Inteligência Artificial

Guilherme Henrique de Souza Nakahata

Universidade Estadual do Paraná - Unespar

13 de Abril de 2024

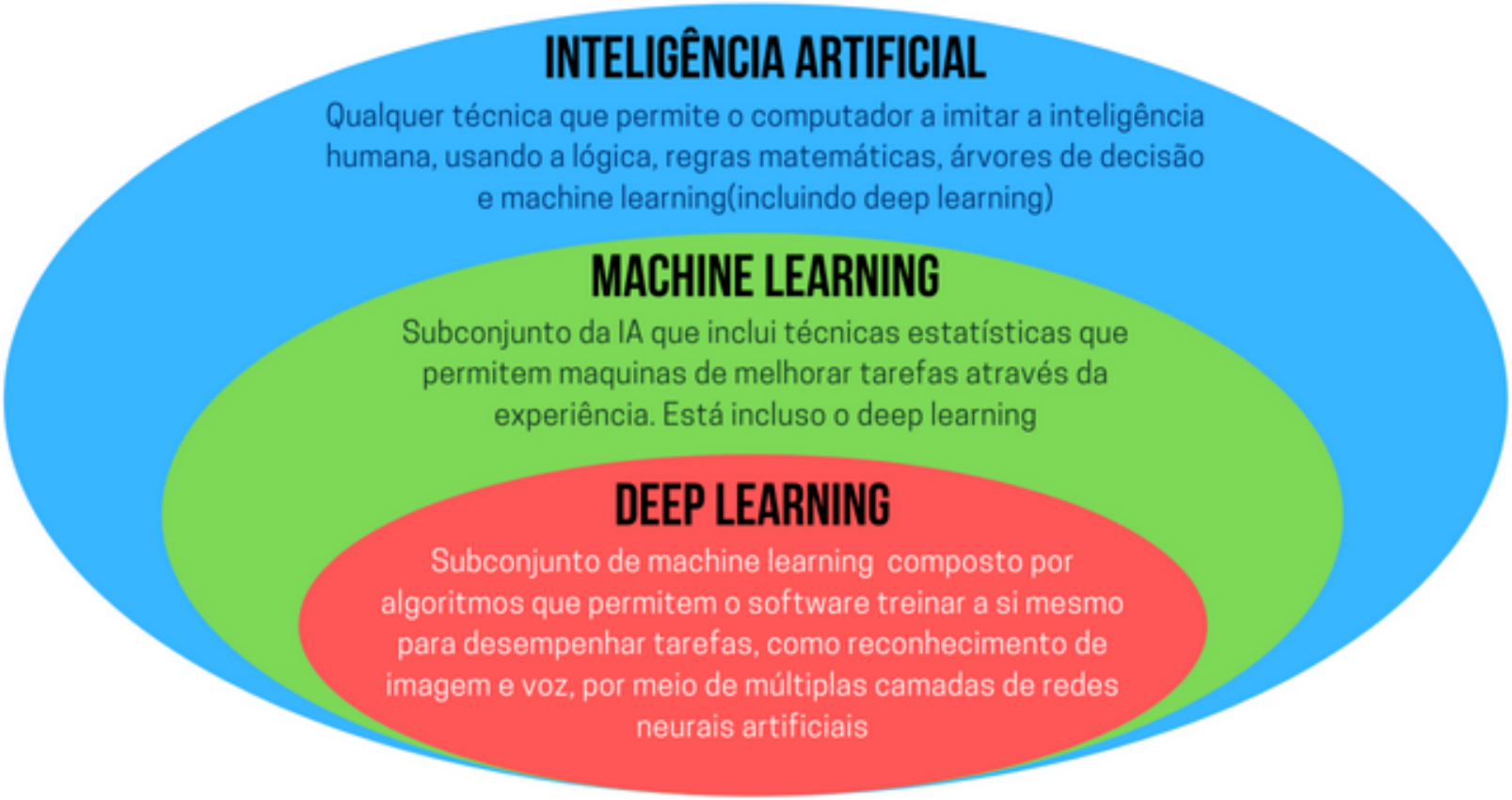
Aprendizado de máquina

- Subárea da Inteligência Artificial;
- Desenvolver modelos;
- Técnicas;
- Construção de **sistemas**;

