-chave:** Modelo Preditivo. Análise de Vendas. Machine Learning. Random Forest.

**SUMÁRIO**

[1. INTRODUÇÃO 4](#_Toc204012216)

[2. OBJETIVOS 4](#_Toc204012217)

[2.1 Objetivos Gerais 4](#_Toc204012218)

[2.2 Objetivos Específicos 4](#_Toc204012219)

[3. REFERENCIAL TEÓRICO 6](#_Toc204012220)

[3.1 Inteligência Artificial e o Campo do Machine Learning 6](#_Toc204012221)

[3.2 O Algoritmo da Árvore de Decisão. 7](#_Toc204012222)

[3.2.1 Anatomia de uma Árvore de Decisão 7](#_Toc204012223)

[3.2.2 O Processo de Construção: Particionamento Recursivo 8](#_Toc204012224)

[3.2.3 Overfitting, Critérios de Parada e Poda 9](#_Toc204012225)

[3.3 Modelos Preditivos e o Método *Random Forest* para Previsão de Vendas 10](#_Toc204012226)

[3.4 Ciência de Dados 11](#_Toc204012227)

[4. METODOLOGIA 12](#_Toc204012228)

[4.1 Área e universo de estudo 13](#_Toc204012229)

[4.2 Fontes de coleta de dados. 13](#_Toc204012230)

[4.3 População e amostra, ou sujeitos da pesquisa 13](#_Toc204012231)

[4.4 Tipo e métodos da pesquisa. 13](#_Toc204012232)

[4.5 Procedimentos e técnicas da pesquisa 13](#_Toc204012233)

[4.6 Variáveis 14](#_Toc204012234)

[5. RESULTADOS E DISCUSSÕES (pode ser discriminado separadamente caso julgue necessário) 14](#_Toc204012235)

[6. CONCLUSÕES 14](#_Toc204012236)

[7. EXPECTATIVAS FUTURAS 14](#_Toc204012237)

[8. REFERÊNCIAS BIBLIOGRÁFICAS (NBR6023) 14](#_Toc204012238)

[9. APÊNDICE (Elemento opcional.) 14](#_Toc204012239)

[10. ANEXO (ELEMENTO OPCIONAL.) 14](#_Toc204012240)

# INTRODUÇÃO

Nos últimos anos, a inteligência artificial (IA) tem revolucionado a forma como as empresas analisam e tomam decisões com base em dados. A utilização de modelos preditivos tornou-se uma ferramenta estratégica para antecipar tendências, otimizar recursos e maximizar lucros, principalmente em setores altamente competitivos como o varejo. Segundo Lima (2020), “a análise preditiva possibilita às organizações compreender comportamentos futuros com base em dados históricos, criando uma vantagem competitiva essencial”.

No contexto brasileiro, o uso de técnicas de machine learning (aprendizado de máquina) tem se expandido, facilitando a aplicação de algoritmos como o Random Forest, que se destaca por sua robustez, acurácia e capacidade de lidar com variáveis complexas e dados ruidosos (Breiman, 2001; James et al., 2021).

O presente projeto visa desenvolver um modelo preditivo de vendas com aplicação prática, alinhando teoria e prática no campo da IA e introduzindo o aluno nos fundamentos da ciência de dados, aprendizado supervisionado e avaliação de modelos.

# OBJETIVOS

Nesta seção são apresentados os objetivos gerais e específicos que guiarão o desenvolvimento deste projeto.

## Objetivos Gerais

Desenvolver um modelo preditivo de vendas utilizando algoritmos de machine learning, com foco no método Random Forest, introduzindo os conhecimentos fundamentais de inteligência artificial, aprendizado supervisionado e modelos preditivos, com aplicação prática e análise crítica dos resultados.

