REDES NEURAIS ARTIFICIAIS E

DEEP LEARNING / T2025/S2

APRESENTAÇÃO

NOSSAS AULAS SERÃO COM ENCONTROS REMOTOS NAS SEGUINTES DATAS:

17/01, 31/01, 21/02 e 07/03 NO HORÁRIO DAS 8:00 ÁS 17:15; TEREMOS UM PERIODO DE ALMOÇO DE UMA HORA E DOIS INTERVALOS DE 15 MINUTOS.

O CONTEÚDO DAS AULAS SERÃO GRAVADOS E O MATERIAL DISPONIBILIZADO PELO CANVAS;

TODOS OS EXERCÍCIOS SERÃO REALIZADOS DURANTE O TEMPO DOS QUATRO ENCONTROS;

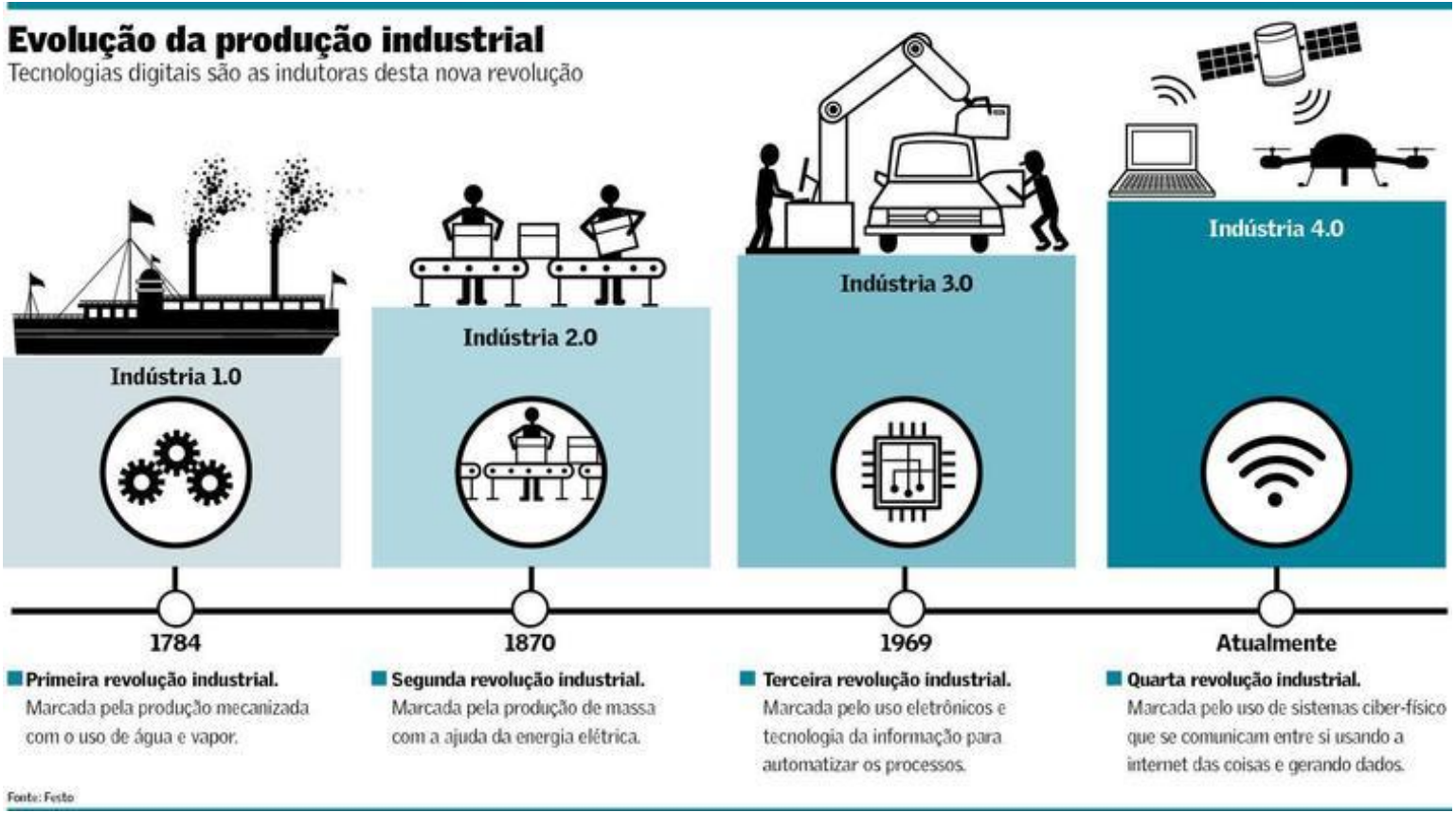
TODAS AS ENTREGAS DEVEM SER FEITAS PELO CANVAS;

ALÉM DOS EXERCÍCIOS HAVERÁ UMA ENTREGA FINAL.

REVISÃO

Teve como objetivo estabelecer a base conceitual, técnica e metodológica necessária para o desenvolvimento das atividades ao longo do curso, promovendo o alinhamento entre fundamentos de Inteligência Artificial, análise de dados e tomada de decisão organizacional baseada em evidências.

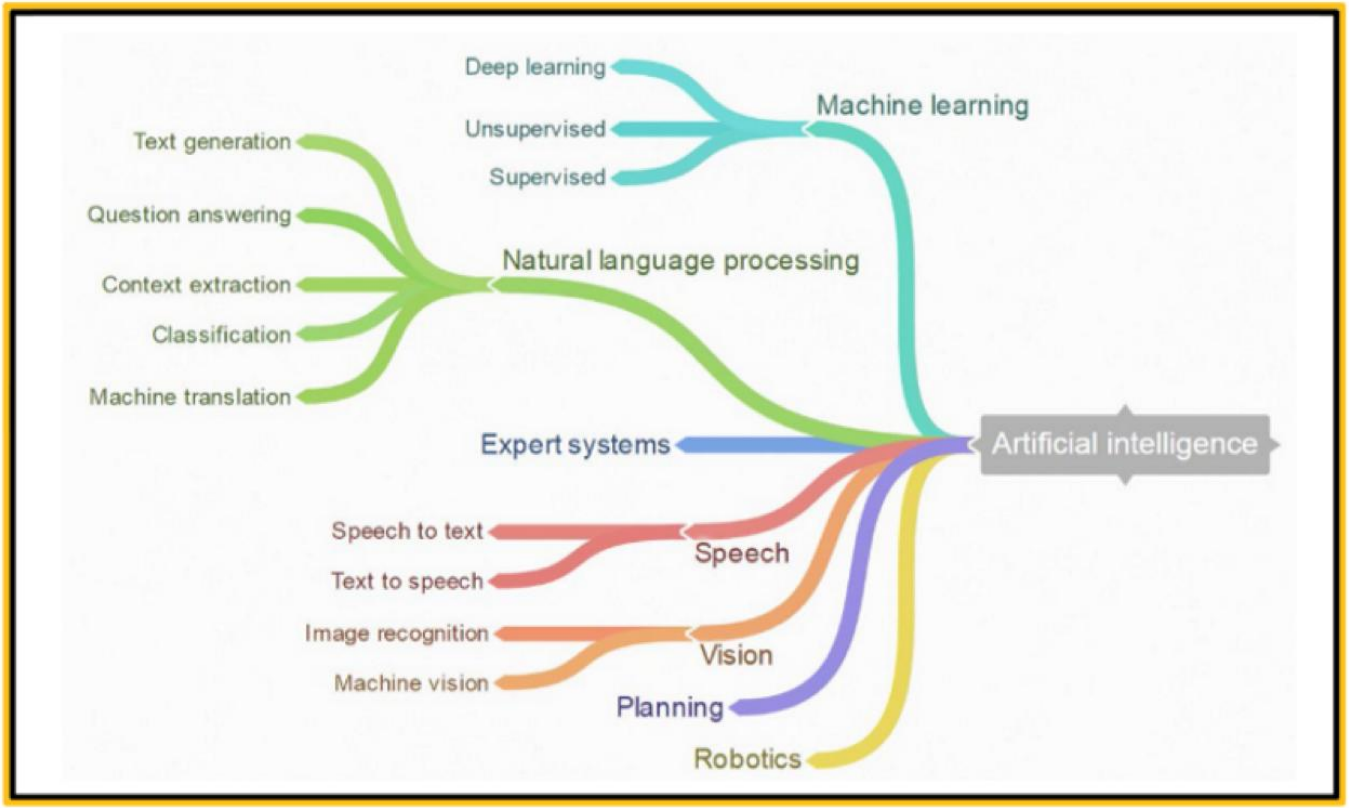
REVISÃO


Evolução da produção industrial
Tecnologias digitais são as indutoras desta nova revolução
Indústria 1.0
Indústria 2.0
Indústria 3.0
Indústria 4.0
1784
1870
1969
Atualmente
Primeira revolução industrial.
Marcada pela produção mecanizada com o uso de água e vapor.
Segunda revolução industrial.
Marcada pela produção de massa com a ajuda da energia elétrica.
Terceira revolução industrial.
Marcada pelo uso eletrônicos e tecnologia da informação para automatizar os processos.
Quarta revolução industrial.
Marcada pelo uso de sistemas ciber-físico que se comunicam entre si usando a internet das coisas e gerando dados.
Fonte: Festo

REVISÃO

Os conceitos introdutórios de Inteligência Artificial, Machine Learning e Deep Learning, destacando suas diferenças, aplicações e papéis no contexto organizacional contemporâneo. Discutiu-se a relevância do uso de dados como ativo estratégico, bem como a importância de uma abordagem analítica e crítica na aplicação de técnicas de IA em cenários reais.

REVISÃO



Abordagens para Inteligência Artificial (MARCONI, 2016).

REVISÃO

Otimização de Processos;

Previsão de Demanda;

Manutenção Preventiva.

EXEMPLO

Exemplo: Otimização do Controle de Qualidade em uma Linha de Produção de Automóveis

Imagine uma fábrica de automóveis que busca melhorar seu processo de controle de qualidade para garantir a excelência dos veículos produzidos. Neste caso, a inteligência artificial pode desempenhar um papel fundamental na otimização desse processo. O exemplo a seguir é uma dica para que você possa compreender melhor como funciona um processo de otimização com IA.

Coleta de Dados;

Análise de Dados com IA;

Tomada de Decisões Automatizada;

Feedback e Aprendizado Contínuo.

EXEMPLO

Coleta de Dados:

São instalados sensores em pontos estratégicos da linha de produção para coletar dados em tempo real sobre parâmetros como dimensões, peso, temperatura, vibração, entre outros. Os dados coletados são armazenados em um banco de dados centralizado para análise posterior.

Análise de Dados com IA:

Algoritmos de inteligência artificial, como redes neurais ou algoritmos de aprendizado de máquina, são aplicados aos dados coletados para identificar padrões, correlações e anomalias. Por exemplo, a IA pode detectar padrões sutis nas medições que indicam um desvio de tolerância nas dimensões de uma peça ou um aumento na vibração em um determinado componente da linha de montagem.

EXEMPLO

Tomada de Decisões Automatizada:

Com base na análise dos dados, a IA é capaz de tomar decisões automatizadas em tempo real para corrigir ou ajustar o processo de produção. Por exemplo, se a IA detectar que uma máquina está produzindo peças fora das especificações, ela pode acionar automaticamente um mecanismo de correção, como ajustar a configuração da máquina ou interromper temporariamente a produção até que o problema seja resolvido.

Feedback e Aprendizado Contínuo:

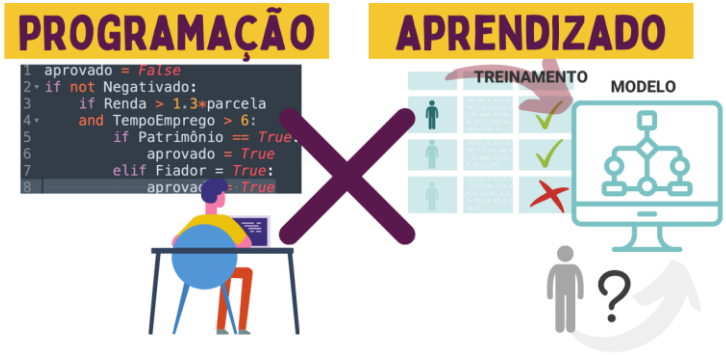
A IA é capaz de aprender com suas próprias decisões e correções ao longo do tempo, refinando continuamente seus modelos e algoritmos com base em novos dados e experiências. Por exemplo, se uma determinada ação corretiva não produzir os resultados esperados, a IA poderá ajustar sua abordagem para lidar melhor com situações semelhantes no futuro.

