**RELATÓRIO DE INICIAÇÃO CIENTÍFICA**

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

**Aluno:** Jesua Isaque Calefi da Silva

**Curso:** Análise e Desenvolvimento de Sistemas

**Ano/semestre:** 2025/2

**Situação: [ X ] Relatório Final**

**[ ] Relatório Parcial**

**Período deste Relatório:** 01 de julho/2025 até 01 de dezembro/2025

**Carga Horária deste relatório:** 12 horas semanais – Total de 264 horas

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

**Carga Horária Total:** 12 horas semanais – Total de 264 horas

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

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

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

**Aluno:** Jesua Isaque Calefi da Silva

**e-mail:** jesua.silva@fatec.sp.gov.br **cel.:** (19) 98202-3894

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

**Propriedade Intelectual:**

**[ ] Sim, caráter de inovação.**

**[ ] Sim, caráter de criação.**

**[ X ] Não.**

**Potencial para Patente:**

**[ ] Sim, caráter invenção.**

**[ ] Sim, caráter modelo de utilidade.**

**[ X ] Não.**

**[ ] Sim, caráter modelo de utilidade.**

**[ X ] Não.**

CENTRO DE EDUCAÇÃO ESTADUAL TECNOLÓGICA “PAULA SOUZA”

FACULDADE DE TECNOLOGIA DE MOGI MIRIM

FATEC “ARTHUR DE AZEVEDO”

CURSO SUPERIOR DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

**Modelos Preditivos para Análise de Dados de Vendas**

JESUA ISAQUE CALEFI DA SILVA

**MOGI MIRIM**

**Dezembro/2025**

**SUMÁRIO**

[1. INTRODUÇÃO 4](#_Toc210900834)

[1.1 O Problema 4](#_Toc210900835)

[1.2 Justificativa 5](#_Toc210900836)

[1.3 Objetivos 5](#_Toc210900837)

[1.3.1 Objetivos Gerais 5](#_Toc210900838)

[1.3.2 Objetivos Específicos 6](#_Toc210900839)

[1.4 Hipóteses 6](#_Toc210900840)

[2. METODOLOGIA 7](#_Toc210900841)

[2.1 Área e universo de estudo 7](#_Toc210900842)

[2.2 Fontes de coleta de dados. 7](#_Toc210900843)

[2.3 População e amostra, ou sujeitos da pesquisa 8](#_Toc210900844)

[2.4 Tipo e métodos da pesquisa. 8](#_Toc210900845)

[2.5 Procedimentos e técnicas da pesquisa 8](#_Toc210900846)

[2.6 Variáveis 10](#_Toc210900847)

[2.7 Orçamento 10](#_Toc210900848)

[2.8 Cronograma 10](#_Toc210900849)

[3. REFERENCIAL TEÓRICO 12](#_Toc210900850)

[3.1 Inteligência Artificial e o Campo do Machine Learning 12](#_Toc210900851)

[3.2 O Algoritmo da Árvore de Decisão 13](#_Toc210900852)

[3.2.1 Anatomia de uma Árvore de Decisão 13](#_Toc210900853)

[3.2.2 O Processo de Construção: Particionamento Recursivo 14](#_Toc210900854)

[3.2.3 *Overfitting*, Critérios de Parada e Poda 15](#_Toc210900855)

[3.3 Modelos Preditivos e o Método *Random Forest* para Previsão de Vendas 16](#_Toc210900856)

[3.3.1 A Força dos Métodos Ensemble 17](#_Toc210900857)

[3.3.2 Medindo o Impacto: A Importância das Variáveis 18](#_Toc210900858)

[3.4 Ciência de Dados 19](#_Toc210900859)

[4. DESENVOLVIMENTO 20](#_Toc210900860)

[5. RESULTADOS E DISCUSSÕES 21](#_Toc210900861)

[6. DISCUSSÕES FINAIS 22](#_Toc210900862)

[REFERÊNCIAS 23](#_Toc210900863)

[APÊNDICE 24](#_Toc210900864)

[ANEXO 24](#_Toc210900865)

# 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 visou o desenvolvimento de 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.