## Objetivos Específicos

Como objetivos específicos, elencam-se:

* Coletar e tratar um conjunto de dados reais ou simulados de vendas;
* Identificar as variáveis mais relevantes para o processo preditivo;
* Aplicar técnicas de pré-processamento e análise exploratória de dados;
* Treinar modelos preditivos supervisionados, com ênfase no *Random Forest*;
* Avaliar o desempenho do modelo com métricas estatísticas como RMSE, MAE ou R²;
* Apresentar visualizações gráficas que interpretem os resultados obtidos;
* Produzir relatório final com discussões, conclusões e recomendações.

# REFERENCIAL TEÓRICO

Este referencial teórico apresenta os conceitos fundamentais para a construção do modelo preditivo. Serão abordados, em sequência, os temas de Inteligência Artificial (IA) e *Machine Learning*, o algoritmo da Árvore de Decisão e o método Random Forest, que é a técnica central do projeto. Por fim, contextualiza-se o campo da Ciência de Dados, que engloba as abordagens anteriores.

## Inteligência Artificial e o Campo do Machine Learning

A busca por criar sistemas capazes de simular a capacidade humana de raciocinar e aprender deu origem à Inteligência Artificial (IA). Este campo da ciência da computação não se resume a robôs humanoides, como o imaginário popular sugere, mas a sistemas que podem agir e pensar racionalmente. De acordo com Russell e Norvig (2013), a IA pode ser vista sob quatro abordagens: sistemas que pensam como humanos, sistemas que agem como humanos, sistemas que pensam racionalmente e sistemas que agem racionalmente. Para o contexto de análise de dados e tomada de decisão, a abordagem de "agir racionalmente" é a mais pertinente, pois se concentra na criação de agentes que tomam a melhor ação possível para atingir um objetivo.

A Inteligência Artificial (IA) é um campo vasto que engloba desde a automação de processos até a criação de sistemas capazes de aprender e tomar decisões. Um agente racional é aquele que age para alcançar o melhor resultado ou, quando há incerteza, o melhor resultado esperado (RUSSELL; NORVIG, 2013).

Dentro do grande campo da IA, surge o Machine Learning (Aprendizado de Máquina) como uma abordagem essencial para criar sistemas inteligentes sem que sejam explicitamente programados para cada tarefa. A ideia central é que os algoritmos possam "aprender" padrões diretamente dos dados. Tom Mitchell, um dos pioneiros na área, oferece uma definição clássica e precisa:

Diz-se que um programa de computador aprende da experiência E com respeito a alguma classe de tarefas T e medida de desempenho P, se seu desempenho em tarefas em T, medido por P, melhora com a experiência E (MITCHELL, 1997, p. 2, tradução nossa).

No contexto deste trabalho, a tarefa (T) é prever o volume de vendas. A experiência (E) é o conjunto de dados históricos de vendas, que inclui variáveis como preço, promoções, data etc. O desempenho (P) será medido por métricas estatísticas que avaliam a acurácia das previsões do modelo, como o Erro Quadrático Médio (RMSE).

O aprendizado de máquina é habitualmente dividido em três categorias: supervisionado, não supervisionado e por reforço. Este projeto se enquadra no aprendizado supervisionado, pois utilizará um conjunto de dados rotulados (dados históricos onde o volume de vendas já é conhecido) para treinar um modelo capaz de fazer previsões sobre dados futuros e não vistos.

## O Algoritmo da Árvore de Decisão.

O algoritmo da Árvore de Decisão é um dos métodos de aprendizado supervisionado mais intuitivos e poderosos. Sua estrutura se assemelha a um fluxograma invertido, onde cada nó representa uma decisão sobre um atributo, cada ramo denota um resultado dessa decisão, e cada nó folha representa o resultado final (seja uma classe ou um valor numérico). Por sua natureza transparente, é frequentemente classificado como um modelo "caixa-branca" (*white-box*), pois a lógica que leva a uma previsão pode ser facilmente visualizada e compreendida por humanos (HAN; KAMBER; PEI, 2012).