REVISÃO

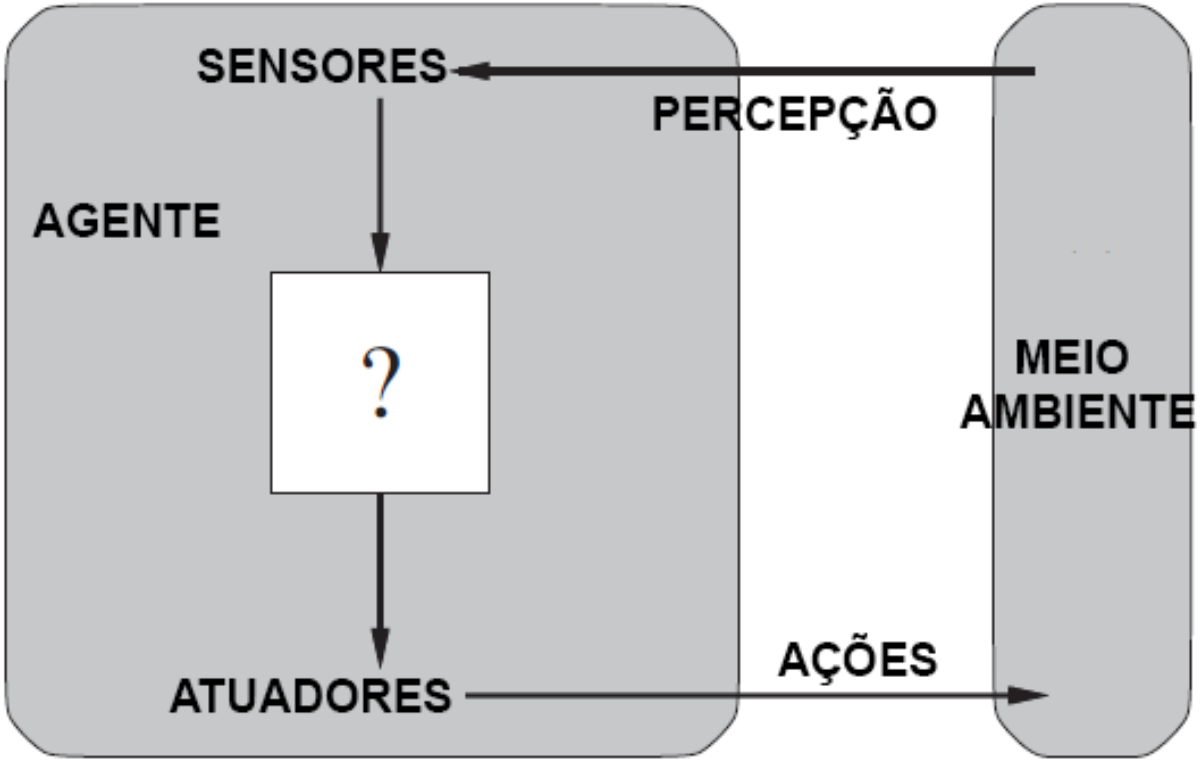
SCHLEDER, Gabriel R.; FAZZIO, Adalberto. Machine Learning na Física, Química, e Ciência de Materiais: Descoberta e Design de Materiais. Revista Brasileira de Ensino de Física, v. 43, n. Suppl 1, p. e20200407, 2021.

APRENDIZADO DE MÁQUINA

O aprendizado de máquina é um subcampo da inteligência artificial (IA) que se concentra no desenvolvimento de algoritmos e modelos computacionais capazes de aprender padrões e realizar tarefas específicas a partir de dados, sem a necessidade de instruções explícitas programadas por humanos. Em outras palavras, em vez de serem explicitamente programadas para realizar uma tarefa, as máquinas são treinadas usando grandes conjuntos de dados para reconhecer padrões e fazer previsões ou tomar decisões com base nesses padrões (MICHALSKI; CARBONELL; MITCHELL, 2013).



AGENTES REATIVOS



Agentes com reflexo simples. (RUSSEL; NORVIG, 2003).

BASE DE DADOS (DATASET)

Característica	Dados Estruturados	Dados Não Estruturados	Dados Semiestruturados
Natureza	Usualmente quantitativa	Usualmente qualitativa	Pode ser quantitativa e qualitativa
Modelo	Predefinido e é difícil alterá-lo	Modelo bem flexível	Tem flexibilidade, mas também possui estrutura
Formato	Número limitado de formato de dados	Grande variedade de formatos de dados	Diversa variedade de formatos
Banco de dados	Bancos baseados em SQL são utilizados	Bancos de dados NoSQL	Bancos de dados relacionais
Pesquisa	Fácil e rápido para localizar e pesquisar esses dados	Como não tem estruturas, é muito difícil procurar esses dados	Pesquisa difícil, mas não tanto quanto nos não estruturados
Análise	Fácil análise	Análise difícil	Análise difícil

MODELO

Paciente	Idade	IMC	Medicamentos	Quedas	Doenças Crônicas	Visitas ao Posto
1	85	22.5	3	1	2	5
2	78		5	0	1	3
3	90	26.7	4	2		7
4	72	29.4	2	-1	1	2
5	88	31.0	6	3	3	
6		27.3	3	0	2	4
7	82	24.8	4		1	6
8	91	33.5		2	2	9
9	150	28.6	3	0	2	3
10	86	25.4	4	2	3	6

MODELO

Tratar valores ausentes (quando fizer sentido)

Para variáveis numéricas (Idade, IMC, Visitas):

Média: se distribuição simétrica

Mediana: se assimétrica (mais comum em saúde)

Modelo preditivo: em bases maiores

Para variáveis discretas (Medicamentos, Doenças): Moda ou Valor clínico plausível (com justificativa)

ATIVIDADE

Atividade Prática – Análise de Dados para Geração de Insights e Propostas de Melhoria Organizacional

Objetivo

Esta atividade visa desenvolver a capacidade do aluno de utilizar dados como base para a tomada de decisão organizacional, promovendo a análise crítica de informações, a geração de percepções estratégicos e a proposição de melhorias fundamentadas tanto em evidências empíricas (dados) quanto em evidências científicas (literatura acadêmica).

Descrição

O aluno deverá selecionar uma base de dados relacionada ao seu contexto profissional ou área de atuação (ex.: vendas, marketing, logística, atendimento ao cliente, produção, tecnologia, saúde, educação, finanças, UX, entre outras) e realizar uma análise exploratória e interpretativa, para identificar problemas, oportunidades, padrões ou tendências que possam subsidiar propostas de melhoria para a empresa ou organização analisada.

ATIVIDADE

A base de dados poderá ser obtida a partir de fontes públicas ou privadas, conforme descrito a seguir.

Fonte dos Dados

Opção 1 – Base de Dados Pública

O aluno poderá utilizar bases de dados públicas provenientes de plataformas especializadas, tais como:

1. Kaggle
2. Portais de dados abertos governamentais
3. Repositórios acadêmicos ou institucionais

A base escolhida deve apresentar relação clara com o problema organizacional analisado.

ATIVIDADE

Opção 2 – Base de Dados Privada (Dados Reais da Empresa)

O uso de dados reais da empresa onde o aluno atua será permitido quando forem rigorosamente atendidas as seguintes condições:

- Anonimização completa dos dados, assegurando que:
  - Nenhuma informação sensível ou pessoal seja exposta (ex.: nomes, CPF, e-mails, telefones, endereços, dados financeiros pessoais, entre outros);
  - Os dados estejam conforme a Lei Geral de Proteção de Dados (LGPD).
- Autorização formal do gestor, supervisor ou responsável pela área, permitindo o uso acadêmico dos dados, mesmo que de forma anonimizada.
- A organização não deverá ser identificada nominalmente, sendo referenciada de forma genérica (ex.: Empresa A, Organização X, Empresa do setor Y).

ATIVIDADE

Etapas Obrigatórias do Trabalho

Antes do início das etapas abaixo todos os alunos deveram apresentar no fórum com título Fórum: Atividade 1 os seguintes elementos:

- Artigo encontrado para fundamentação da sua argumentação e sua referência bibliográfica conforme a NBR6023;
- Dataset apresentado com o link de acesso caso seja público, se for privado somente uma descrição do conteúdo;
- Justificativa da escolha do dataset, artigo e análise a ser feita.
- O preenchimento é obrigatório e valida a presença de alunos que faltaram ou ausentaram durante o período de aula.

ATIVIDADE

O FÓRUM FICARÁ DISPONÍVEL ATÉ O DIA 24/01, PREECHIMENTOS APÓS A DATA SERÃO APLICADO FALTA PARA OS ALUNOS.

Após o devido preenchimento, sua atividade deve seguir os pontos abaixo:

Contextualização do Problema

- Descrição da área ou setor analisado;
- Justificativa da escolha da base de dados;
- Identificação do problema, desafio ou oportunidade organizacional.

Descrição da Base de Dados

- Origem dos dados (pública ou privada);
- Tipo de dados (quantitativos, qualitativos, categóricos, temporais);
- Principais variáveis e volume aproximado da base.

ATIVIDADE

Análise Exploratória dos Dados

- Aplicação de estatísticas descritivas e/ou visualizações gráficas;
- Identificação de padrões, correlações, tendências ou anomalias;
- Interpretação dos resultados obtidos.

Geração de Insights

- Principais descobertas oriundas da análise dos dados;
- Impactos potenciais desses insights para a organização.

Propostas de Melhoria

- Sugestões de ações, mudanças de processo, estratégias ou decisões;
- Justificativa clara e objetiva, fundamentada nos dados analisados.

ATIVIDADE

Fundamentação Científica das Propostas

- O aluno deverá realizar pesquisa de artigos científicos por meio do Google Acadêmico;
- Os artigos devem estar diretamente relacionados ao problema analisado, às técnicas utilizadas ou ao setor de atuação;
- A literatura científica deverá ser utilizada para reforçar, justificar ou comparar as propostas de melhoria sugeridas;
- Utilizar no mínimo 3 (três) referências científicas, citadas ao longo do texto e listadas ao final do trabalho, conforme o padrão acadêmico adotado pela disciplina.

Considerações Éticas

- Descrição das medidas adotadas para proteção e anonimização dos dados (quando aplicável);
- Reflexão sobre o uso responsável e ético das informações.

ATIVIDADE

Entrega da Apresentação

- Apresentação em slides, no formato em PDF a ser enviado nesta atividade;
- Linguagem técnica, clara e objetiva;
- Inclusão de gráficos, tabelas e interpretações textuais;
- Referências bibliográficas organizadas conforme o padrão definido pelo professor.

Apresentação da Atividade

Além da apresentação criada, o aluno deverá realizar uma apresentação oral, conforme os critérios abaixo:

- Criação de slides de apresentação;
- Tempo máximo de 5 (cinco) minutos por aluno;

A apresentação deve contemplar, objetivamente:

- Contexto do problema;
- Base de dados utilizada;
- Principais insights obtidos;
- Propostas de melhoria fundamentadas nos dados e na literatura científica;
- Apresentação a ser realizada na próxima aula;.
- O controle do tempo faz parte da avaliação.

ATIVIDADE

Critérios de Avaliação

- Adequação da base de dados ao contexto profissional
- Qualidade da análise exploratória
- Clareza e relevância dos insights
- Coerência e viabilidade das propostas de melhoria
- Uso adequado da fundamentação científica
- Conformidade ética no uso dos dados
- Organização, escrita técnica e qualidade da apresentação oral

EXEMPLO

Contextualização do Problema

O setor de serviços digitais por assinatura enfrenta, atualmente, um dos principais desafios estratégicos relacionados à retenção de clientes, especialmente em mercados altamente competitivos. A perda recorrente de clientes (churn) impacta diretamente indicadores financeiros, custos de aquisição e sustentabilidade do negócio.

Diante desse cenário, este trabalho tem como objetivo analisar dados históricos de clientes de uma empresa de serviços digitais (denominada, para fins acadêmicos, Empresa X), buscando identificar padrões comportamentais associados ao cancelamento de serviços, bem como oportunidades de melhoria organizacional baseadas em evidências empíricas e científicas.

A escolha do problema justifica-se pela relevância prática do tema, amplamente discutido tanto no mercado quanto na literatura acadêmica sobre Customer Analytics e Data-Driven Decision Making.

EXEMPLO

2. Descrição da Base de Dados

2.1 Origem dos Dados

A base de dados utilizada é de fonte pública, obtida por meio da plataforma Kaggle, sendo amplamente empregada em estudos acadêmicos e aplicações educacionais voltadas à análise de churn em serviços digitais. Link de acesso: https://www.kaggle.com/datasets/blastchar/telco-customer-churn

2.2 Tipo e Estrutura dos Dados

- **Tipo de dados:** Quantitativos (ex.: tempo de contrato, valor mensal) e Qualitativos/categóricos (ex.: tipo de contrato, forma de pagamento)
- **Volume:** Aproximadamente 7.000 registros de clientes.
- **Variável alvo:** *Churn* (Sim / Não)

As variáveis representam informações contratuais, financeiras e de relacionamento com o cliente, permitindo uma análise abrangente do comportamento de cancelamento.

EXEMPLO

3. Análise Exploratória dos Dados

A análise exploratória foi conduzida por meio de **estatísticas descritivas** e **visualizações gráficas**, visando identificar padrões, tendências e possíveis relações entre as variáveis.

3.1 Principais Análises Realizadas

- Distribuição da variável *Churn*
- Comparação entre clientes que cancelaram e não cancelaram o serviço
- Análise da relação entre:
- Tempo de contrato (*tenure*)
- Tipo de contrato
- Valor mensal cobrado
- Forma de pagamento

WA_Fn-UseC_-Telco-Customer-Churn.csv

1 to 10 of 7043 entries Filter

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank transfer (automatic)	42.3	1840.75	No
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-month	Yes	Electronic check	70.7	151.65	Yes
9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to-month	Yes	Electronic check	99.65	820.5	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	No	Month-to-month	Yes	Credit card (automatic)	89.1	1949.4	No
6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes	No	No	No	No	No	Month-to-month	No	Mailed check	29.75	301.9	No
7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes	Month-to-month	Yes	Electronic check	104.8	3046.05	Yes
6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank transfer (automatic)	56.15	3487.95	No

Show 10 per page

1 2 10 100 600 700 705

EXEMPLO

3.2 Principais Padrões Identificados

- Clientes com **contratos mensais** apresentam maior taxa de churn em comparação com contratos anuais ou bienais.
- Clientes com **menor tempo de permanência** tendem a cancelar com maior frequência.
- Valores mensais mais elevados estão associados a maior probabilidade de churn, especialmente nos primeiros meses de contrato.
- Métodos de pagamento automáticos apresentam menor taxa de cancelamento.
- Esses padrões indicam que o churn está fortemente associado a **fatores contratuais e financeiros**, além da maturidade do relacionamento com o cliente.