Aprendizado de máquina

- Subárea da Inteligência Artificial;
- Desenvolver modelos;
- Técnicas;
- Construção de **sistemas computacionais**;
- Relacionados à capacidade de aquirir conhecimento a partir de dados;

Aprendizado de máquina



Aprendizado de máquina

- Em geral usamos Aprendizagem de máquina em diferentes situações:
  - Existe um padrão a ser aprendido;
  - Existe dados disponíveis sobre o problema em questão;

Aprendizado de máquina

- Aprendendo a partir de dados;
- Dados são baratos e abundantes;
  - Transações de todos os clientes de um banco;
- Conhecimento é caro e escasso;
  - Qual é o perfil de clientes com baixo risco para a concessão de crédito?

Aprendizado de máquina

- Algoritmos de AM são indutores:
  - Induzem uma função ou hipótese capaz de resolver um problema;
  - Empregam um princípio de inferência (**indução**);
  - A partir de dados do problema já resolvido se obtém uma função que será utilizado para resolver instâncias futuras não vistas.

Tipos de Aprendizagem

- Supervisionada;
- Não supervisionada;
- Semi-supervisionada;
- Aprendizagem por Reforço;

Tipos de Aprendizagem

- Supervisionada
  - Treina um modelo usando um conjunto de dados rotulados;
  - O objetivo é prever a saída a partir de novas entradas com base nas entradas e saídas anteriores;
- Não supervisionada
  - Treina um modelo usando dados que não possuem rótulos;
  - O objetivo é encontrar padrões ou estruturas ocultas nos dados;
- Semi-supervisionada
  - Utiliza uma pequena quantidade de dados rotulados e uma grande quantidade de dados não rotulados;
  - Combinando os benefícios de aprendizagem supervisionada e não supervisionada.
- Aprendizagem por Reforço
  - Treina um agente para tomar decisões em um ambiente;
  - Recebendo recompensas ou penalidades com base em suas ações;
  - Ajusta suas estratégias para maximizar a recompensa total.

Preparação de dados

- Preparação de dados;
- Transformação de dados brutos;
- Formato mais apropriado;
- Aprendizado;

Conjunto de dados

- Formados por objetos;
- Representam entidades físicas ou abstratas;
- Podem ser gerados por diversos processos:
  - Transações financeiras;
  - Monitoramento ambiental;
  - Registros clínicos;
  - Navegação;
- Assumir vários formatos:
  - Valores numéricos:
  - Simbólicos;
  - Textos;
  - Imagens;
  - Áudios.

Instância

- Os objetos de um conjunto são chamados de instâncias, amostras ou exemplos;
- Conjunto de atributos ou características;
- Vetor de atributos o conjunto de valores que definem os atributos de um exemplo.

Atributo

- Atributo ou Característica descreve uma propriedade do exemplo:
  - Pode ser valor categórico ou número;
  - Pode ter valores desconhecidos, ou que não-se-aplica.
- Atributos numéricos podem ser discretos ou contínuos:
  - **Discreto**: Assume um número finito de valores entre quaisquer dois valores;
  - **Contínuo**: Assume um número infinito de valores entre quaisquer dois valores;

Exemplo

	Idade	*Diagnóstico*	*Astigmatismo*	*Taxa lacrimal*	*Lente*
1	infantil	miopia	não	reduzida	nenhuma
2	infantil	miopia	sim	normal	gelatinosa
3	infantil	hipermetropia	não	normal	gelatinosa

Atributo Meta

- Tarefas preditivas possuem atributo especial denominado atributo meta ou atributo alvo;
- Descreve o que se deseja aprender e fazer previsões;
- Classificação Binária ou Multi classe;
- Classificação multi rótulo;

Id	Débito	Colateral	Risco
1	Alto	Nenhum	Alto
2	Alto	Nenhum	Alto
3	Baixo	Nenhum	Moderado
4	Baixo	Nenhum	Alto
5	Baixo	Nenhum	Baixo
6	Baixo	Adequado	Baixo
7	Baixo	Nenhum	Alto
8	Baixo	Adequado	Moderado

Treino e Teste

- A indução do modelo é chamado de **treinamento**;
- Objetivo é aprender a partir dos exemplos;
- A avaliação do modelo é chamado de **teste**;
- Prever atributos meta dos exemplos de testes;
- Métricas de desempenho são coletadas;
- Estimar o desempenho do modelo na previsão de casos novos.

