

UNIDADE II

Ciência de Dados

Prof. Me. João Cruz

Relembrando

O que é Ciência de Dados?

 A Ciência de Dados é um campo interdisciplinar que combina técnicas estatísticas, matemáticas e programação para extrair insights e conhecimentos úteis a partir de conjuntos de dados complexos.

Etapas

- Coleta dos dados;
- Limpeza dos dados;
- Exploração.

Objetivo

Conseguir insigths.

Antes de falarmos em pré-processamento...



Fonte: https://br.freepik.com/vetores-gratis/conceito-de-pesquisa-para-a-pagina-de-destino_5060354.htm#query=INVESTIGADOR &position=0&from_view=search&track=sph

Principais fontes de dados



Bases de dados internas



Bases de dados públicas



Redes sociais e plataformas on-line



Sensores e dispositivos conectados



Dados de pesquisas e estudos



Dados externos de terceiros



Web scraping

Antes de falarmos em pré-processamento...

Preparando os dados para a apresentação.



Coleta de Dados



Limpeza de Dados



Valores Ausentes



Valores Ruidosos



Valores Inconsistentes



Organização dos dados



Apresentação dos dados

Como elaborar a preparação e o pré-processamento de dados

Objetivo

 Transformar e organizar os dados brutos em um formato adequado para análise e modelagem.

Tarefas da preparação e pré-processamento de dados:

- Limpeza de dados;
- Normalização e padronização;
- Codificação de variáveis categóricas;
- Redução de dimensionalidade;
- Tratamento de outliers;
- Divisão de dados;
- Remoção de stopwords;
- Stemming.

Limpeza de dados

Objetivo

Identificação e tratamento de dados ausentes, inconsistentes ou incorretos.

Como fazer

- Dados ausentes: imputação.
- Dados inconsistentes ou incorretos:
 Corrigidos ou removidos.

1 2 3 4

1 2 3 4

Normalização e padronização

Objetivo

 Definir um padrão entre dados de origens diferentes, ou representados em escadas diferentes.

Como fazer

1..n

1..n

1..n*(-1)

1

Normalização e padronização

Normalização Min-Max

Os valores dos dados são ajustados para um intervalo específico, geralmente entre 0 e 1.

Técnicas de normalização

Normalização Z-score

Ajusta os dados de forma que tenham uma média 0 e 1 com desvio padrão de 1.

Normalização por escala logarítmica

A normalização por escala logarítmica pode ser aplicada para reduzir a variação e tornar os dados mais comparáveis

Remoção de stopwords

Objetivo

- Desenvolver um padrão para um conjunto de palavras que são focadas apenas em melhorar a compreensão fonética das frases.
- Normalmente, são <u>artigos</u>, <u>preposições</u>, <u>pronomes</u> e outras <u>palavras</u> <u>de ligação</u>.

Como fazer

Frase original

"O cachorro pulou sobre a cerca."

Stopword

"O cachorro pulou sobre a cerca."

"Cachorro pulou cerca."

Stemming

Objetivo

 Redução de palavras ao seu "stem" (raiz), removendo os sufixos e prefixos para encontrar a forma básica da palavra.

Lematização

Objetivo

- Redução de palavras à sua forma básica, considerando a morfologia da palavra.
- Mais complexa que o stemming.

Como fazer

CORRENDO Stemming CORR Lematização CORRER

Métodos de raspagem (web scraping)

Objetivo

- Técnica utilizada para extrair informações de páginas da web de forma automatizada.
- Essa técnica permite coletar dados de diferentes fontes on-line, de maneira eficiente e rápida, sem a necessidade de intervenção manual.

Exemplos de técnicas do método de raspagem

Análise de HTML

Leitura do código HTML das páginas.

Web scraping baseado em API

Informações constantes nas requisições feitas pelas API.

Uso de bibliotecas específicas

 Bibliotecas amplamente conhecidas como a Scrapy e a Selenium.

Raspagem de dados de mídias sociais

 Facebook, Twitter, Instagram, possuem API próprias para auxiliar na obtenção e análise de dados quantitativos.