WA_Fn-UseC_-Telco-Customer-Churn.csv

1 to 10 of 7043 entries Filter

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank transfer (automatic)	42.3	1840.75	No
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-month	Yes	Electronic check	70.7	151.65	Yes
9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to-month	Yes	Electronic check	99.65	820.5	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	No	Month-to-month	Yes	Credit card (automatic)	89.1	1949.4	No
6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes	No	No	No	No	No	Month-to-month	No	Mailed check	29.75	301.9	No
7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes	Month-to-month	Yes	Electronic check	104.8	3046.05	Yes
6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank transfer (automatic)	56.15	3487.95	No

Show 10 per page

1 2 10 100 600 700 705

EXEMPLO

4. Geração de Insights

A partir da análise dos dados, foram identificados os seguintes insights estratégicos:

1. Fase crítica do cliente: Os primeiros meses de relacionamento representam um período crítico para retenção.
2. Modelo contratual como fator decisivo: Contratos de curto prazo apresentam maior vulnerabilidade ao churn.
3. Sensibilidade ao preço: Clientes com maior custo mensal demonstram maior propensão ao cancelamento quando não percebem valor proporcional ao serviço.
4. Pagamentos automáticos como fator de retenção: A adoção de pagamentos recorrentes automatizados reduz a taxa de cancelamento.
5. Esses insights fornecem subsídios claros para ações organizacionais orientadas à retenção e à experiência do cliente.

WA_Fn-UseC_-Telco-Customer-Churn.csv

1 to 10 of 7043 entries Filter

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank transfer (automatic)	42.3	1840.75	No
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-month	Yes	Electronic check	70.7	151.65	Yes
9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to-month	Yes	Electronic check	99.65	820.5	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	No	Month-to-month	Yes	Credit card (automatic)	89.1	1949.4	No
6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes	No	No	No	No	No	Month-to-month	No	Mailed check	29.75	301.9	No
7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes	Month-to-month	Yes	Electronic check	104.8	3046.05	Yes
6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank transfer (automatic)	56.15	3487.95	No

Show 10 per page

1 2 10 100 600 700 705

EXEMPLO

5. Propostas de Melhoria Organizacional

Com base nos dados analisados, propõem-se as seguintes ações:

5.1 Ações de Curto Prazo

- Implementação de programas de acompanhamento intensivo nos primeiros 90 dias de contrato.
- Oferta de incentivos para migração de contratos mensais para planos de longo prazo.
- Revisão da comunicação de valor para clientes com mensalidades mais elevadas.

5.2 Ações de Médio Prazo

- Desenvolvimento de modelos preditivos de churn, permitindo intervenções proativas.
- Segmentação de clientes com base em risco de cancelamento.
- Campanhas personalizadas baseadas no perfil comportamental identificado.

Essas propostas visam reduzir custos de aquisição, aumentar a retenção e melhorar a sustentabilidade do negócio.

EXEMPLO

6. Fundamentação Científica das Propostas

As propostas apresentadas são sustentadas por evidências da literatura científica, conforme destacado a seguir:

Verbeke et al. (2012) demonstram que modelos analíticos baseados em dados históricos são eficazes na identificação de churn e suporte à tomada de decisão estratégica. Idris, Khan e Lee (2012) destacam a importância da segmentação de clientes e da análise comportamental para ações de retenção.

Huang et al. (2015) evidenciam que fatores contratuais e financeiros exercem influência significativa sobre o cancelamento de serviços.

Esses estudos corroboram os achados empíricos observados na análise exploratória realizada.

7. Considerações Éticas

Embora a base utilizada seja pública, foram respeitados princípios éticos fundamentais:

- Nenhuma informação pessoal sensível foi utilizada ou exposta.
- A análise respeitou os princípios da **Lei Geral de Proteção de Dados (LGPD)**, utilizando dados anonimizados.
- Os resultados foram interpretados de forma responsável, evitando generalizações indevidas ou uso discriminatório das informações.
- A reflexão ética reforça a importância do uso consciente de dados no contexto organizacional e acadêmico.

EXEMPLO

8. Considerações Finais

Este trabalho demonstrou como a análise exploratória de dados, aliada à fundamentação científica, pode gerar insights relevantes e propostas de melhoria organizacional. Os resultados evidenciam o papel estratégico da análise de dados como base para decisões informadas e sustentáveis.

Além disso, a atividade reforça a importância de uma abordagem ética, crítica e fundamentada no uso de dados, preparando o profissional para atuar de forma responsável em projetos de Inteligência Artificial e Machine Learning.

9. Referências Bibliográficas (NBR 6023)

- VERBEKE, W.; DEJAEGER, K.; MARTENS, D.; HUR, J.; BAESE, B. New insights into churn prediction in the telecommunication sector. *Expert Systems with Applications*, v. 39, n. 1, p. 1–11, 2012.
- IDRIS, A.; KHAN, A.; LEE, Y. Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost based ensemble classification. *Applied Soft Computing*, v. 12, n. 11, p. 3696–3706, 2012.
- HUANG, B.; KECHADI, M.; BUCKLEY, B. Customer churn prediction in telecommunications. *Expert Systems with Applications*, v. 39, n. 1, p. 1414–1425, 2015.

DADOS SUPERVISIONADO E NÃO SUPERVISIONADOS

Os conceitos de **dados supervisionados** e **não supervisionados** estão no cerne do Aprendizado de Máquina (Machine Learning), definindo como os algoritmos aprendem a partir das informações fornecidas. A principal diferença reside na presença de **rótulos** (labels) ou respostas corretas nos dados de treinamento

DADOS SUPERVISIONADO E NÃO SUPERVISIONADOS

Dados Supervisionados (Supervised Learning)

No aprendizado supervisionado, o modelo é treinado usando um conjunto de dados que contém tanto as entradas (features) quanto as saídas corretas desejadas (labels/rótulos). Pense nisso como aprender com um professor: o computador recebe exemplos rotulados, prevê a saída e ajusta-se com base no erro.

- Característica: Dados rotulados (Input + Output).
- Objetivo: Mapear uma entrada para uma saída (previsão).

Tipos de Tarefa:

- Classificação: Prever uma categoria ou classe (ex: Spam/Não Spam, Gato/Cachorro).
- Regressão: Prever um valor numérico contínuo (ex: preço de uma casa, previsão de temperatura).

Exemplos de Uso: Reconhecimento de imagem, detecção de fraude bancária, diagnóstico médico.

DADOS SUPERVISIONADO E NÃO SUPERVISIONADOS

Dados Não Supervisionados (Unsupervised Learning)

No aprendizado não supervisionado, o modelo trabalha com dados que não possuem rótulos ou respostas corretas pré-definidas. O algoritmo deve explorar os dados por conta própria para identificar estruturas ocultas, padrões ou agrupamentos.

- Característica: Dados não rotulados (Apenas Input).
- Objetivo: Encontrar estruturas, padrões ocultos ou agrupamentos.

Tipos de Tarefa:

- Agrupamento (Clustering): Agrupar dados similares (ex: K-means).
- Associação: Encontrar regras que descrevem grandes porções dos dados.
- Redução de Dimensionalidade: Simplificar dados complexos sem perder informações essenciais.

Exemplos de Uso: Segmentação de clientes para marketing, sistemas de recomendação, detecção de anomalias.

MODELAGEM PREDITIVA E MACHINE LEARNING

1. Formulação do Problema (Conceitual, mas Aplicada)

1.1. O que é a formulação de um problema?

Antes de implementar qualquer modelo de Machine Learning, é essencial entender o problema que queremos resolver. Isso envolve:

- Identificar o objetivo.
- Compreender os dados disponíveis.
- Levantar as hipóteses importantes.
- Garantir que o problema seja bem estruturado para aprendizado.

MODELAGEM PREDITIVA E MACHINE LEARNING

1.2. Exemplo prático da formulação de problema:

Caso: Determinar o risco de inadimplência para clientes de um banco.

1. **Objetivo:** Predizer se um cliente será inadimplente nos próximos 6 meses (classificação binária).
2. **Contexto:** Dados históricos dos clientes, incluindo renda, histórico de pagamentos e limitações de crédito.
3. **Impacto:** Reduzir perdas financeiras e otimizar os recursos de cobrança.

MODELAGEM PREDITIVA E MACHINE LEARNING

1.2. Exemplo prático da formulação de problema:

Caso: Determinar o risco de inadimplência para clientes de um banco.

1. **Objetivo:** Predizer se um cliente será inadimplente nos próximos 6 meses (classificação binária).
2. **Contexto:** Dados históricos dos clientes, incluindo renda, histórico de pagamentos e limitações de crédito.
3. **Impacto:** Reduzir perdas financeiras e otimizar os recursos de cobrança.

1.3. Perguntas críticas para formular o problema:

- Qual é o objetivo principal do modelo?
- A variável-alvo está bem definida?
- As features disponíveis são relevantes para o problema?
- Existem vieses potenciais nos dados que precisam ser mitigados?

MODELAGEM PREDITIVA E MACHINE LEARNING

2. Variável Alvo (Target)

2.1. Definição e Importância

A variável-alvo é o "resultado" que queremos prever.

Exemplos de variáveis-alvo para problemas específicos:

- Saúde: Probabilidade de ser internado (0 para não, 1 para sim).
- Educação: Risco de evasão escolar (baixa, média, alta).
- Negócios: Chance de churn (cancelar assinatura).

2.2. Características de uma boa variável-alvo

- **Clareza:** Deve ser bem definida e mensurável.
- **Relevância:** Deve ser representativa do problema.
- **Estimabilidade:** Com base nos dados históricos, deve ser possível inferi-la.

MODELAGEM PREDITIVA E MACHINE LEARNING

2.3. Exemplos práticos de definição de target:

Saúde: Variável-alvo binária: "Paciente será internado nos próximos 30 dias" (0 ou 1).

Educação: Variável-alvo multiclasse: "Probabilidade de um estudante evadir" (0 para baixa, 1 para média, 2 para alta).

Negócios: Regressão: "Valor de débito previsto para um cliente no próximo trimestre".

MODELAGEM PREDITIVA E MACHINE LEARNING

3. Features vs Leakage

3.1. O que são features (variáveis explicativas)?

As features são as variáveis de entrada que o modelo usa para prever a variável-alvo. Ex.: Idade, renda mensal, histórico de crédito.

3.2. O problema de leakage (vazamento de informações)

Leakage ocorre quando uma ou mais features contêm informações que só estariam disponíveis **após o evento a ser previsto**. Isso resulta em um modelo que tem desempenho irreal durante o treino e acaba inútil no mundo real.

MODELAGEM PREDITIVA E MACHINE LEARNING

3.3. Exemplos concretos de leakage:

Usar resultados de exames realizados após o processo de internação para prever se o paciente será internado. Também podemos utilizar na educação onde as notas finais dos alunos para prever se eles evadirão antes da conclusão do curso. Ou como vimos no exemplo desta aula os dados foram usados para o cancelamento de assinatura como feature para prever o churn.

3.4. Boas práticas para evitar leakage:

Excluir variáveis que contenham informações posteriores ao evento-alvo. Dividir corretamente os dados para treino, validação e teste (ex. estratificação temporal).

MODELAGEM PREDITIVA E MACHINE LEARNING

4. Tipos de Problemas de Machine Learning

4.1. Classificação Binária

Definição: Problema em que a variável-alvo tem dois possíveis valores.

Exemplos Reais:

Saúde: Prever se um paciente será hospitalizado (Sim/Não).

- **Problema:** Prever o risco de internação.
- **Tipo:** Classificação Binária.
- **Target:** Internação (Sim/Não).
- **Features:** Idade, histórico de doenças, resultados de exames laboratoriais anteriores.
- **Benefício:** Antecipar internações para otimizar recursos hospitalares.

MODELAGEM PREDITIVA E MACHINE LEARNING

4.2. Classificação Multiclasse

Definição: Problema em que a variável-alvo pode assumir mais de dois valores discretos.

Exemplos Reais:

Educação: Classificar alunos em "baixo", "médio" ou "alto" risco de evasão.

- **Problema:** Prever evasão de alunos.
- **Tipo:** Classificação Multiclasse.
- **Target:** Risco de evasão (Baixo, Médio, Alto).
- **Features:** Frequência, notas, envolvimento em atividades acadêmicas.
- **Benefício:** Estruturar estratégias de apoio e retenção de estudantes.

MODELAGEM PREDITIVA E MACHINE LEARNING

4.3. Regressão

Definição: Problema em que a variável-alvo assume valores contínuos.

Exemplos Reais:

- **Saúde (Predizer custos hospitalares):** Algoritmos de regressão (como Regressão Linear Múltipla) analisam dados demográficos, histórico de doenças e tipos de procedimentos para estimar os custos hospitalares de um paciente. Isso ajuda hospitais a planejar recursos e a definir orçamentos.
- **Negócios (Estimar faturamento):** Regressão é usada para modelar o impacto de investimentos em marketing, sazonalidade e preços sobre o volume de vendas. Isso permite prever o faturamento de uma empresa no próximo trimestre ou ano.
- **Educação (Prever desempenho):** Modelos de regressão analisam variáveis como horas de estudo, frequência às aulas e notas prévias para prever a média final de um aluno ou a probabilidade de aprovação.