Avaliação de modelos

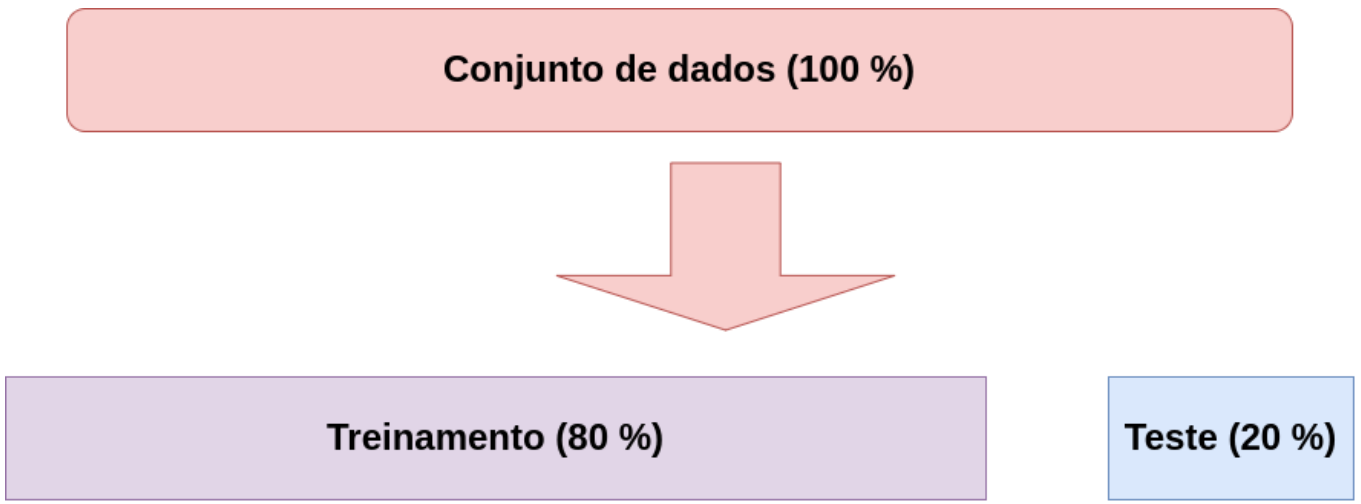
- Não é utilizado os mesmos exemplos utilizado no aprendizado e na avaliação;
- Precisamos dividir os exemplos em conjuntos de treino e teste;
- Duas estratégias mais utilizadas nos modelos supervisionados:
  - Holdout;
  - Validação cruzada;
- Ambos são relacionados à amostragem, visando a avaliação.

Holdout

- Consiste em dividir os exemplos em uma porcentagem fixa:
  - Exemplos;
  - Treinamentos;
- Valores comuns de divisão:
  - 80 % para treino e 20 % de teste;
  - 66 % para treino e 33 % de teste;
- Valores empíricos;
- Quanto maior a amostra de teste, maior a confiança;
- Não é uma boa estrategia para conjuntos pequenos;

Holdout

- Os exemplos do conjuntos de testes devem ser selecionadas aleatoriamente;
- Simples;
- Estratificada;