## O Problema

Empresas dos mais variados setores enfrentam dificuldades recorrentes na previsão precisa de vendas, o que compromete diretamente a eficiência de suas operações. A ausência de previsões confiáveis impacta negativamente o planejamento de estoques, a alocação de recursos, a programação de produção e as campanhas de marketing. Por exemplo, uma superestimação da demanda pode levar a excesso de estoque, custos de armazenagem e perdas por obsolescência, enquanto a subestimação pode resultar em ruptura de estoque, perda de vendas e insatisfação do cliente.

No setor varejista, esse desafio se intensifica devido à sazonalidade, variações no comportamento do consumidor e influência de fatores externos, como promoções ou eventos econômicos. Um estudo de Simchi-Levi, Kaminsky e Simchi-Levi (2008) destaca que falhas em previsões de demanda podem gerar ineficiências logísticas de até 30%, afetando significativamente a lucratividade das organizações.

Diante desse cenário, surge a seguinte questão central: é possível utilizar algoritmos de aprendizado de máquina, como o *Random Forest*, para melhorar significativamente a acurácia das previsões de vendas com base em dados históricos?

A busca por essa resposta justifica a proposta deste projeto, que visa aliar fundamentos técnicos da ciência de dados à resolução de um problema concreto de relevância mercadológica.

## Justificativa

Com o crescimento dos dados digitais e a competitividade empresarial, torna-se essencial o desenvolvimento de técnicas eficientes para previsão e tomada de decisão. A aplicação de modelos como o *Random Forest* permite não apenas gerar previsões mais precisas, mas também identificar os fatores que mais impactam nas vendas.

Além disso, o projeto justifica-se como ferramenta de iniciação científica, contribuindo para a formação técnica e acadêmica do aluno, que será inserido nos conceitos e práticas da inteligência artificial e do aprendizado de máquina, desenvolvendo competências valiosas no mercado de trabalho e em futuras pesquisas.

"Escrever cientificamente é mais do que comunicar resultados — é organizar o raciocínio, argumentar com clareza e permitir a reprodutibilidade do conhecimento produzido."(MARCONI; LAKATOS, 2010, p. 30)

## Objetivos

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

### Objetivos Gerais

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

### Objetivos Específicos

Como objetivos específicos, elencaram-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.

## Hipóteses

Com a obtenção dos modelos de *machine learning* aplicados corretamente a dados históricos de vendas, possibilitar o usuário obter previsões com alta acurácia e confiabilidade, otimizando o processo decisório em ambientes empresariais.

# METODOLOGIA

A metodologia deste projeto foi 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 foi 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, foram 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 foram utilizados equipamentos físicos, sendo todos os testes conduzidos em ambiente computacional local com ferramentas de código aberto. O processo metodológico seguiu fundamentos teóricos consagrados da ciência de dados e da estatística aplicada, e buscou 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 foi 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 foi 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.

Para garantir que a avaliação do modelo seja robusta e que seu desempenho não seja resultado de uma divisão favorável entre treino e teste, será empregada uma técnica de validação apropriada para séries temporais. Diferente da validação cruzada tradicional, que embaralha os dados aleatoriamente, a análise de dados temporais exige que a ordem cronológica seja preservada. Para isso, será utilizada a abordagem de Validação Cruzada *Walk-Forward*, na qual o modelo é treinado em um período inicial de dados e validado no período imediatamente seguinte. Progressivamente, a janela de treinamento é expandida para incorporar os dados de validação, e o modelo é testado em um novo período futuro. Este processo simula de forma mais realista como o modelo seria utilizado na prática: fazendo previsões para o futuro com base em todo o histórico disponível até o momento.

A performance de um modelo de Random Forest é sensível a um conjunto de configurações internas conhecidas como hiperparâmetros, como o número de árvores na floresta (*n\_estimators*), a profundidade máxima de cada árvore (*max*\_*depth*) e o número mínimo de amostras em um nó folha (*min*\_*samples*\_*leaf*). Para encontrar a combinação ótima desses hiperparâmetros que maximize a performance do modelo para este conjunto de dados específico, será utilizada a técnica de *Randomized Search Cross-Validation.* Este método explora de forma eficiente diferentes combinações de hiperparâmetros a partir de um espaço de busca pré-definido, utilizando a validação cruzada para avaliar cada combinação. O objetivo é identificar os hiperparâmetros que minimizem o erro de previsão (MAE ou RMSE), resultando em um modelo final mais preciso e generalizável.