MODELAGEM PREDITIVA E MACHINE LEARNING

Regressão Linear

A regressão linear é uma técnica estatística fundamental para modelagem e análise de dados que busca encontrar a relação entre uma variável dependente (ou resposta) e uma ou mais variáveis independentes (ou preditoras), assumindo uma relação linear entre elas. O objetivo da regressão linear é encontrar a "melhor" linha reta que descreve a relação entre as variáveis independentes e a variável dependente. Essa linha é chamada de "linha de regressão" e é representada pela equação:

- $Y = ß_0 + ß_1X_1 + ß_2X_2 + ... ß_nX_n + \mathcal{E}$
- Y é a variável dependente (ou resposta);
- $X_1, X_2, ... X_n$ são os coeficientes que representam a interceptação da linha de regressão com o eixo Y e as inclinações da linha em relação a cada uma das variáveis independes;
- $\mathcal{E}$ é o termo de erro, representando a diferença entre os valores observados da variável dependente e os valores previstos pela equação da regressão.

MODELAGEM PREDITIVA E MACHINE LEARNING

Exemplo de Regressão Linear

Previsão de novos valores: Depois de obter a equação da regressão, você pode usar essa equação para prever novos valores de y para um dado x. Basta substituir os valores conhecidos de x na equação para calcular os valores preditos de y.

Por exemplo, se você tem uma equação de regressão linear simples $y=2x+3$ e deseja prever o valor de y quando $x=5$, você substitui $x=5$ na equação:

$y=2\times5+3=13$

O modelo prevê que y será igual a 13 quando x é igual a 5. Este é um exemplo simples; em situações reais, o processo pode ser mais complexo, especialmente em regressões múltiplas com várias variáveis independentes.

MODELAGEM PREDITIVA E MACHINE LEARNING

Previsão de Vendas: Uma empresa pode usar a regressão linear para prever as vendas futuras com base em variáveis como publicidade, preço do produto, época do ano, entre outras.

Análise Financeira: Os analistas financeiros podem usar a regressão linear para prever o preço de uma ação com base em variáveis como lucros da empresa, indicadores econômicos, volumes de negociação, entre outros.

Previsão de Produção Agrícola: Os agricultores podem usar a regressão linear para prever a produção agrícola com base em variáveis como temperatura, umidade, tipo de solo, entre outros.

Análise de Marketing: As empresas podem usar a regressão linear para entender como diferentes variáveis de marketing, como gastos com publicidade, influenciam as métricas de desempenho, como o número de clientes adquiridos.

Previsão de Consumo de Energia: As empresas de serviços públicos podem usar a regressão linear para prever o consumo de energia com base em variáveis como temperatura, temporada, demografia, entre outros.

MODELAGEM PREDITIVA E MACHINE LEARNING

Análise de Marketing

Uma empresa quer prever suas vendas mensais com base nos gastos com publicidade em diferentes canais, como televisão, rádio e jornal. Eles coletaram dados de vendas mensais (variável dependente) e gastos em publicidade em cada canal (variáveis independentes) nos últimos meses. Agora, eles desejam usar esses dados para prever as vendas futuras com base nos gastos com publicidade.

Aqui está um conjunto hipotético de dados:

MÊS	GASTOS TV (MILHARES DE DÓLARES)	GASTOS RÁDIO (MILHARES DE DÓLARES)	GASTOS JORNAL (MILHARES DE DÓLARES)	VENDAS (MILHARES DE DÓLARES)
1	230	37	69	480
2	44	39	45	200
3	17	45	69	150
4	200	45	69	700
5	60	48	69	400

MODELAGEM PREDITIVA E MACHINE LEARNING

$Y = ß_0 + ß_1X_1 + ß_2X_2 + \ldots ß_nX_n + \mathcal{E}$

$Vendas = ß_0 + ß_1 * Gastos_{TV} + ß_2 * Gastos_{Rádio} + ß_3 * Gastos_{Jornal} + \mathcal{E}$

Como precisamos calcular os coeficientes $ß_0$, $ß_1$, $ß_2$, $ß_3$ que são minimizam a soma dos quadrados dos resíduos. Podemos considerar que os valores de cada um dos coeficientes são:

$ß_0 = 50$; $ß_1 = 2$; $ß_2 = 3$; $ß_3 = 1$

Método dos Mínimos Quadrados Ordinários (OLS):

O método dos mínimos quadrados ordinários (OLS) é uma técnica comum para ajustar os coeficientes da regressão. OLS encontra os coeficientes que minimizam a soma dos quadrados dos resíduos (diferenças entre os valores observados e os valores previstos). Esses coeficientes podem ser calculados usando técnicas matemáticas, como álgebra linear.

MODELAGEM PREDITIVA E MACHINE LEARNING

MÊS	GASTOS TV (MILHARES DE DÓLARES)	GASTOS RÁDIO (MILHARES DE DÓLARES)	GASTOS JORNAL (MILHARES DE DÓLARES)	VENDAS (MILHARES DE DÓLARES)
1	230	37	69	480
2	44	39	45	200
3	17	45	69	150
4	200	45	69	700
5	60	48	69	400

$$Vendas = ß_0 + ß_1 * Gastos_{TV} + ß_2 * Gastos_{Rádio} + ß_3 * Gastos_{Jornal} + \mathcal{E}$$

$ß_0 = 50; ß_1 = 2; ß_2 = 3; ß_3 = 1$

$$Vendas_{previstas} = 50 + 2 * 230 + 3 * 37 + 1 * 69$$

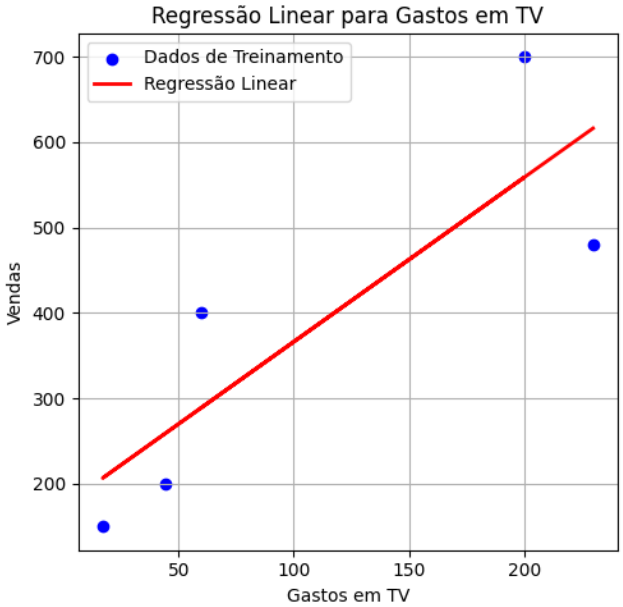
$$Vendas_{previstas} = 50 + 460 + 111 + 69$$

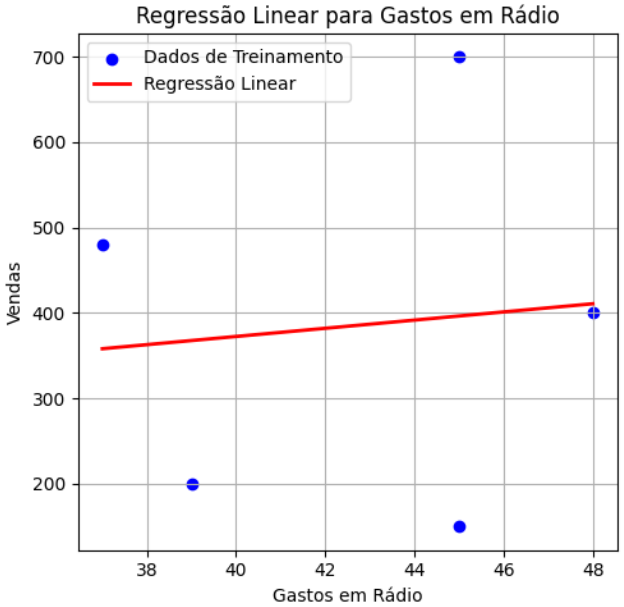
$$Vendas_{previstas} = 690$$

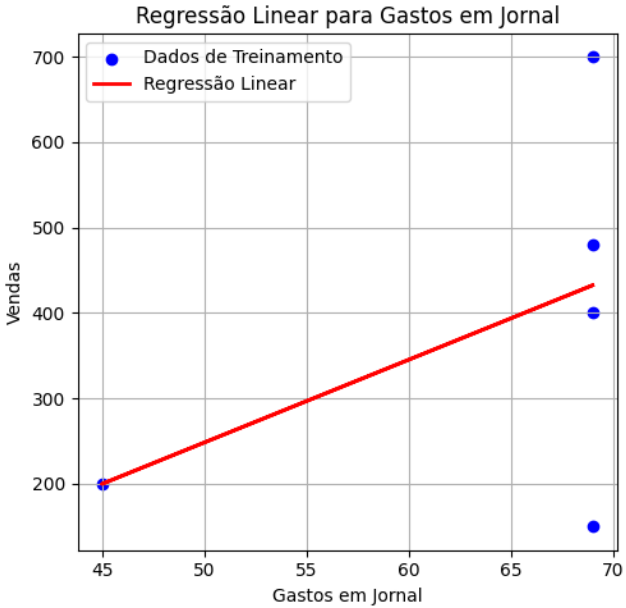
As previsões de venda no primeiro mês será de R$690,00.

```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression

# Dados de exemplo
gastos_tv = np.array([230, 44, 17, 200, 60]).reshape(-1, 1)  # Reshape para uma matriz 2D
gastos_radio = np.array([37, 39, 45, 45, 48]).reshape(-1, 1)
gastos_jornal = np.array([69, 45, 69, 69, 69]).reshape(-1, 1)
vendas = np.array([480, 200, 150, 700, 400])

# Lista de tuplas para iterar sobre os dados de gastos
dados_gastos = [("TV", gastos_tv), ("Rádio", gastos_radio), ("Jornal", gastos_jornal)]

# Criar o modelo de regressão linear
modelo_regressao = LinearRegression()

# Criar subplots
fig, axs = plt.subplots(1, 3, figsize=(15, 5))

# Iterar sobre os dados de gastos
for i, (tipo_gasto, gastos) in enumerate(dados_gastos):
    # Ajustar o modelo aos dados
    modelo_regressao.fit(gastos, vendas)

    # Coeficientes do modelo
    coeficiente = modelo_regressao.coef_[0]
    intercepto = modelo_regressao.intercept_

    # Plotar o gráfico de dispersão dos dados e a linha de regressão
    axs[i].scatter(gastos, vendas, color='blue', label='Dados de Treinamento')
    axs[i].plot(gastos, modelo_regressao.predict(gastos), color='red', linewidth=2, label='Regressão Linear')
    axs[i].set_title(f'Regressão Linear para Gastos em {tipo_gasto}')
    axs[i].set_xlabel(f'Gastos em {tipo_gasto}')
    axs[i].set_ylabel('Vendas')
    axs[i].legend()
    axs[i].grid(True)

plt.tight_layout()
plt.show()
```

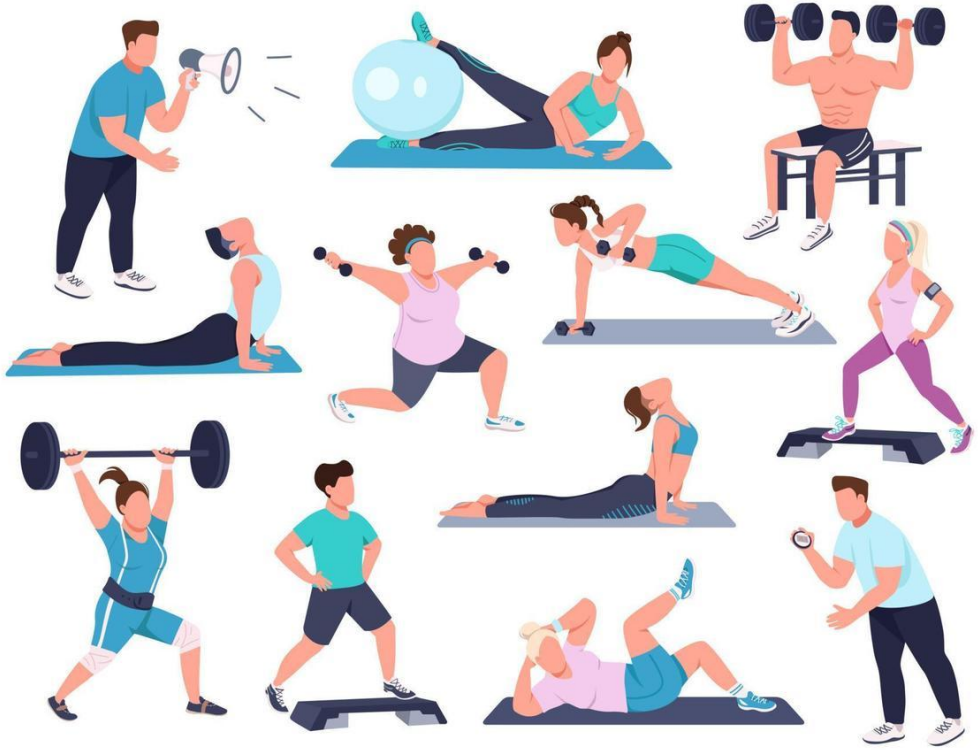
MODELAGEM PREDITIVA E MACHINE LEARNING







EXERCÍCIO



MODELAGEM PREDITIVA E MACHINE LEARNING

Uma árvore de decisão é um modelo de aprendizado de máquina supervisionado utilizado para classificação e regressão. Ele funciona como uma estrutura em forma de árvore, na qual cada nó representa uma característica (ou atributo), cada ramo representa uma decisão baseada nessa característica, e cada folha representa o resultado da decisão (uma classe ou um valor numérico).