Tabulação cruzada (tabulação de dados)

O que é

• É uma técnica estatística usada para resumir e analisar dados em forma de tabelas.

Objetivo

 Apresentar as frequências e proporções de cada combinação de valores das variáveis, permitindo identificar padrões, tendências e associações.

Onde é usada

 Essa técnica é utilizada em pesquisas de opinião, estudos de mercado, análise de dados quantitativos e em várias outras áreas.

Tabulação cruzada (tabulação de dados)

O processo de tabulação envolve os seguintes passos:



Identificação das variáveis



Organização dos dados



Contagem de frequências



Cálculo de proporções



Apresentação dos resultados

Seleção de atributos (seleção de variáveis)

O que é

Etapa que envolve a escolha dos atributos mais relevantes e informativos em um conjunto de dados para construir modelos mais eficientes e precisos.

Objetivo

Reduzir o número de atributos a serem considerados no modelo, sem comprometer a qualidade da análise.

Atributos

- Cor;
- Formato;
- Distância;
- Proximidade de centro.



Seleção de atributos (seleção de variáveis)

Existem diferentes técnicas de seleção de atributos, que podem ser divididas em **três categorias principais**:

Métodos baseados em filtros

Avaliar a relação entre cada atributo e a variável alvo, sem levar em consideração o modelo de aprendizado.

Exemplos: Análise de correlação, testes de independência, ganho de informação e índice de relevância.

Métodos baseados em wrappers

Um modelo de aprendizado é treinado de forma iterativa avaliando a performance com a modificação dos atributos.

Exemplos: Modelo de busca exaustiva, modelo de busca para frente e busca para trás.

Métodos incorporados

Incorpora a seleção de atributos diretamente no processo de treinamento do modelo.

Exemplo: Em algoritmo de árvores de decisão, quando são utilizados mecanismos internos para seleção automática dos atributos mais relevantes.



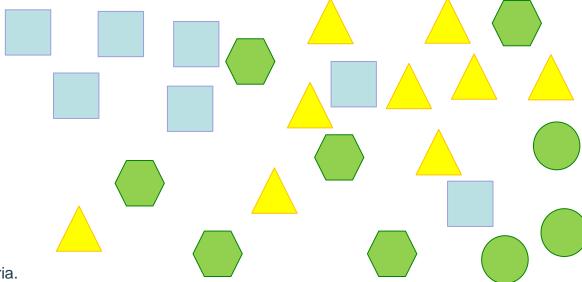
Engenharia de características

Objetivo

 Transformar dados brutos em um conjunto de características relevantes para a análise dos dados.

Como fazer

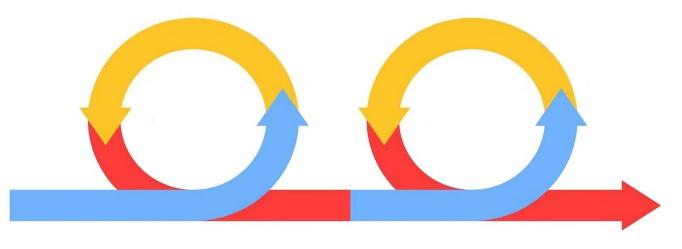
- Primeira abordagem: seleção de características dos dados para criação de subconjuntos relevantes.
- Segunda abordagem: criação de novas características com base no subconjunto.



Fonte: Autoria própria.

Engenharia de características

- A engenharia de características é um processo iterativo e criativo.
- Requer um bom entendimento do domínio do problema, além de uma análise aprofundada dos dados.



Fonte: Autoria própria.

Sempre avalie o desempenho dos modelos com diferentes conjuntos de características, utilizando técnicas de validação cruzada, para selecionar as melhores características e evitar overfitting.

Dados ausentes

O que são

 Quando não há valores disponíveis para uma ou mais variáveis em determinadas observações.

Consequência

Comprometer a qualidade da análise e dos modelos de aprendizado de máquina.

Razões

 Por causa de erros de coleta, problemas técnicos, recusa dos respondentes em fornecer certas informações, ou simplesmente porque os dados não foram registrados.

Dados ausentes

Como lidar com dados ausentes

- Exclusão de casos ou variáveis;
- Preenchimento com valor fixo;
- Preenchimento com valor estimado;
- Modelagem especializada.