## Variáveis

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

## Orçamento

Como se trata de projeto sem financiamento (não bolsista), os custos são reduzidos, limitando-se a:

* Acesso à internet: R$ 0 (estrutura pessoal do aluno ou da instituição)
* Computador pessoal ou da instituição: R$ 0
* *Softwares* e bibliotecas: R$ 0 (todos gratuitos/*open source*)

Total estimado: R$ 0,00

## Cronograma

O desenvolvimento e execução do projeto foi dentro do campus da Fatec Arthur de Azevedo de Mogi Mirim ao longo dos dias comerciais. Aos sábados o aluno realizou os estudos em sua residência.

O projeto foi executado entre 1º de julho e 1º de dezembro de 2025, com 12 horas semanais de dedicação, incluindo 1 hora semanal de reunião com o orientador para discutir resultados, receber instruções e alinhar os próximos passos. O cronograma de estudos e desenvolvimento do projeto está presente no quadro 1.

**Quadro 1:** Cronograma de desenvolvimento do projeto.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Mês** | **07** | **08** | **09** | **10** | **11** | **12** |
| **Descrição** |  |  |  |  |  |  |
| Desenvolvimento do projeto e envio à CEPE |  |  |  |  |  |  |
| Estudo teórico |  |  |  |  |  |  |
| Coleta dos dados, tratamento e análises iniciais |  |  |  |  |  |  |
| Construção do modelo preditívo |  |  |  |  |  |  |
| Testes e ajustes do modelo |  |  |  |  |  |  |
| Redação do relatório e revisão final |  |  |  |  |  |  |
| Entrega do relatório e apresentação |  |  |  |  |  |  |

**Fonte:** Próprio Autor.

# REFERENCIAL TEÓRICO

Para a fundamentação deste projeto, o referencial teórico será dividido em dois tópicos principais. O primeiro abordará os conceitos fundamentais de Inteligência Artificial e *Machine Learning*, estabelecendo a base conceitual da área de estudo. O segundo tópico se aprofundará nos modelos preditivos, com foco em regressão para previsão de vendas, e apresentará o método de *Random Forest*, recomendado para a aplicação prática deste trabalho.

## 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.

### A Força dos Métodos Ensemble

O poder preditivo do Random Forest não reside em uma única árvore de decisão, mas na aplicação de uma técnica poderosa conhecida como ensemble learning (aprendizagem de conjunto). A premissa fundamental dos métodos ensemble é a de que a combinação das previsões de múltiplos modelos tende a produzir um resultado mais preciso e robusto do que qualquer um dos modelos individuais poderia alcançar, uma manifestação da "sabedoria da multidão" aplicada ao aprendizado de máquina (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

O Random Forest é um exemplo da técnica de *ensemble* chamada *Bagging*, um acrônimo para *Bootstrap Aggregating*. O processo de *bagging* consiste em duas etapas principais:

1. ***Bootstrap*:** Múltiplas amostras do conjunto de dados de treinamento original são criadas por meio de amostragem com reposição. Isso significa que cada nova amostra tem o mesmo tamanho da original, mas algumas observações podem se repetir enquanto outras podem ser omitidas. O resultado é um conjunto de amostras de treinamento ligeiramente diferentes entre si.
2. ***Aggregating*:** Um modelo de Árvore de Decisão é treinado de forma independente em cada uma dessas amostras de dados. Para uma tarefa de regressão, como a previsão de vendas, a "agregação" final é feita calculando a média das previsões de todas as árvores individuais.

Essa abordagem diminui drasticamente a variância do modelo e reduz o risco de overfitting, que é a principal desvantagem de uma única árvore de decisão. Ao treinar cada árvore com uma amostra de dados e um subconjunto de variáveis diferentes, o bagging garante que as árvores sejam descorrelacionadas, tornando a previsão final mais estável e generalizável (JAMES et al., 2021).