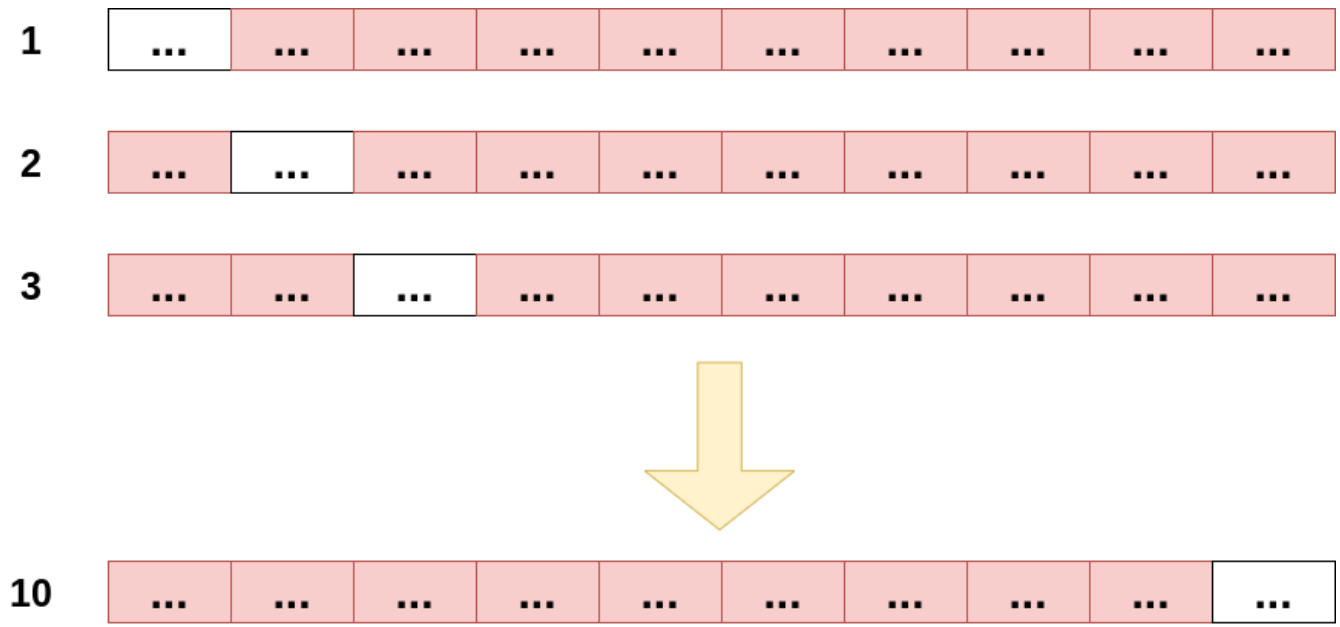
Validação Cruzada

- Diminuir o impacto das diferentes amostras;
- Executar k rodadas de holdout;
- Média dos k resultados;
- Validação cruzada sistematiza a execução;
- Amostra de teste diferentes;
- Garante que ao final do processo todos os exemplos terão sido utilizados para teste;

Validação Cruzada

- É o método de avaliação mais adequado para pequenos conjuntos;
- Dividir os exemplos em k partições de tamanho próximos;
- Utilizar os *k-1* fold para treinamento e avaliar o modelo com o fold remanescente;
- Repetir o processo k vezes;
- Valor comum para k é 10;

Avaliação Cruzada



Avaliação Cruzada

- Processo de amostragem:
  - Simples;
  - Estratificada;
- Validação cruzada estratificada;
- As classes são consideradas durante a amostragem das partições;
- As partições terão valores próximos durante a distribuição das classes;

Análise de dados

- Informações úteis;
- Extraídas do conjunto de dados por meio de exploração;
- Podendo ser utilizada no:
  - Pré processamento;
  - Aprendizado;
  - Interpretação dos resultados;
- Estatística descritiva:
  - Resumem de forma quantitativa as principais características do conjunto:
    - Frequência;
    - Tendência central;
    - Dispersão.

Frequência

- Mede a proporção de vezes que um atributo assume um dado valor;
- Tarefas de classificação a frequência das classes é importante;
- Frequências muito diferentes, o conjunto é **desbalanceado**;
- Distribuição das classes do conjunto.

Distribuição de Classes

- Exemplo: um conjunto de 150 exemplos com 100 exemplos da classe A e 50 exemplos da classe B, tem a seguinte distribuição de classes:
  - distribuição$(A, B) = (0,67, 0,33) = (67\%, 33\%)$
- A classe A é chamada de **majoritária**;
- A classe B é chamada de **minoritária**;