Interatividade

Quais dos passos a seguir fazem parte do pré-processamento usando na tabulação cruzada?

- a) Padronização e indexação.
- b) Contagem de frequência e cálculo de proporções.
- c) Mapeamento de dados e limpeza de dados.
- d) Exercícios de sistemas e avaliação de probabilidades.
- e) Apresentação de tabelas e normalização de dados.

Resposta

Quais dos passos a seguir fazem parte do pré-processamento usando na tabulação cruzada?

- a) Padronização e indexação.
- b) Contagem de frequência e cálculo de proporções.
- c) Mapeamento de dados e limpeza de dados.
- d) Exercícios de sistemas e avaliação de probabilidades.
- e) Apresentação de tabelas e normalização de dados.

O que são

 São algoritmos ou sistemas que utilizam dados históricos para fazer previsões ou estimativas sobre eventos futuros.

Objetivo

 Identificar padrões e relações nos dados disponíveis e usá-los para prever o resultado de um evento futuro.

Para que servem

 Aprendizado de máquina, estatística de dados, inteligência artificial, obtenção de insights e tomar decisões informadas.

Exemplos de modelos preditivos:







Arvores de Decisão



Redes **Neurais**



Máquinas de Vetores de **Suporte**



Naive Bayes



Random **Forest**

Fonte: Autoria própria.

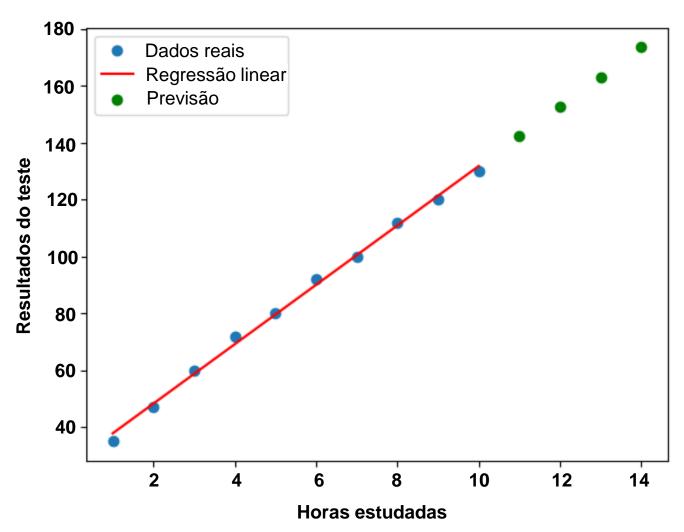
 Esses modelos podem ser aplicados a uma ampla variedade de problemas, como previsão de vendas, detecção de fraudes, previsão de demanda, diagnóstico médico, recomendação de produtos e muito mais.

Desafios da implementação dos modelos preditivos

- Criação de modelos treinados a partir de dados históricos;
- Compreender as vantagens e desvantagens de cada modelo;
- A necessidade de identificar quais variáveis ou recursos relevantes para tarefas de previsão;
- Seleção cuidadosa dos recursos que resultem em um modelo preciso e eficiente;
- Um trabalho de transformação e pré-processamento com muita qualidade;
- Cuidado para evitar overfitting e underfitting;
- Avaliação do modelo e dos resultados do treinamento.

Técnicas relacionadas ao modelo preditivo

- Regressão linear simples;
- Ajuste com mínimos quadrados;
- Gradiente descendente;
- Regressão linear múltipla;
- Regressão logística.



Fonte: Livro-texto.