### Anatomia de uma Árvore de Decisão

Uma árvore de decisão é composta por elementos específicos que definem seu fluxo e estrutura:

* **Nó Raiz:** É o nó inicial no topo da árvore, representando o conjunto completo de dados antes de qualquer divisão.
* **Nó de Decisão:** É um nó interno que se ramifica em outros nós. Ele representa um teste ou uma "pergunta" sobre o valor de um determinado atributo.
* **Ramo ou Aresta:** Conecta um nó a outro, representando o resultado do teste no nó de decisão. Por exemplo, para um atributo "Salário", os ramos poderiam ser " < R$ 5.000" e " >= R$ 5.000".
* **Nó Folha:** São os nós finais da árvore que não se dividem mais. Eles contêm a predição final do modelo. Em uma tarefa de classificação, a folha contém a classe majoritária; em uma tarefa de regressão, contém um valor numérico, como a média.

### O Processo de Construção: Particionamento Recursivo

A construção de uma árvore de decisão ocorre por meio de um processo chamado particionamento recursivo. É um algoritmo de abordagem *greedy* (gulosa), o que significa que, em cada etapa, ele busca a melhor divisão possível naquele momento, sem considerar o impacto global dessa decisão nos passos futuros (JAMES et al., 2021). O processo se inicia no nó raiz com todos os dados e segue os seguintes passos:

1. Para cada atributo disponível, o algoritmo avalia todas as divisões possíveis.
2. Ele seleciona o atributo e o ponto de corte que resultam na "melhor" divisão, ou seja, que criam os subgrupos mais homogêneos possíveis.
3. O conjunto de dados é dividido nos subgrupos (nós filhos) com base nessa regra.
4. O processo é repetido recursivamente para cada nó filho até que um critério de parada seja atingido.

A definição de "melhor" divisão depende da natureza da tarefa (classificação ou regressão).

**Para Árvores de Classificação**, o objetivo é maximizar a pureza dos nós.

As métricas mais comuns para isso são:

* **Índice Gini:** Mede a probabilidade de um elemento, escolhido aleatoriamente no nó, ser classificado incorretamente. Um valor de 0 indica pureza total (todos os elementos pertencem à mesma classe).
* **Entropia:** Originária da teoria da informação, a entropia mede o grau de desordem ou incerteza em um nó. Um valor de 0 também indica pureza total. A divisão escolhida é aquela que proporciona o maior Ganho de Informação, que é a redução na entropia após a divisão.

**Para Árvores de Regressão**, como a utilizada neste projeto, o objetivo não é separar classes, mas sim criar grupos com valores numéricos o mais próximos possível. A métrica padrão para isso é a Redução do Erro Quadrático Médio (MSE). O algoritmo busca a divisão que minimiza a soma dos erros quadrados entre os pontos de dados e a média de seu respectivo nó. Em outras palavras, ele tenta minimizar a variância dentro dos nós filhos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

### Overfitting, Critérios de Parada e Poda

Uma árvore pode teoricamente crescer até que cada folha contenha uma única observação. No entanto, uma árvore tão complexa se ajustaria perfeitamente ao ruído dos dados de treinamento, um fenômeno conhecido como superajuste (*overfitting*). Um modelo superajustado tem um desempenho excelente nos dados que já viu, mas falha drasticamente ao generalizar para novos dados.

Para combater o overfitting, utilizam-se técnicas de pré e pós-poda:

* **Critérios de Parada (Pré-poda):** Impede o crescimento excessivo da árvore, definindo limites. Os hiperparâmetros mais comuns são a profundidade máxima da árvore (max\_depth), o número mínimo de amostras para dividir um nó (min\_samples\_split) e o número mínimo de amostras em um nó folha (min\_samples\_leaf).
* **Poda (Pós-poda):** Consiste em deixar a árvore crescer completamente e, em seguida, remover ramos que fornecem pouco poder preditivo, substituindo-os por uma folha. Esta técnica pode levar a resultados mais robustos, embora seja computacionalmente mais custosa (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A principal desvantagem de uma única árvore de decisão, sua instabilidade e tendência ao *overfitting*, foi o que motivou o desenvolvimento de métodos de *ensemble* como o Random Forest, que será abordado na seção seguinte.