### Medindo o Impacto: A Importância das Variáveis

Além de sua acurácia, uma das vantagens mais significativas do Random Forest é sua capacidade de fornecer uma medida da importância de cada variável (*feature importance*) para a previsão. Compreender quais fatores mais influenciam as vendas é um *insight* estratégico valioso para qualquer negócio. Existem duas abordagens principais para calcular essa importância:

1. **Redução Média da Impureza (*Mean Decrease in Impurity*):** Esta é a abordagem mais comum, implementada pela biblioteca Scikit-Learn. Durante a construção das árvores, sempre que um nó é dividido, a divisão é feita na variável que mais reduz a "impureza" do nó. Para a regressão, essa impureza é tipicamente o Erro Quadrático Médio (MSE). A importância de uma variável é calculada como a média da redução de impureza que ela proporciona em todas as árvores da floresta. Embora seja rápido de calcular, esse método pode ser enviesado, favorecendo variáveis com alta cardinalidade (muitos valores únicos).
2. **Importância por Permutação (*Permutation Importance*):** Considerada uma técnica mais robusta, a importância por permutação é calculada após o treinamento completo do modelo. O processo consiste em, para cada variável: (a) medir o desempenho do modelo no conjunto de validação; (b) embaralhar aleatoriamente os valores apenas daquela variável, mantendo as outras intactas; (c) medir novamente o desempenho do modelo. A queda na performance do modelo após o embaralhamento indica o quão dependente o modelo era daquela variável. Uma queda acentuada significa que a variável é muito importante.

A análise dessas métricas permite não apenas validar o modelo, mas também extrair conhecimento acionável do negócio, alinhando os objetivos técnicos da ciência de dados com os objetivos estratégicos da organização.

## 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.

# DESENVOLVIMENTO

AQUI VC VAI TER QUE FALAR SOBRE A MODELAGEM REALIZADA

APRESENTAR DETALHADAMENTE AS HIPÓTESES

APRESENTAR O MODELO

APRESENTAR A IMPLEMENTAÇÃO

# RESULTADOS E DISCUSSÕES

AQUI VC VAI TER QUE APRESENTAR OS RESULTADOS OBTIDOS AO RODAR O MODELO

DISCUTIR OS RESULTADOS OBTIDOS

# DISCUSSÕES FINAIS

AQUI VC VAI TER QUE RETOMAR A PROBLEMÁTICA< DIZER O QUE UTILIZOU PARA RESOLVER O PROBLEMA E RESUMIR OS RESULTADOS. APRESENTAR TRABALHOS FUTUROS

# REFERÊNCIAS

BREIMAN, Leo. **Random Forests. Machine Learning**, v. 45, n. 1, p. 5-32, 2001.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining: Concepts and Techniques.** 3. ed. Waltham: Morgan Kaufmann, 2012.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. New York: Springer, 2009.

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. **An Introduction to Statistical Learning: with Applications in R**. 2. ed. New York: Springer, 2021.

LIMA, F. T. **Inteligência artificial no varejo brasileiro**. São Paulo: Atlas, 2020.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Metodologia científica**. 6. ed. São Paulo: Atlas, 2010.

MITCHELL, Tom M. **Machine Learning.** New York: McGraw-Hill, 1997.

MONTEIRO, Danilo. **Cientista de dados: o profissional mais sexy do século XXI**. São Paulo: Casa do Código, 2020.

OLIVEIRA, T. S.; LIMA, J. R. **Ciência de dados aplicada: fundamentos, ferramentas e aplicações.** Rio de Janeiro: LTC, 2021.

PROVOST, Foster; FAWCETT, Tom. **Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking.** Sebastopol: O'Reilly Media, 2013.

RUSSELL, Stuart J.; NORVIG, Peter**. Inteligência Artificial.** 3. ed. Rio de Janeiro: Elsevier, 2013.

SIMCHI-LEVI, David; KAMINSKY, Philip; SIMCHI-LEVI, Edith. **Designing and Managing the Supply Chain: Concepts, Strategies and Case Studies**. 3. ed. New York: McGraw-Hill, 2008.

# APÊNDICE

Não se aplica neste momento.

# ANEXO

Não se aplica neste momento.