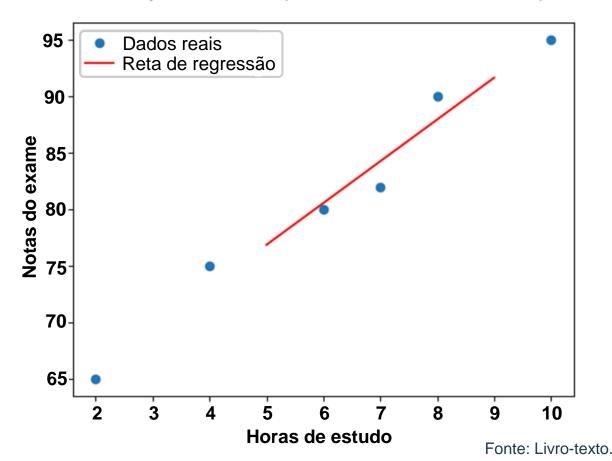
Regressão linear simples

O que é

Técnica estatística que busca estabelecer uma relação linear entre duas variáveis:
 variável dependente (alvo de previsão) e variável independente (variável de entrada).

Objetivo

 Encontrar a melhor reta que representa essa relação entre as variáveis e prever outros valores com base nesta reta.



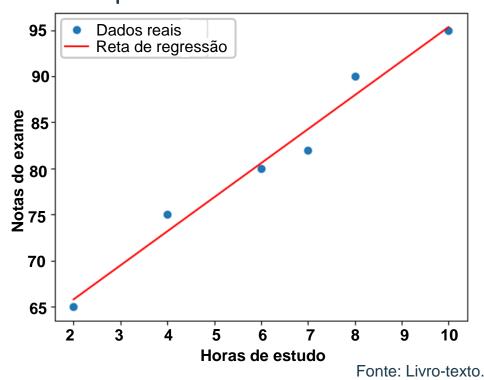
Ajuste com mínimos quadrados

O que é

- Técnica estatística utilizada para encontrar os parâmetros de um modelo matemático que melhor se ajustam a um conjunto de dados observados.
- Essa abordagem busca minimizar a soma dos quadrados dos resíduos, que são as diferenças entre os valores observados e os valores previstos pelo modelo.

Onde é usado

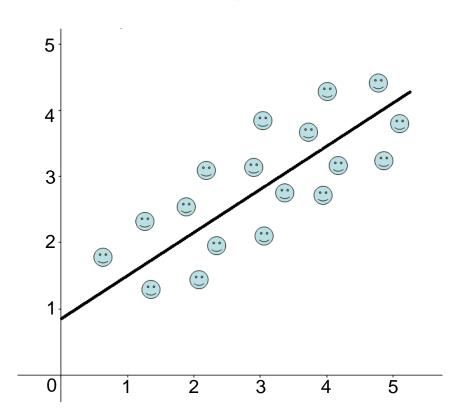
- Econometria, estatística, ciências matemáticas e nas ciências de dados.
 - Regressões cujo objetivo é encontrar funções entre variáveis dependentes e independentes.



Gradiente descendente

O que é

- Algoritmo de otimização utilizado para encontrar o mínimo de uma função de perda, ou seja, para minimizar a diferença entre os valores previstos e os valores reais.
- O gradiente indica a direção e a magnitude do maior aumento da função de perda.
- Ao caminhar na direção oposta ao gradiente, é possível aproximar-se do mínimo global ou local da função.



Fonte: Autoria própria.

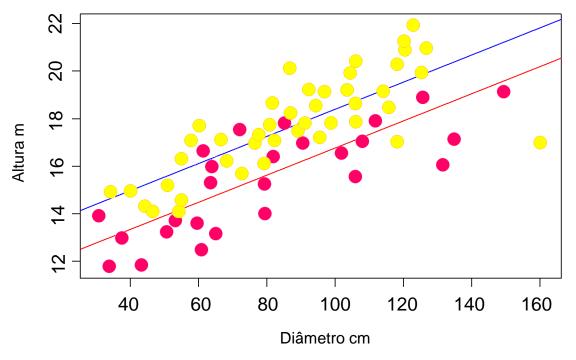
Regressão linear múltipla

O que é

- É uma técnica de aprendizado de máquina que visa modelar a relação entre uma variável dependente contínua e várias variáveis independentes.
- É uma extensão da regressão linear simples.

Objetivo

 Encontrar uma equação que represente a relação linear entre as variáveis independentes e a variável dependente.



Fonte: Autoria própria.

Regressão logística

O que é

- É um modelo estatístico utilizado para realizar análise de classificação binária, ou seja, para prever a probabilidade de um evento pertencer a uma das duas categorias possíveis.
- O modelo de regressão logística, na verdade, é um modelo de classificação.