## Modelos Preditivos e o Método *Random Forest* para Previsão de Vendas

A previsão de vendas é uma aplicação clássica de modelos preditivos e de extrema importância para a estratégia de negócios de uma empresa. Previsões minuciosas permitem otimizar a gestão de estoque, o planejamento de campanhas de marketing, a atribuição de recursos financeiros e a logística de distribuição. Do ponto de vista técnico, prever um valor contínuo, como o volume de vendas, é um problema de regressão, uma das principais tarefas do aprendizado supervisionado.

Existem diversos algoritmos de regressão, desde os mais simples, como a Regressão Linear, até os mais complexos. Para este projeto, será utilizado o método Random Forest (Floresta Aleatória), devido à sua alta performance, robustez e capacidade de capturar relações não-lineares complexas nos dados, o que é comum em cenários de vendas.

O algoritmo *Random Forest*, proposto por Leo Breiman (2001), funciona através da construção de múltiplas árvores de decisão durante o treinamento. Uma árvore de decisão é um modelo que utiliza uma série de "perguntas" (ou regras) sobre as variáveis de entrada para chegar a uma decisão de previsão. O problema de uma única árvore de decisão é sua tendência ao superajuste (*overfitting*), ou seja, ela se ajusta tão bem aos dados de treino que perde a capacidade de generalizar para novos dados.

O *Random Forest* supera essa limitação ao criar um "comitê" de árvores. Cada árvore é treinada com uma amostra aleatória dos dados de treino e, em cada nó da árvore, apenas um subconjunto aleatório de variáveis é considerado para a divisão. Para uma tarefa de regressão, a previsão final do modelo é a média das previsões de todas as árvores individuais que compõem a floresta (JAMES et al., 2021).

Essa abordagem confere ao *Random Forest* vantagens significativas para a previsão de vendas:

* Alta Acurácia: Geralmente apresenta um desempenho preditivo superior a muitos outros algoritmos.
* Robustez a *Outliers* (valores atípicos) e Ruídos: A natureza de agregação de múltiplas árvores torna o modelo menos sensível a dados discrepantes.
* Análise de Importância das Variáveis: O modelo permite calcular a "importância" de cada variável (exemplo: preço, dia da semana, investimento em marketing), ajudando a entender quais fatores mais influenciam as vendas.

A aplicação do *Random Forest* permitirá, portanto, não apenas desenvolver um modelo para prever o volume de vendas de um produto, mas também fornecer *insights* valiosos sobre os principais impulsionadores dessas vendas, validando a hipótese de que algoritmos de *machine learning* são ferramentas eficazes para a análise de dados comerciais.

## Ciência de Dados

Segundo Provost e Fawcett (2013), a ciência de dados fornece um processo sistemático para a extração de conhecimento prático a partir de grandes volumes de dados, permitindo transformações estratégicas nas organizações. Trata-se de uma área multidisciplinar que integra estatística, computação, matemática e conhecimento do domínio específico, com o objetivo de extrair valor de conjuntos de dados complexos, estruturados ou não estruturados. A crescente digitalização de processos e a coleta massiva de informações tornaram a ciência de dados uma peça-chave na inteligência empresarial, contribuindo para a tomada de decisões baseadas em evidências.

Dentro da ciência de dados, destacam-se tarefas como mineração de dados, análise preditiva, visualização e comunicação de resultados, todas apoiadas por ferramentas computacionais avançadas. Segundo Oliveira e Lima (2021), o diferencial competitivo proporcionado pela análise de dados reside na capacidade de identificar padrões e antecipar comportamentos de mercado. Isso se aplica de forma direta ao contexto de vendas, onde modelos preditivos permitem prever a demanda futura com base em dados históricos, otimizando estoques, logística e estratégias de marketing.