Medidas de posição

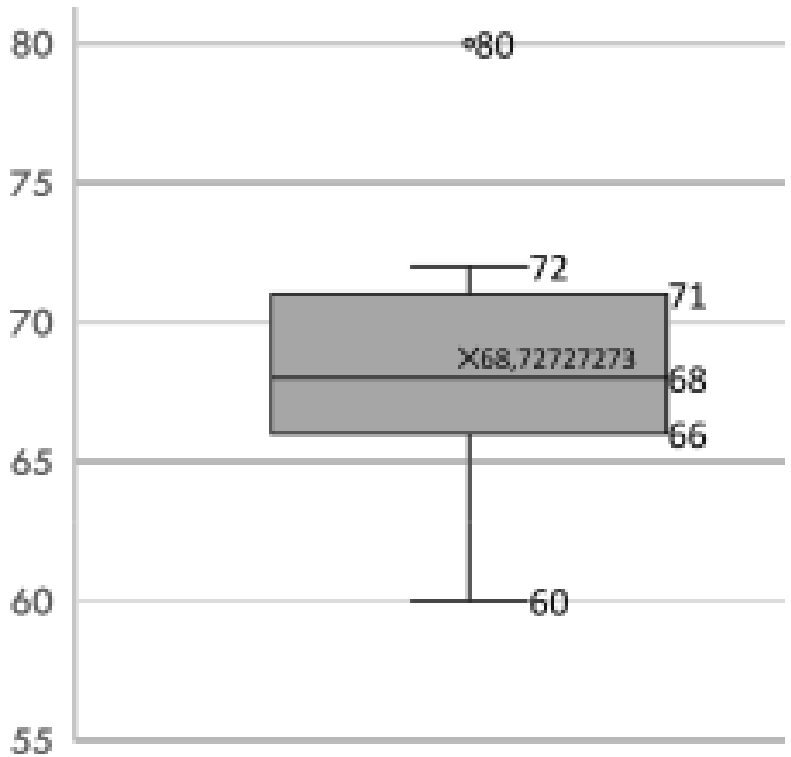
- Média;
- Mediana;
- Moda;
- Separatrizes;
- Quartil;
  - 1° Quartil: 25 %;
  - 2° Quartil: Mediana;
  - 3° Quartil: Valor que tem 75 % dos demais valores abaixo dele;
- Percentil.

Medidas de dispersão

- Medem a variabilidade (espalhamento):
  - Amplitude;
  - Variância;
  - Desvio padrão;
  - Intervalo interquartil (IQR).

Boxplot

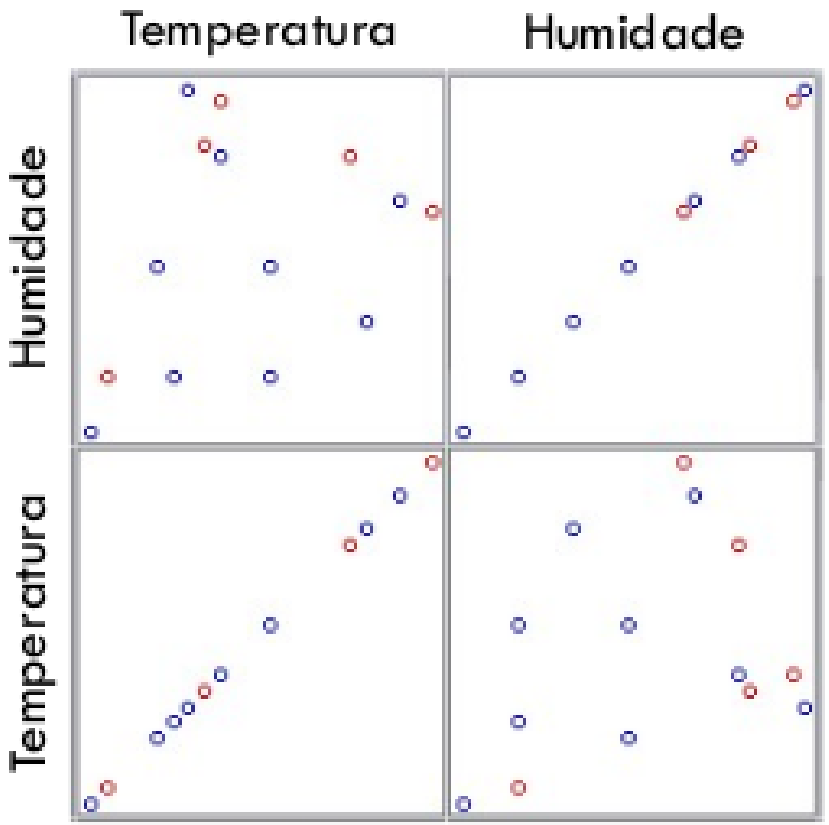
- Diagramas de caixa (Boxplot);
- Representa a distribuição empírica dos dados;
- As caixas representam o 1° e 3° quartil e a mediana;
- Hastes inferiores e superiores estendem até o limite inferior e superior;



Dados multivariados

- Análise de cada atributo de forma individual;
- Analisar a relação de par-a-par entre atributos;
- Medidas de dispersão;
- Covariância;
  - Mede o grau com que esses atributos variam juntos;
- Correlação;
  - Corrige a covariância retirando a influência da magnitude dos atributos.
- Gráfico para visualizar correlação;
- Scatter plot.