Objetivo

- Estimar a probabilidade de um evento ocorrer com base em um conjunto de variáveis independentes.
 - Neste modelo, a variável dependente assume sempre um valor binário (0 ou 1, ou Verdadeiro ou Falso, ou outro).

Como funciona

- Utiliza a função logística (sigmoide) para mapear a saída linear da combinação linear das variáveis independentes em um intervalo entre 0 e 1.
- Essa transformação permite interpretar a saída como uma probabilidade.

Interatividade

Qual dos modelos abaixo é o recomendado quando se pretende minimizar a diferença entre os valores previstos e os valores reais?

- a) Regressão linear simples.
- b) Regressão linear múltipla.
- c) Regressão logística.
- d) Regressão descendente.
- e) Ajuste com mínimos quadrados.

Resposta

Qual dos modelos abaixo é o recomendado quando se pretende minimizar a diferença entre os valores previstos e os valores reais?

- a) Regressão linear simples.
- b) Regressão linear múltipla.
- c) Regressão logística.
- d) Regressão descendente.
- e) Ajuste com mínimos quadrados.

Planejamento de experimentos (design of experiments, DoE)

O que é

- É uma abordagem estatística para investigar e otimizar processos, sistemas ou produtos por meio de experimentos controlados.
- Envolve a seleção cuidadosa das variáveis de interesse, a definição das configurações ou níveis dessas variáveis e a realização de experimentos controlados para coletar os dados necessários.

Objetivo

 Identificar fatores que afetam a resposta desejada e determinar as melhores configurações desses fatores para atingir os resultados desejados.

Planejamento de experimentos (design of experiments, DoE)

Exemplos de tipos de designs de experimentos

Design fatorial

É um tipo de design em que todas as combinações possíveis dos níveis dos fatores são testadas. Isso permite investigar o efeito de cada fator individualmente, bem como as interações entre os fatores.

Design fracionado

É um design que permite testar apenas uma fração dos possíveis tratamentos. Isso é útil quando o número de combinações possíveis é muito grande e não é possível realizar todos os experimentos.

Design de blocos

É um design em que os experimentos são divididos em blocos ou grupos. Dentro de cada bloco, os tratamentos são atribuídos de forma aleatória.

Design central composto

É um design que combina um design fatorial completo com pontos adicionais no centro do espaço de trabalho.

Planejamento de experimentos

Split de dados – Treino, teste e validação

O que é

 Técnica usada em aprendizado de máquina para dividir o conjunto de dados em três partes distintas: conjunto de treinamento, conjunto de teste e conjunto de validação.

Objetivo

 Avaliar e validar o desempenho do modelo de aprendizado de máquina de forma adequada.

Planejamento e experimentos

Split de dados – Treino, teste e validação



Conjunto de treinamento

Conjunto de dados usado para treinar o modelo de aprendizado de máquina.
O modelo é ajustado com base nessas amostras de treinamento, aprendendo os padrões e as relações entre as variáveis.



Conjunto de teste

O conjunto de teste é usado para medir o desempenho do modelo em dados independentes, ou seja, dados que o modelo não encontrou durante o treinamento.



Conjunto de validação

O conjunto de validação é usado para avaliar diferentes configurações do modelo e selecionar aquela que apresenta o melhor desempenho em termos de métricas específicas.

Split de dados – Treino, teste e validação

Tipos de split de dados

Split de dados tradicional

Cross-validation (Validação cruzada)

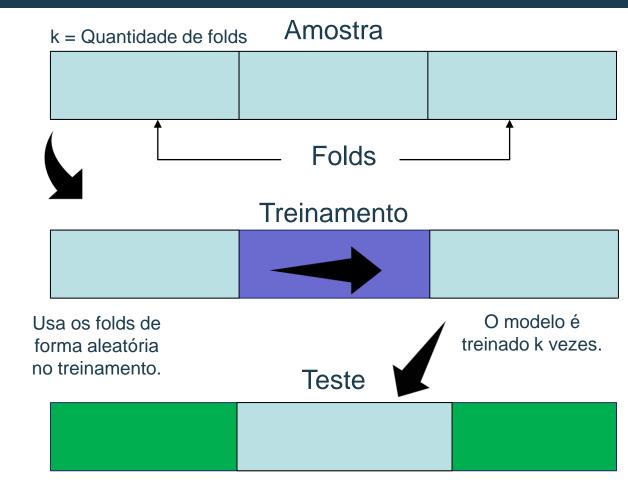
Leave-one-out (LOO)