O modelo de árvores de decisão tendem a ser propensas a overfitting (sobreajuste) em conjuntos de dados complexos. Para mitigar esse problema, técnicas como poda da árvore, uso de árvores ensemble (como Random Forests) e ajuste de hiperparâmetros são comumente empregadas.

MODELAGEM PREDITIVA E MACHINE LEARNING

Um exemplo simples de árvore de decisão para classificar se uma fruta é uma laranja ou uma maçã, com base em duas características: cor e textura da casca. Suponha que tenhamos o seguinte conjunto de dados de treinamento:

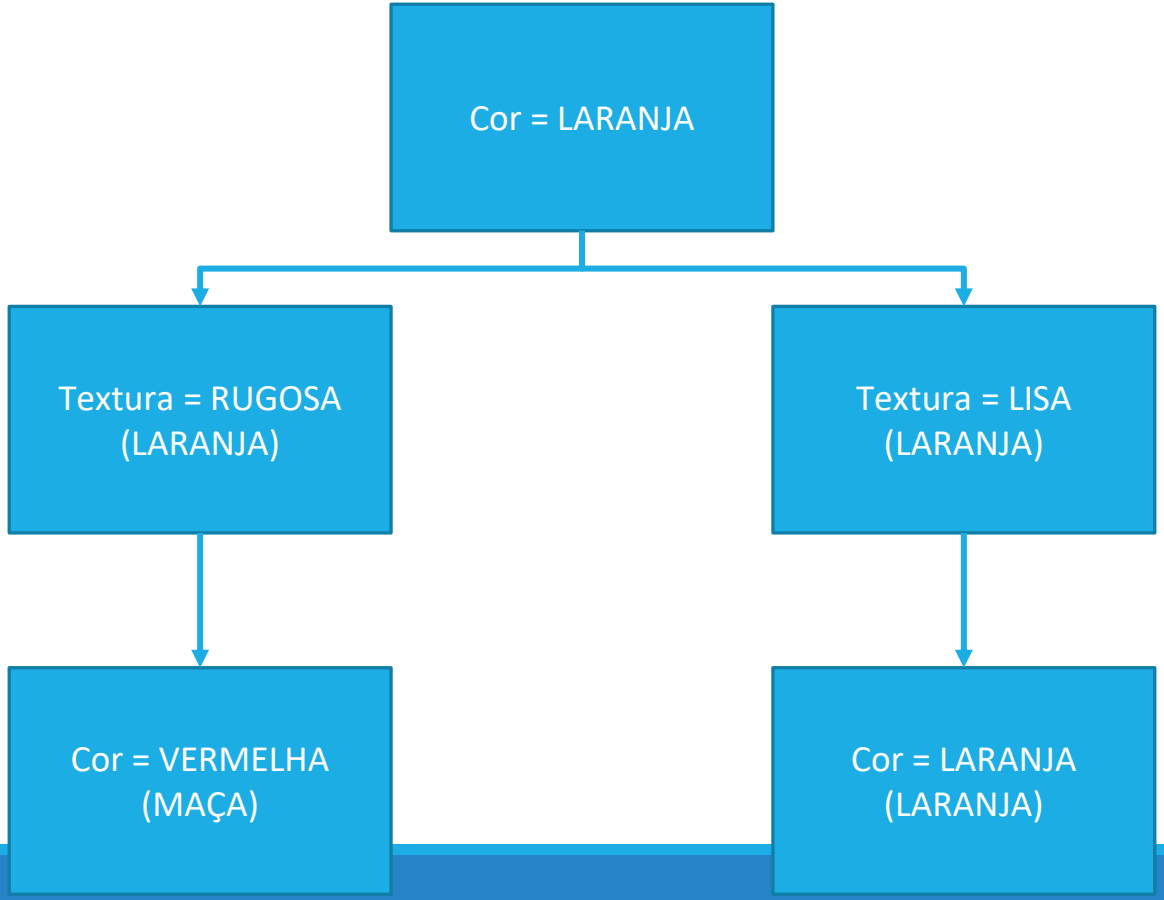
AMOSTRA	COR	TEXTURA	FRUTAS
1	LARANJA	RUGOSA	LARANJA
2	LARANJA	LISA	LARANJA
3	VERMELHO	RUGOSA	MAÇÃ
4	VERMELHO	LISA	MAÇÃ
5	LARANJA	RUGOSA	LARANJA

MODELAGEM PREDITIVA E MACHINE LEARNING

1. No nó raiz, escolhemos a característica mais discriminativa. Neste caso, vamos escolher a cor. Dividimos o conjunto de dados com base na cor.
2. No primeiro nível, temos dois nós filhos: um para "Cor = Laranja" e outro para "Cor = Vermelha".
3. Para o nó "Cor = Laranja", verificamos a textura. Se a textura for "Rugosa", classificamos como "Laranja". Se for "Lisa", classificamos como "Laranja".
4. Para o nó "Cor = Vermelha", seguimos o mesmo procedimento: se a textura for "Rugosa", classificamos como "Maçã". Se for "Lisa", classificamos como "Maçã".

Essa árvore de decisão nos permite classificar novas amostras de frutas com base em sua cor e textura da casca. Por exemplo, uma fruta com cor laranja e textura lisa seria classificada como laranja, enquanto uma fruta com cor vermelha e textura rugosa seria classificada como maçã.

MODELAGEM PREDITIVA E MACHINE LEARNING



MODELAGEM PREDITIVA E MACHINE LEARNING

```python
import pandas as pd
import numpy as np
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt

# Gerar dados de exemplo
np.random.seed(42)  # Para reproduzibilidade
cores = np.random.choice(['Laranja', 'Vermelha'], size=90000)
texturas = np.random.choice(['Rugosa', 'Lisa'], size=90000)
frutas = np.random.choice(['Laranja', 'Maçã'], size=90000)

# Criar DataFrame
df = pd.DataFrame({'Cor': cores, 'Textura': texturas, 'Fruta': frutas})

# Mapear dados categóricos para numéricos
df['Cor'] = df['Cor'].map({'Laranja': 0, 'Vermelha': 1})
df['Textura'] = df['Textura'].map({'Rugosa': 0, 'Lisa': 1})
df['Fruta'] = df['Fruta'].map({'Laranja': 0, 'Maçã': 1})

# Separar features e target
X = df[['Cor', 'Textura']]
y = df['Fruta']

# Dividir conjunto de dados em treino e teste
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Criar e treinar o modelo de árvore de decisão
modelo = DecisionTreeClassifier()
modelo.fit(X_train, y_train)
```

MODELAGEM PREDITIVA E MACHINE LEARNING

```
# Fazer previsões no conjunto de teste
previsoes = modelo.predict(X_test)

# Calcular a acurácia do modelo
acuracia = accuracy_score(y_test, previsoes)
print("Acurácia do modelo: {:.2f}".format(acuracia))

# Contagem de amostras de cada classe
contagem_classes = df['Fruta'].value_counts()

# Gráfico de barras
plt.figure(figsize=(8, 6))
contagem_classes.plot(kind='bar', color=['orange', 'red'])
plt.title('Distribuição das classes')
plt.xlabel('Classe')
plt.ylabel('Quantidade')
plt.xticks([0, 1], ['Laranja', 'Maçã'], rotation=0)
plt.show()
```

MODELAGEM PREDITIVA E MACHINE LEARNING

O Random Forest é um algoritmo de aprendizado de máquina que pertence à categoria de métodos de ensemble, ou seja, ele combina múltiplos modelos de aprendizado de máquina para realizar uma tarefa específica, como classificação ou regressão. A ideia básica por trás do Random Forest é construir uma "floresta" de árvores de decisão durante o treinamento.

Cada árvore de decisão é construída de forma independente, utilizando uma amostra aleatória dos dados de treinamento e um subconjunto aleatório das características (variáveis) disponíveis. Durante a previsão, as predições de cada árvore são combinadas (geralmente por média ou votação) para produzir uma predição final.

MODELAGEM PREDITIVA E MACHINE LEARNING

Para o exemplo vamos usar o conjunto de dados Iris, que é um conjunto de dados clássico em aprendizado de máquina, e vamos construir um modelo Random Forest para classificar as diferentes espécies de flores com base em suas características.

A base de dados Iris é um conjunto de dados clássico frequentemente usado em aprendizado de máquina e estatísticas para fins de demonstração e teste de algoritmos. Essa base de dados foi introduzida em 1936 pelo estatístico britânico Ronald Fisher em seu artigo "The Use of Multiple Measurements in Taxonomic Problems" e é amplamente utilizada como exemplo em diversas áreas da ciência de dados.

O conjunto de dados Iris é composto por 150 amostras, com 50 amostras de cada espécie. Cada amostra é representada por um vetor de características com as medidas mencionadas acima, juntamente com um rótulo indicando a espécie da flor correspondente.

MODELAGEM PREDITIVA E MACHINE LEARNING

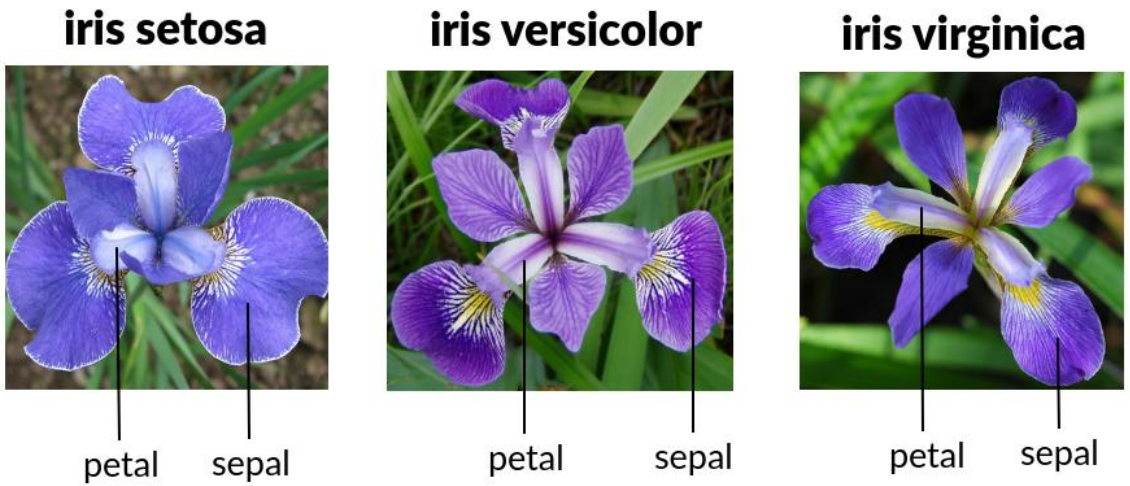
A base de dados Iris consiste em amostras de três espécies de flores Iris: Iris setosa, Iris versicolor e Iris virginica. Para cada espécie, foram medidas quatro características morfológicas das flores:

Comprimento da sépala (em centímetros)

Largura da sépala (em centímetros)

Comprimento da pétala (em centímetros)

Largura da pétala (em centímetros)