Scatter plot



Pré-Processamento

- Melhorar a qualidade dos dados;
- Possibilita a indução de modelos melhores;
- Reduz complexidade computacional;
- Adequar os dados para determinado algoritmo;
  - Algoritmos que só aceitar atributos numéricos;

Pré-processamento

- Amostragem;
- Transformação;
- Limpeza;
- Seleção de atributos;

Pré-processamento

- Amostragem;
  - Dificuldades em lidar com volume alto de instâncias;
  - Baseados em distâncias;
  - Problemas de memória;
  - Maior quantidade de dados tende a aumentar o desempenho;
  - Diminui a eficiência computacional do processo indutivo;
- Amostragem aleatória simples:
  - Escolhe aleatoriamente n elementos que farão parte da amostra;
- Amostragem aleatória estratificada:
  - Respeita a mesma distribuição de classes.

Dados desbalanceados

- Comum haver dados desbalanceados;
- Os algoritmos têm problema de desempenho com dados desbalanceados;
- Tendem a favorecer a classificação na classe majoritária;
- Reamostragem:
  - Subamostragem: Remove instâncias da classe majoritária;
  - Sobreamostragem: Adiciona instâncias da classe minoritária;

Transformação

- Vários algoritmos de aprendizagem de máquina são limitadas à manipulação de determinados tipos:
  - Valores números;
  - Valores simbólicos;
- Conversão de tipos;

Transformação

- Codificação one-hot;

Exemplo	Cor	Fabricante	#portas
#1	Branco	Honda	4P
#2	Verde	Honda	2P
#3	Preto	Ford	4P
#4	Preto	GM	4P

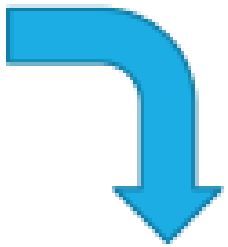
Codificaçãc
one-hot

Exemplo	Cor = Branco	Cor = Verde	Cor = Preto	Fabr = Honda	Fabr = Ford	Fabr = GM	#porta = 4P	#porta = 2P
#1	1	0	0	1	0	0	1	0
#2	0	1	0	1	0	0	0	1
#3	0	0	1	0	1	0	1	0
#4	0	0	1	0	0	1	1	0

Transformação

- Codificação one-hot;

Exemplo	Temperatura
#1	Baixa
#2	Média
#3	Alta
#4	Muito Alta



Codificação termômetro

Exemplo	Temp1	Temp2	Temp3
#1	0	0	0
#2	0	0	1
#3	0	1	1
#4	1	1	1

Limpeza de dados

- Amenizar problemas advindos de processos imprecisos de aquisição de dados;
  - Valores ausentes;
  - Valores ruidosos;
- Remoção das instâncias com valores ausentes;
- Preenchimento manual dos valores;
  - Reexecutar o processo;
  - Especialista do domínio;
- Preenchimento automático dos valores;
  - Atribuição de um valor constante (?, desconhecido, !);
  - Valor médio;
  - Utilizar algum modelo preditivo.

Valores ruidosos

- Consistem em valores muito diferentes dos demais;
- Casos atípicos;
- Erros de aquisição;
- Inspeção e correção manual;
- Identificação e limpeza automática;
- Redundância;
- Eliminação de exemplos redundantes;
- Atributos redundantes;

Atividade

- Implementar um algoritmo que:
  - Leia um arquivo CSV;
  - Identifique e remova instâncias repetidas;
  - Trate valores ausentes (Imputação ou remoção);
  - Transforme atributos categóricos;
  - Divida dos dados em conjuntos de treinamento e teste;

Obrigado!
Dúvidas?

Guilherme Henrique de Souza Nakahata

guilhermenakahata@gmail.com

https://github.com/GuilhermeNakahata/UNESPAR-2024