Time-series split (Divisão de séries temporais) Stratified sampling (Amostragem estratificada)

Validação cruzada (cross-validation)

O que é

- Técnica utilizada na avaliação de modelos de aprendizado de máquina.
- Muito útil quando o conjunto de dados disponível é limitado e é necessário obter uma estimativa mais confiável do desempenho do modelo.



Os folds são usados também nos testes de forma aleatória e k vezes.

 Todos os dados são utilizados tanto para treinamento quanto para teste em algum momento.

Validação cruzada (cross-validation)

Métricas de desempenho de validação cruzada:

Acurácia

Mede a proporção de predições corretas em relação ao total de predições.

Precisão

Mede a proporção de exemplos de verdadeiros positivos em relação a todos os classificados como positivos (verdadeiros positivos mais falsos positivos).

Recall

É a proporção de exemplos verdadeiros positivos em relação a todos os exemplos que realmente pertencem à classe positiva (verdadeiros positivos mais falsos negativos).

F1-Score

É a média harmônica entre essas duas métricas e é especialmente útil quando você deseja encontrar um equilíbrio entre precisão e recall.

Benchmarking

O que é

- Prática utilizada para comparar o desempenho de um sistema, processo ou modelo em relação a outros similares.
- No contexto da ciência de dados, é frequentemente aplicado para avaliar o desempenho de algoritmos de aprendizado de máquina ou de técnicas de processamento de dados em tarefas específicas.

Objetivo

 Fornecer uma referência ou ponto de comparação para medir o desempenho relativo de diferentes abordagens.

Benchmarking

Etapas do processo de benchmarking:

- Definição do problema;
- Seleção de benchmarks;
- Escolha de técnicas de referência;
- Execução dos experimentos;
- Análise dos resultados.



Fonte: https://br.freepik.com/vetores-gratis/analise-dopainel-avaliacao-do-desempenho-do-computador-graficona-tela-analise-estatistica-avaliacao-de-infografico-relatoriocomercial-em-exibicao-ilustracao-isolada-da-metafora-doconceito_10783211.htm#query=Benchmarking%20software &position=17&from_view=search&track=ais

Interatividade

Indique qual das alternativas abaixo melhor completa a lacuna.

A (O) ______ é a métrica mais simples e direta. No entanto, em situações de desequilíbrio de classes, esta métrica pode ser enganosa, pois o modelo pode ficar viciado para prever a classe majoritária.

- a) Precisão.
- b) Aproximação.
- c) Acurácia.
- d) F1-score.
- e) Recall.

Resposta

Indique qual das alternativas abaixo melhor completa a lacuna.

A (O) ______ é a métrica mais simples e direta. No entanto, em situações de desequilíbrio de classes, esta métrica pode ser enganosa, pois o modelo pode ficar viciado para prever a classe majoritária.

- a) Precisão.
- b) Aproximação.
- c) Acurácia.
- d) F1-score.
- e) Recall.

Aplicações avançadas de aprendizado de máquina

Aprendizado profundo (deep learning)

• É uma área avançada de aprendizado de máquina que utiliza redes neurais artificiais profundas para aprender representações complexas e realizar tarefas sofisticadas, como processamento de imagens, reconhecimento de fala e tradução automática.

Aprendizado por reforço

- Técnica que permite que um agente aprenda a tomar decisões em um ambiente dinâmico, recebendo feedback em forma de recompensas.
- É aplicado em problemas de tomada de decisão, como jogos, controle de robôs e otimização de recursos.