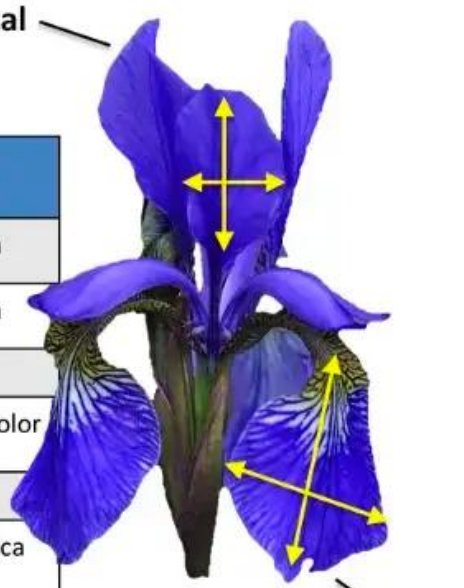
MODELAGEM PREDITIVA E MACHINE LEARNING

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

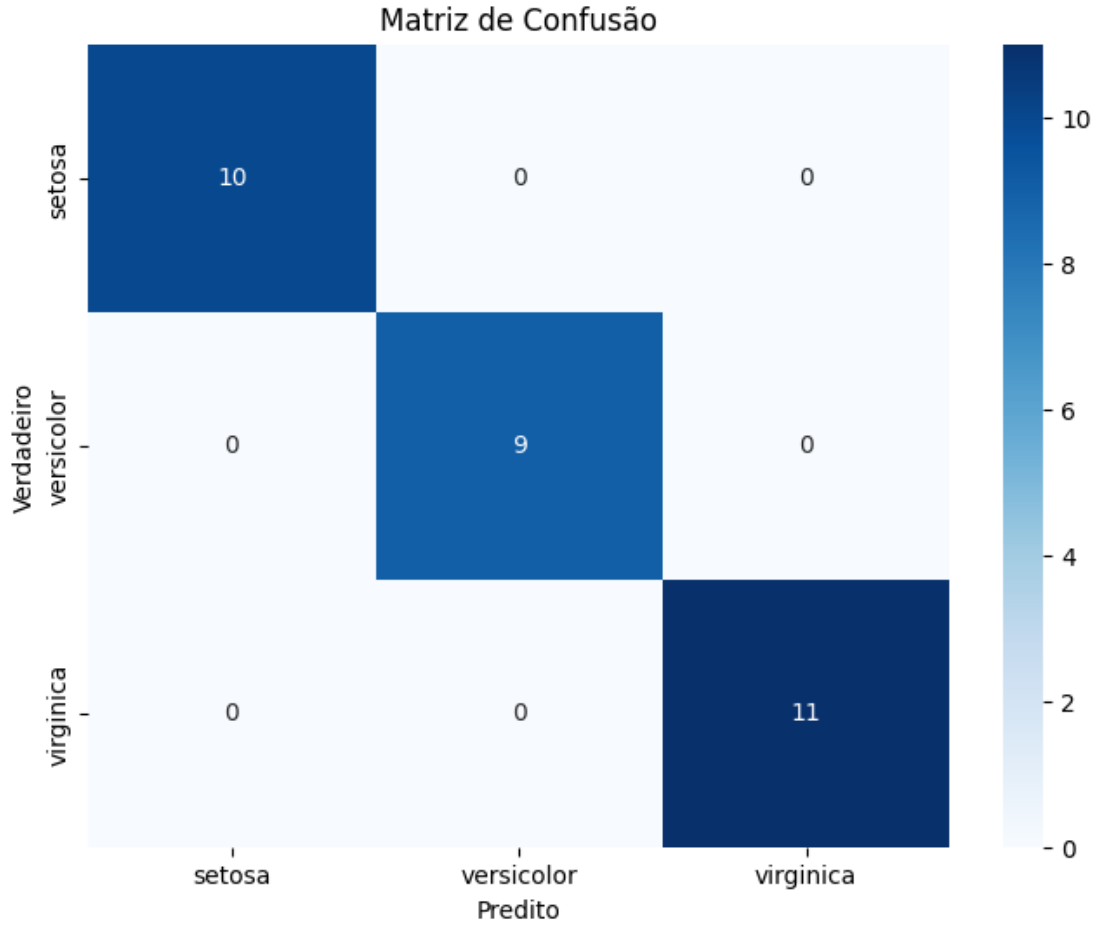
Features
(attributes, measurements, dimensions)

Class labels
(targets)

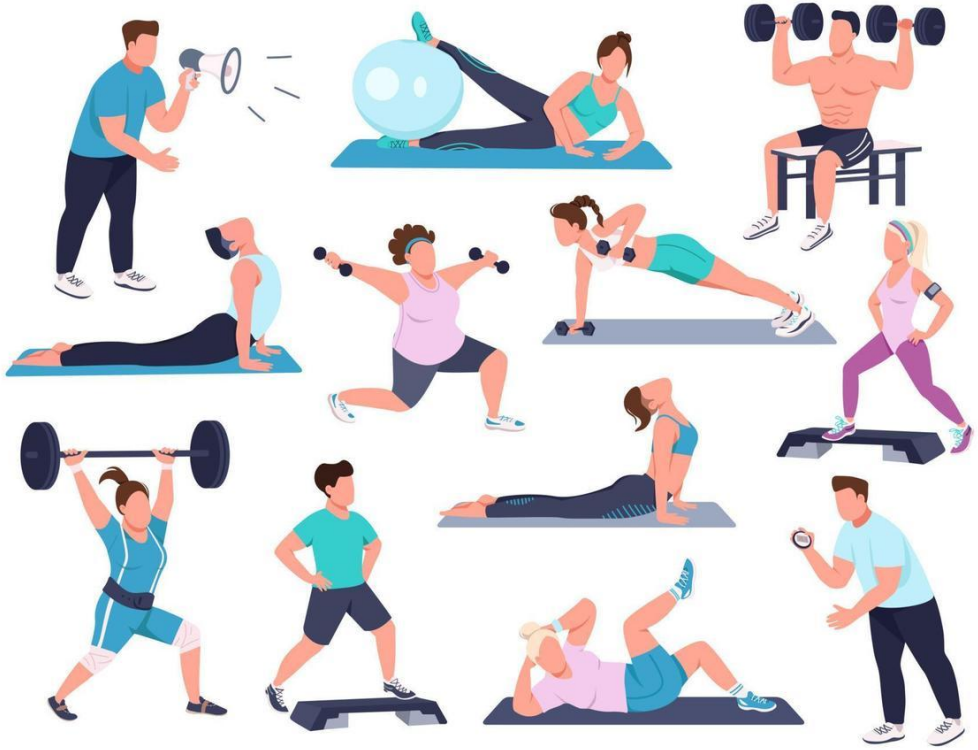


```python
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix

# Carregando o conjunto de dados Iris
iris = load_iris()
X = iris.data  # características
y = iris.target  # rótulos

# Dividindo os dados em conjunto de treinamento e teste
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Criando e treinando o modelo Random Forest
rf_classifier = RandomForestClassifier(n_estimators=100, random_state=42)
rf_classifier.fit(X_train, y_train)

# Fazendo previsões no conjunto de teste
y_pred = rf_classifier.predict(X_test)

# Avaliando a precisão do modelo
accuracy = accuracy_score(y_test, y_pred)
print("Precisão do modelo Random Forest:", accuracy)

# Calculando a matriz de confusão
conf_matrix = confusion_matrix(y_test, y_pred)

# Plotando a matriz de confusão
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix, annot=True, cmap='Blues', fmt='g', xticklabels=iris.target_names, yticklabels=iris.target_names)
plt.xlabel('Predito')
plt.ylabel('Verdadeiro')
plt.title('Matriz de Confusão')
plt.show()
```

MACHINE LEARNING



EXERCÍCIO



MODELAGEM PREDITIVA E MACHINE LEARNING

A matriz de confusão é uma tabela que é usada em problemas de classificação para descrever o desempenho de um modelo de aprendizado de máquina. Ela mostra o número de observações em cada classe que foram corretamente ou incorretamente classificadas pelo modelo.

A matriz de confusão é uma ferramenta útil para avaliar o desempenho de um modelo de classificação, pois fornece uma visão detalhada dos tipos de erros que o modelo está cometendo. Com base na matriz de confusão, várias métricas de avaliação, como precisão, recall, F1-score, entre outras, podem ser calculadas para entender melhor o desempenho do modelo em diferentes aspectos.

MODELAGEM PREDITIVA E MACHINE LEARNING

Para um problema de classificação com duas classes (positivo e negativo), uma matriz de confusão teria a seguinte estrutura:

	Positivo (Previsto)	Negativo (Previsto)
Positivo (Real)	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Negativo (Real)	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Verdadeiro Positivo (VP): Número de observações corretamente classificadas como positivas.

Falso Negativo (FN): Número de observações que foram erroneamente classificadas como negativas (quando na verdade eram positivas).

Falso Positivo (FP): Número de observações que foram erroneamente classificadas como positivas (quando na verdade eram negativas).

Verdadeiro Negativo (VN): Número de observações corretamente classificadas como negativas.

MODELAGEM PREDITIVA E MACHINE LEARNING

Para um problema de classificação com duas classes (positivo e negativo), uma matriz de confusão teria a seguinte estrutura:

	Positivo (Previsto)	Negativo (Previsto)
Positivo (Real)	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Negativo (Real)	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Verdadeiro Positivo (VP): Número de observações corretamente classificadas como positivas.

Falso Negativo (FN): Número de observações que foram erroneamente classificadas como negativas (quando na verdade eram positivas).

Falso Positivo (FP): Número de observações que foram erroneamente classificadas como positivas (quando na verdade eram negativas).

Verdadeiro Negativo (VN): Número de observações corretamente classificadas como negativas.

MODELAGEM PREDITIVA E MACHINE LEARNING

Uma Máquina de Vetores de Suporte (SVM - Support Vector Machine) é um algoritmo de aprendizado supervisionado usado principalmente para classificação e regressão. O objetivo principal do SVM é encontrar um hiperplano no espaço de alta dimensão que possa separar os exemplos de diferentes classes com a maior margem possível.

Os SVMs são amplamente utilizados em uma variedade de campos, incluindo reconhecimento de padrões, bioinformática, reconhecimento de escrita à mão, entre outros, devido à sua capacidade de lidar com problemas complexos de classificação e regressão.

```python
# Importando as bibliotecas necessárias
import numpy as np
import matplotlib.pyplot as plt
from sklearn import svm
from sklearn.datasets import make_blobs

# Gerando um conjunto de dados sintético para classificação binária
X, y = make_blobs(n_samples=50, centers=2, random_state=6)

# Criando o classificador SVM
clf = svm.SVC(kernel='linear')

# Treinando o modelo SVM
clf.fit(X, y)

# Plotando os pontos de dados
plt.scatter(X[:, 0], X[:, 1], c=y, s=30, cmap=plt.cm.Paired)

# Plotando o hiperplano de separação
# Obtendo os coeficientes do hiperplano
w = clf.coef_[0]
b = clf.intercept_[0]

# Gerando os valores para a linha de separação
x_plot = np.linspace(np.min(X[:, 0]), np.max(X[:, 0]), 100)
y_plot = - (w[0] * x_plot + b) / w[1]

# Plotando a linha de separação
plt.plot(x_plot, y_plot, 'k-')

# Plotando os vetores de suporte
plt.scatter(clf.support_vectors_[:, 0], clf.support_vectors_[:, 1], s=100,
            linewidth=1, facecolors='none', edgecolors='k')

plt.title('SVM Result')
plt.xlabel('Feature 1')
plt.ylabel('Feature 2')
plt.show()
```

MODELAGEM PREDITIVA E MACHINE LEARNING

Este modelo SVM é eficaz para classificar novos pontos de dados não vistos em duas classes com base em suas características. Ele pode ser usado para resolver uma variedade de problemas de classificação binária, desde que os dados possam ser linearmente separáveis ou aproximadamente separáveis por um hiperplano. A definição formal de um hiperplano em um espaço *n*-dimensional é uma superfície que satisfaz a equação:

$a_1x_1+a_2x_2+\ldots+a_nx_n=b$

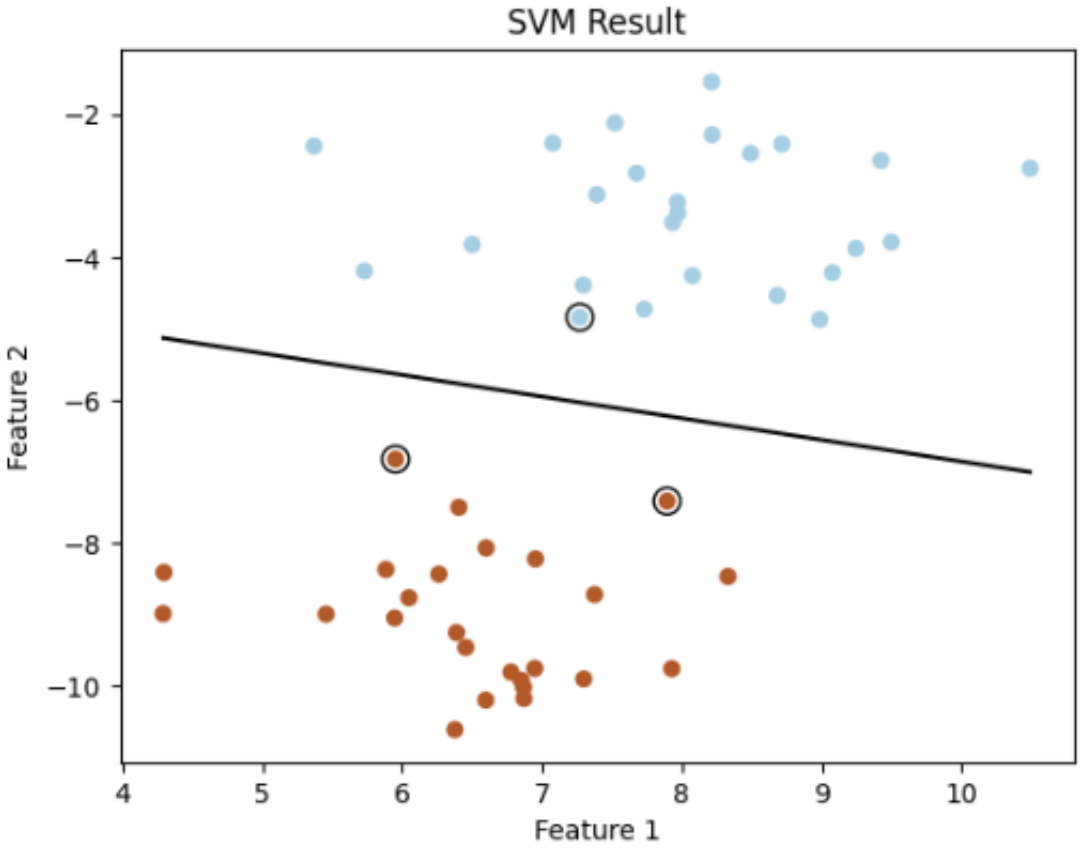
$x_1, \ldots x_n$ são as coordenadas do espaço n-dimensional.

$a_{1, \ldots}\ a_n$ são os coeficienter dos termos em cada dimensão.