Além disso, a ciência de dados não se limita ao tratamento técnico dos dados, mas exige também uma compreensão crítica sobre a qualidade das fontes, o viés dos dados e a interpretação adequada dos resultados. Conforme aponta Monteiro (2020), o cientista de dados deve ser capaz de formular perguntas relevantes, manipular dados com rigor, selecionar os algoritmos adequados e comunicar descobertas de forma clara e impactante. Essa combinação de competências técnicas e analíticas torna a formação nesta área um desafio relevante, mas também uma oportunidade estratégica para profissionais e estudantes.

Neste projeto, a ciência de dados será a base metodológica e conceitual para a construção do modelo de previsão de vendas. Por meio do uso de algoritmos como o *Random Forest*, aliado a técnicas estatísticas e à visualização de dados, busca-se não apenas gerar previsões, mas também compreender os fatores que influenciam os resultados de forma mais significativa. Assim, o projeto se alinha às práticas atuais do mercado e da pesquisa aplicada, ao mesmo tempo em que introduz o aluno às etapas essenciais do ciclo de vida de projetos em ciência de dados.

# METODOLOGIA

A metodologia deste projeto será desenvolvida a partir da aplicação prática de técnicas de machine learning para construção de um modelo preditivo voltado à análise de vendas. O ponto de partida será a obtenção e tratamento de um conjunto de dados históricos — reais ou públicos — contendo variáveis como volume de vendas, preços, datas, promoções e outras informações relevantes. Com o uso da linguagem de programação Python e bibliotecas como Pandas, Scikit-learn e Matplotlib, serão realizadas etapas de pré-processamento, análise exploratória e modelagem supervisionada, utilizando o algoritmo Random Forest como principal abordagem, dada sua robustez e ampla aceitação na literatura especializada. Não serão utilizados equipamentos físicos, sendo todos os testes conduzidos em ambiente computacional local com ferramentas de código aberto. O processo metodológico segue fundamentos teóricos consagrados da ciência de dados e da estatística aplicada, e busca conciliar rigor acadêmico com aplicabilidade prática, permitindo ao aluno compreender desde a organização dos dados até a avaliação e interpretação dos resultados preditivos obtidos.

## Área e universo de estudo

A pesquisa será conduzida em ambiente acadêmico, com aplicação simulada de um modelo preditivo utilizando dados de vendas fictícios ou publicamente disponíveis (ex: *datasets* do Kaggle). O universo simulado representa cenários reais de empresas do setor varejista.

## Fontes de coleta de dados.

Conjuntos de dados históricos de vendas contendo variáveis como preço, data, promoções, dia da semana, campanhas, entre outras.

## População e amostra, ou sujeitos da pesquisa

A população é composta pelos registros de vendas contidos no *dataset* selecionado. A amostra será definida conforme os filtros e segmentações aplicadas durante a análise exploratória.

## Tipo e métodos da pesquisa.

Trata-se de uma pesquisa aplicada e quantitativa, com abordagem experimental e exploratória, por meio da modelagem computacional dos dados.

## Procedimentos e técnicas da pesquisa

* Coleta e organização dos dados em *Python*;
* Análise exploratória com bibliotecas como Pandas, *Matplotlib* e *Seaborn*;
* Aplicação do modelo *Random Forest* com *Scikit-Learn*;
* Avaliação dos resultados com métricas estatísticas;
* Análise dos atributos com maior importância para a previsão.

## Variáveis

* Variáveis independentes: preço, dia da semana, promoções etc.
* Variável dependente: volume de vendas.
* Fatores que podem afetar o modelo: dados ausentes, multicolinearidade, *overfitting*.

# RESULTADOS E DISCUSSÕES (pode ser discriminado separadamente caso julgue necessário)

# CONCLUSÕES

# EXPECTATIVAS FUTURAS

# REFERÊNCIAS BIBLIOGRÁFICAS (NBR6023)

# APÊNDICE (Elemento opcional.)

# ANEXO (ELEMENTO OPCIONAL.)

**ANTES DE IMPRIMIR/ENVIAR, RETIRE TODOS OS MARCA TEXTOS AMARELOS E TEXTOS EXPLICATIVOS.**