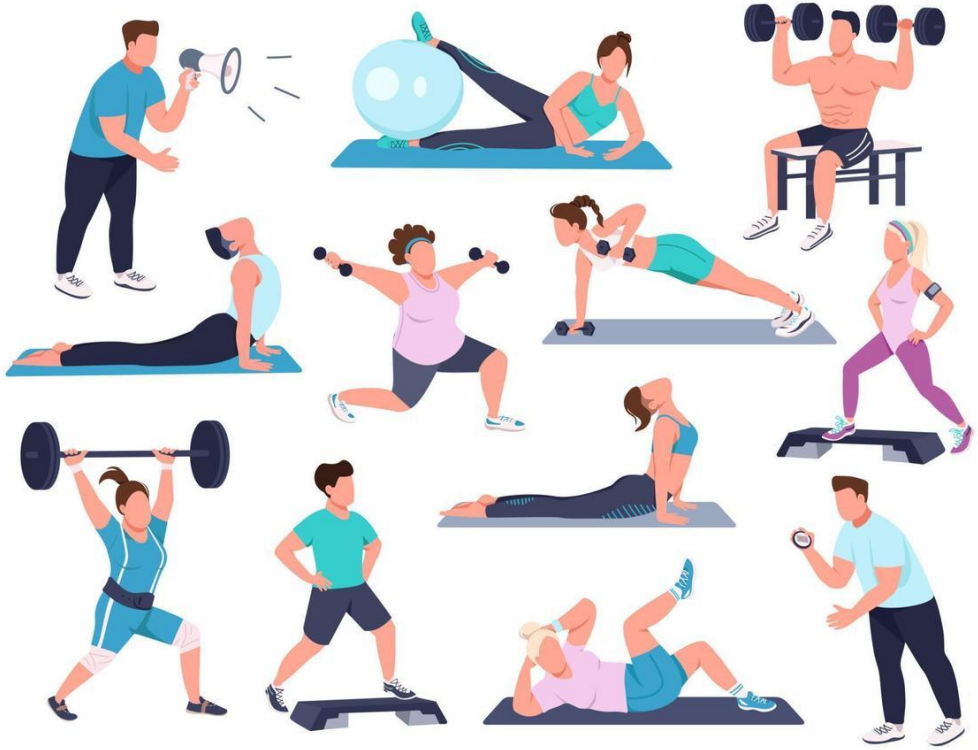
B é uma constante

O hiperplano divide o espaço em duas regiões. Em um espaço *n*-dimensional, ele separa o espaço em duas partes de dimensões *n* - uma em cada lado do hiperplano. Isso é útil em algoritmos de aprendizado de máquina, como as Máquinas de Vetores de Suporte (SVM), onde um hiperplano é usado para separar pontos de dados em diferentes classes.

MODELAGEM PREDITIVA E MACHINE LEARNING



EXERCÍCIO



MODELAGEM PREDITIVA E MACHINE LEARNING

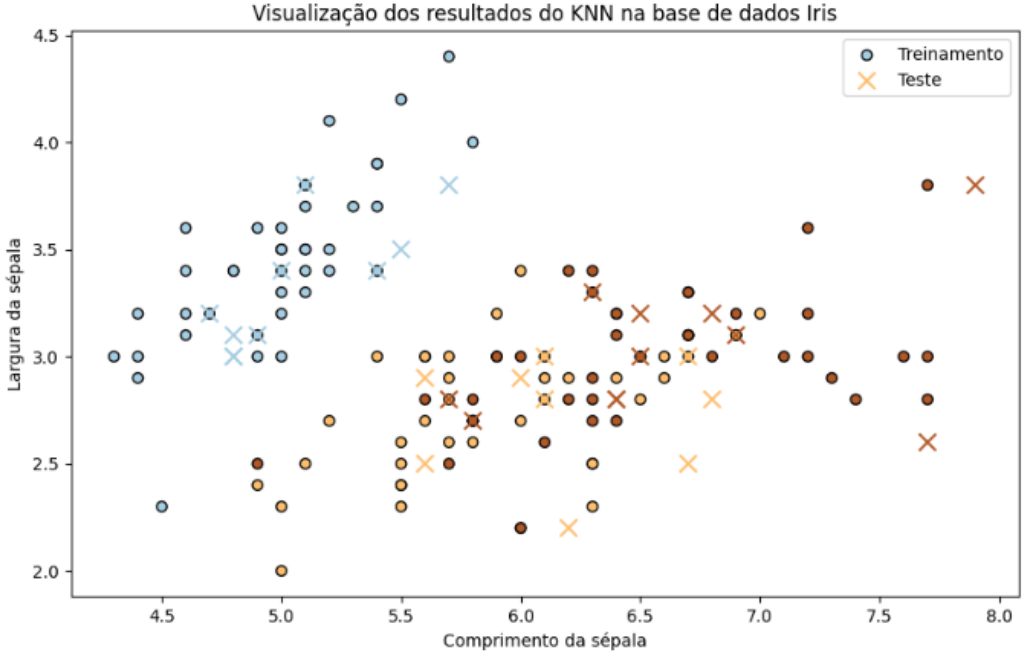
O algoritmo de K-vizinhos mais próximos (KNN - K-Nearest Neighbors) é um método de classificação ou regressão supervisionada, comumente usado em aprendizado de máquina. A ideia básica por trás do KNN é determinar a classe de um ponto de dados desconhecido com base na classe dos pontos de dados vizinhos a ele.

O KNN é um algoritmo simples e intuitivo, adequado para conjuntos de dados pequenos a moderadamente grandes, mas pode ser computacionalmente custoso para grandes conjuntos de dados, pois requer calcular a distância entre o ponto de dados de teste e todos os pontos de dados no conjunto de treinamento.

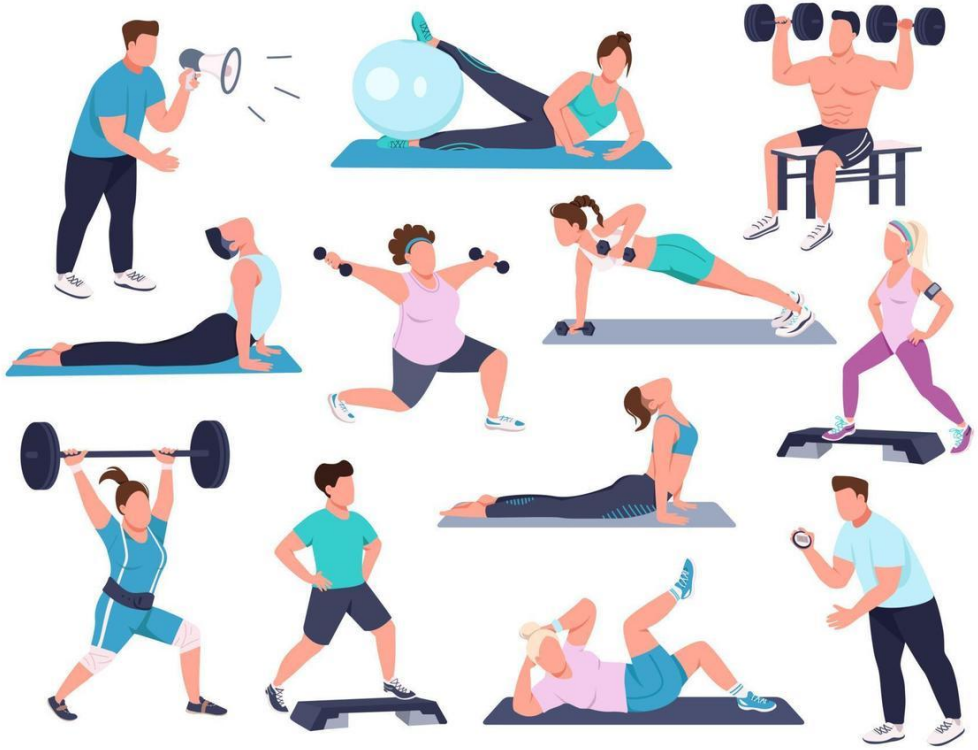
```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn import datasets
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score

# Carregar o conjunto de dados Iris
iris = datasets.load_iris()
X = iris.data[:, :2]  # Pegar apenas as duas primeiras características para simplificar a visualização
y = iris.target

# Dividir o conjunto de dados em conjuntos de treinamento e teste
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Inicializar o classificador KNN
k = 5  # Número de vizinhos
knn = KNeighborsClassifier(n_neighbors=k)

# Treinar o classificador KNN
knn.fit(X_train, y_train)

# Fazer previsões no conjunto de teste
y_pred = knn.predict(X_test)

# Calcular a precisão do modelo
accuracy = accuracy_score(y_test, y_pred)
print("Acurácia do modelo KNN:", accuracy)

# Plotar os resultados
plt.figure(figsize=(10, 6))

# Plotar os pontos de dados de treinamento
plt.scatter(X_train[:, 0], X_train[:, 1], c=y_train, cmap=plt.cm.Paired, label='Treinamento', edgecolors='k')

# Plotar os pontos de dados de teste
plt.scatter(X_test[:, 0], X_test[:, 1], c=y_pred, cmap=plt.cm.Paired, label='Teste', marker='x', s=100)

# Criar uma legenda
plt.legend()

# Adicionar rótulos aos eixos
plt.xlabel('Comprimento da sépala')
plt.ylabel('Largura da sépala')

# Adicionar um título
plt.title('Visualização dos resultados do KNN na base de dados Iris')

plt.show()
```

MODELAGEM PREDITIVA E MACHINE LEARNING

Para uma tarefa de classificação, o KNN calcula a classe mais comum entre os K vizinhos mais próximos e atribui essa classe ao ponto de dados de teste. Para uma tarefa de regressão, em vez de atribuir uma classe, o KNN calcula uma média ou uma média ponderada dos valores dos K vizinhos mais próximos e atribui esse valor ao ponto de dados de teste.



EXERCÍCIO



MODELAGEM PREDITIVA E MACHINE LEARNING

O modelo Naive Bayes é um algoritmo de classificação probabilístico baseado no teorema de Bayes, com uma suposição "ingênua" de independência entre os recursos (preditores) do conjunto de dados. Essa suposição simplifica bastante o cálculo da probabilidade condicional e torna o modelo computacionalmente eficiente, especialmente em conjuntos de dados grandes.

O teorema de Bayes descreve a probabilidade de um evento, com base no conhecimento prévio das condições que podem estar relacionadas a esse evento. Em termos de classificação, o modelo Naive Bayes calcula a probabilidade de que uma determinada instância de dados pertença a uma determinada classe, com base nas probabilidades das características observadas.

MODELAGEM PREDITIVA E MACHINE LEARNING

O algoritmo Naive Bayes funciona com base no teorema de Bayes, que é uma maneira de calcular a probabilidade condicional de um evento ocorrer, dado que outro evento já ocorreu. Aqui está a fórmula do teorema de Bayes:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

P(A|B) é a probabilidade de A ocorrer dado B.

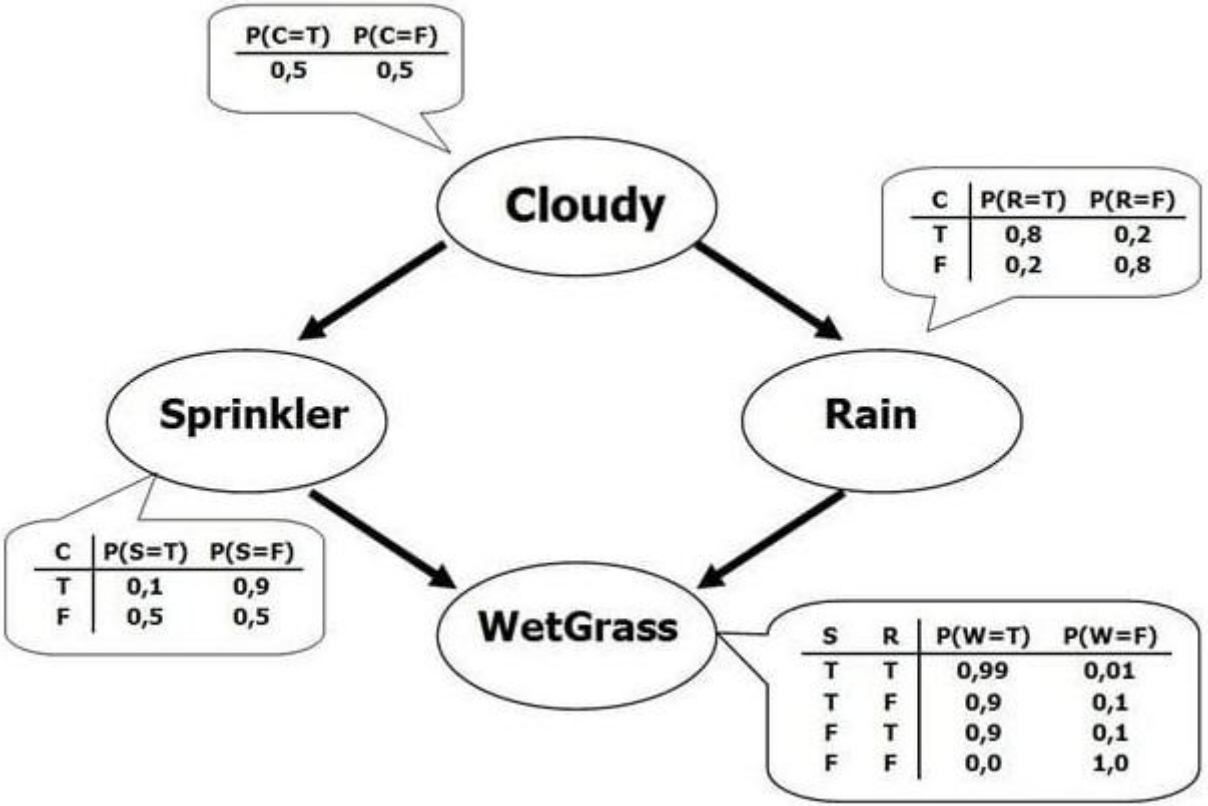
P(B|A) é a probabilidade de B ocorrer dado A.

P(A) e P(B) são as probabilidades marginais de A e B.

Thomas Bayes
1702 - 1761

Naive Bayes assume que as características são condicionalmente independentes, dada a classe. Isso significa que a presença de uma característica em particular em um conjunto de dados não afeta a presença de outra característica. Essa é uma suposição simplificadora, mas muitas vezes razoável em muitos problemas do mundo real.

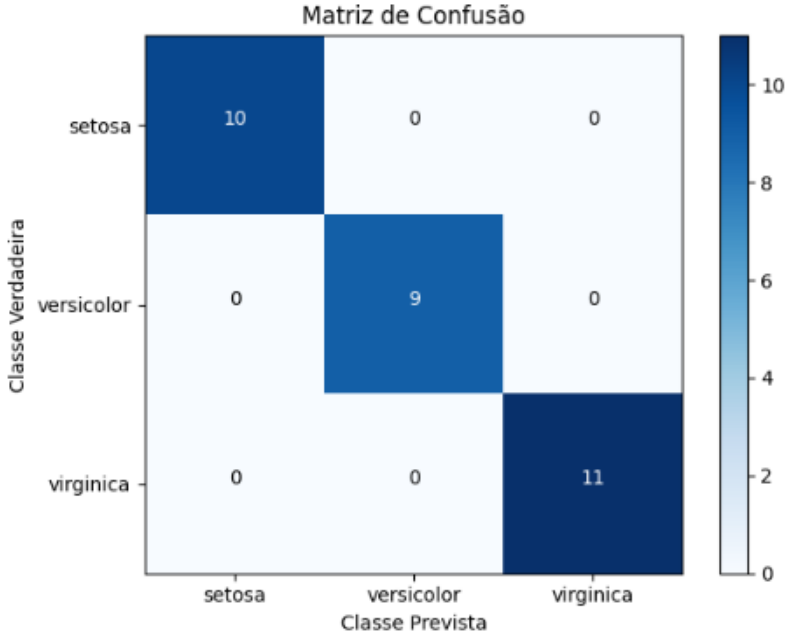
MODELAGEM PREDITIVA E MACHINE LEARNING



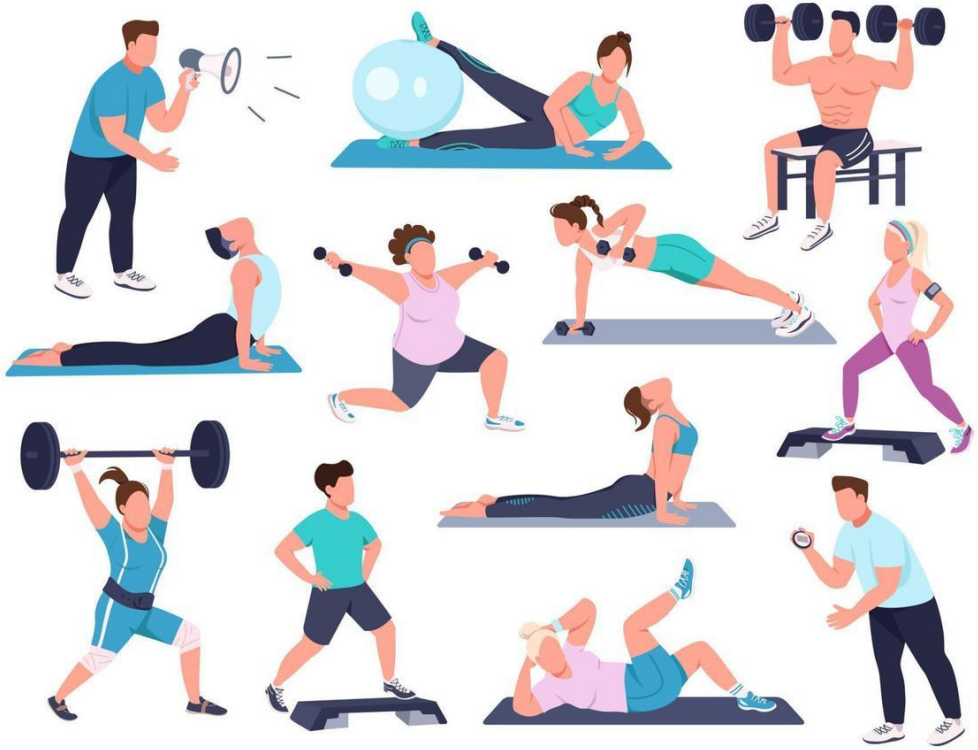
```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score, confusion_matrix

# Carregando o conjunto de dados Iris
iris = load_iris()
X = iris.data
y = iris.target

# Dividindo o conjunto de dados em conjunto de treinamento e teste
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Criando e treinando o modelo Naive Bayes Gaussiano
model = GaussianNB()
model.fit(X_train, y_train)

# Fazendo previsões no conjunto de teste
y_pred = model.predict(X_test)

# Calculando a acurácia do modelo
accuracy = accuracy_score(y_test, y_pred)
print("Acurácia do modelo:", accuracy)

# Plotando a matriz de confusão
conf_matrix = confusion_matrix(y_test, y_pred)
plt.imshow(conf_matrix, interpolation='nearest', cmap=plt.cm.Blues)
plt.title('Matriz de Confusão')
plt.colorbar()
classes = iris.target_names
tick_marks = np.arange(len(classes))
plt.xticks(tick_marks, classes)
plt.yticks(tick_marks, classes)

for i in range(len(classes)):
    for j in range(len(classes)):
        plt.text(j, i, conf_matrix[i, j],
                 horizontalalignment="center",
                 color="white" if conf_matrix[i, j] > conf_matrix.max() / 2 else "black")

plt.ylabel('Classe Verdadeira')
plt.xlabel('Classe Prevista')
plt.tight_layout()
plt.show()
```

MODELAGEM PREDITIVA E MACHINE LEARNING

o algoritmo Naive Bayes calcula a probabilidade de uma instância pertencer a uma classe específica com base nas probabilidades das características observadas, assumindo que as características são independentes entre si, dada a classe. Ele é computacionalmente eficiente e muitas vezes produz bons resultados em uma variedade de problemas de classificação.



EXERCÍCIO



MODELAGEM PREDITIVA E MACHINE LEARNING

Os modelos supervisionados são treinados com pares de entrada-saída para aprender a prever ou classificar novas entradas, enquanto os modelos não supervisionados exploram a estrutura dos dados sem a necessidade de rótulos ou saídas conhecidas.

Supervisionado:

Em modelos supervisionados, o algoritmo é treinado em um conjunto de dados que inclui tanto os inputs (características ou atributos) quanto os outputs (rótulos ou variáveis dependentes) correspondentes.

O objetivo é aprender a relação entre os inputs e os outputs, de modo que, quando apresentado com novos inputs, o modelo possa prever ou classificar os outputs com base no que foi aprendido durante o treinamento.

Exemplos de modelos supervisionados incluem regressão linear, árvores de decisão, random forest, SVM (Support Vector Machines), entre outros.

MODELAGEM PREDITIVA E MACHINE LEARNING

Um exemplo de dados supervisionados seria um conjunto de dados de previsão de preços de imóveis. Neste conjunto de dados, cada instância (linha) representa um imóvel e é composta por características como área, número de quartos, localização, etc. Além disso, cada instância também inclui o preço de venda do imóvel, que é a variável que queremos prever.

Portanto, neste exemplo:

- As características do imóvel (área, número de quartos, localização, etc.) são os atributos de entrada, também chamados de features ou variáveis independentes.
- O preço de venda do imóvel é o atributo de saída, também conhecido como variável dependente ou rótulo.

MODELAGEM PREDITIVA E MACHINE LEARNING

Não supervisionado:

Em modelos não supervisionados, o algoritmo é treinado em um conjunto de dados que contém apenas os inputs, sem os correspondentes outputs ou rótulos. O objetivo é explorar a estrutura ou padrões nos dados, sem guia externo sobre o que procurar. O algoritmo tenta encontrar padrões naturais ou agrupamentos nos dados. Não há uma variável alvo específica a ser prevista ou otimizada durante o treinamento.

Exemplos de modelos não supervisionados incluem o algoritmo k-means para clustering, análise de componentes principais (PCA), algoritmos de redução de dimensionalidade, entre outros.

MODELAGEM PREDITIVA E MACHINE LEARNING

Um exemplo de dados não supervisionados seria um conjunto de dados de segmentação de clientes para uma empresa de varejo. Neste conjunto de dados, cada instância representa um cliente e é composta por características como idade, gênero, renda, histórico de compras, entre outros.

No entanto, ao contrário do exemplo supervisionado, neste caso, não temos um atributo de saída específico que estamos tentando prever. Em vez disso, o objetivo é explorar a estrutura dos dados para identificar grupos naturais de clientes com características semelhantes, sem a necessidade de rótulos ou categorias predefinidas.

Portanto, neste exemplo:

As características dos clientes (idade, gênero, renda, etc.) são os atributos de entrada, também chamados de features ou variáveis independentes.

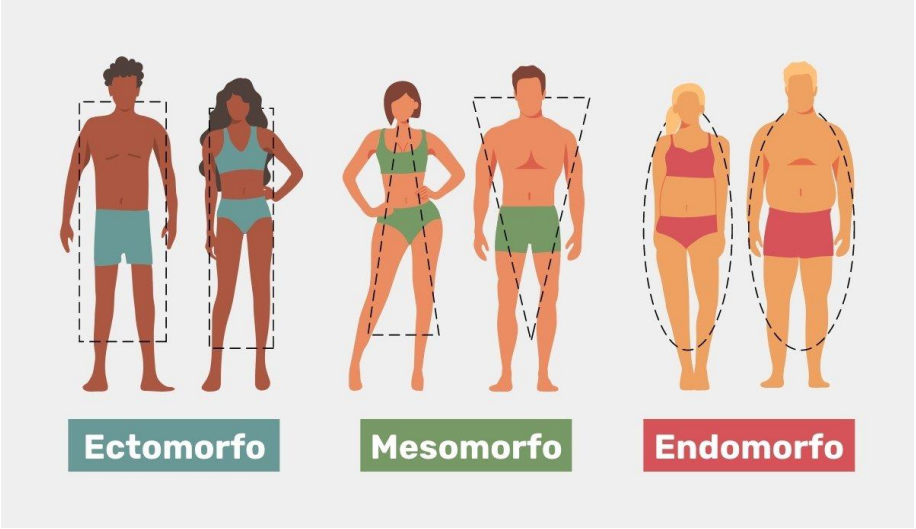
Não há um atributo de saída específico; em vez disso, estamos interessados em encontrar padrões ou grupos nos dados que possam nos fornecer insights sobre o comportamento dos clientes e auxiliar em estratégias de marketing, por exemplo.

MODELAGEM PREDITIVA E MACHINE LEARNING

Esta seção aborda um dos pontos mais críticos onde projetos de Data Science falham: **confiar na métrica errada ou em uma avaliação mal estruturada**, resultando em modelos que parecem ótimos no papel, mas fracassam no mundo real.

Avaliação de modelos

- Holdout vs Cross-validation
- Métricas:
    - Accuracy (quando NÃO usar)
    - Precision / Recall
    - F1-score
    - ROC-AUC
- Matriz de confusão como ferramenta de decisão
- Threshold ≠ 0.5 (impacto organizacional)



MODELAGEM PREDITIVA E MACHINE LEARNING

Overfitting e underfitting (ponte para redes neurais)

- Curvas de aprendizado
- Complexidade do modelo
- Bias–variance tradeoff
- Regularização (conceitual)

DÚVIDAS ou PERGUNTAS?

MUITO OBRIGADO!!!!

daniel.ohata@facens.br

REFERÊNCIAS

BISONG, Ekaba; BISONG, Ekaba. Google colaboratory. **Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners**, p. 59-64, 2019.

DE ANDRADE, Amanda Figueiredo et al. A ÉTICA NO USO DA INTELIGÊNCIA ARTIFICIAL E SEUS RISCOS JURÍDICOS. Revista Acadêmica Online, v. 11, n. 56, p. e1400-e1400, 2025.

ISZCZUK, Ana Claudia Duarte et al. Evoluções das tecnologias da indústria 4.0: dificuldades e oportunidades para as micro e pequenas empresas. **Brazilian Journal of Development**, v. 7, n. 5, p. 50614-50637, 2021.

MARCONI, Francesco. Artificial Intelligence Backbone. < https://twitter.com/fpmarconi/status/794208040207740928 >, 2016 Acesso 01/01/2019.

MICHALSKI, Ryszard Stanislaw; CARBONELL, Jaime Guillermo; MITCHELL, Tom M. (Ed.). **Machine learning: An artificial intelligence approach**. Springer Science & Business Media, 2013.

ROSÁRIO, João Maurício. **Automação industrial**. Editora Baraúna, 2012.

RUSSELL, Stuart J.; NORVIG, Peter. **Artificial intelligence: a modern approach**. Malaysia; Pearson Education Limited, 2016.

SANTOS, Gustavo Soares. Novas Tecnologias Aplicadas na Construção Civil: Conceitos da Indústria 4.0. RCT-Revista de Ciência e Tecnologia, v. 8, 2022.

ZHOU, Zhi-Hua. **Machine learning**. Springer nature, 2021.