Aplicações avançadas de aprendizado de máquina

Aprendizado semissupervisionado

- Abordagem combina o aprendizado supervisionado e não supervisionado, permitindo o uso de uma pequena quantidade de dados rotulados e uma grande quantidade de dados não rotulados.
- É útil quando rotular dados for caro ou demorado, mas há disponibilidade de muitos dados não rotulados.

Aprendizado por transferência (transfer learning)

- O aprendizado por transferência envolve o aproveitamento de conhecimentos adquiridos em uma tarefa para auxiliar o aprendizado em outra tarefa relacionada.
- É útil quando há escassez de dados na tarefa atual, pois permite que o modelo aproveite informações de tarefas anteriores.

Aplicações avançadas de aprendizado de máquina

Aprendizado on-line

- O aprendizado on-line refere-se à capacidade de aprender e atualizar modelos continuamente à medida que novos dados são recebidos em tempo real.
- É aplicado em cenários onde os dados estão em constante mudança e é necessário adaptar-se rapidamente a novas informações.

Visão computacional

O que é

- É uma área da ciência da computação que se dedica ao desenvolvimento de algoritmos e técnicas para permitir que computadores "enxerguem" e compreendam imagens e vídeos.
- Ela busca replicar a capacidade de percepção visual humana por meio da análise e interpretação de dados visuais.

Objetivo

 Envolve o processamento de imagens e vídeos para extrair informações relevantes e realizar tarefas como reconhecimento de objetos, detecção de padrões, segmentação de imagens, reconhecimento facial, entre outras.

Onde está sendo usada?

 Sistemas de vigilância, veículos autônomos, detecção médica, reconhecimento de caracteres em documentos, realidade aumentada e muito mais.

Visão computacional

Conceitos e técnicas utilizados na visão computacional

- Pré-processamento de imagens;
- Extração de características;
- Segmentação de imagens;
- Reconhecimento de padrões;
- Aprendizado de máquina em visão computacional.



Fonte: https://br.freepik.com/fotos-gratis/transformacaodigital-da-biometria-com-midia-remixada-de-microchipfuturista_16016833.htm#query=vis%C3%A3o%20compu tacional&position=3&from_view=search&track=ais

Processamento de linguagem natural (PLN)

O que é

• É uma área da inteligência artificial que se dedica ao estudo e desenvolvimento de algoritmos e técnicas para permitir que computadores compreendam, analisem e gerem linguagem humana de forma natural.

Objetivo

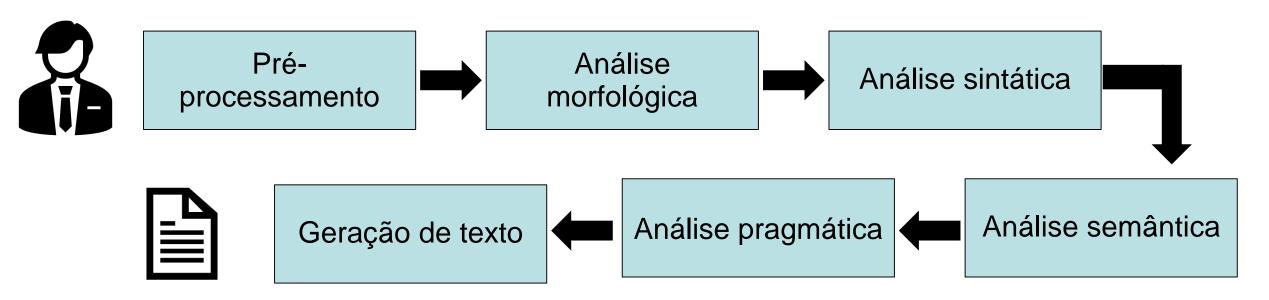
 O PLN lida com a ambiguidade, a variação linguística, a semântica e a estrutura da linguagem humana.

Onde usar?

 Reconhecimento de fala, compreensão de texto, tradução automática, sumarização de textos, geração de texto, resposta a perguntas, análise de sentimentos, entre outras.

Processamento de linguagem natural (PLN)

Etapas do PLN



Reconhecimento de fala – ASR (Automatic speech recognition)

O que é

 E uma área da tecnologia de processamento de linguagem natural que se concentra em converter a fala humana em texto escrito de forma automatizada.

Objetivo

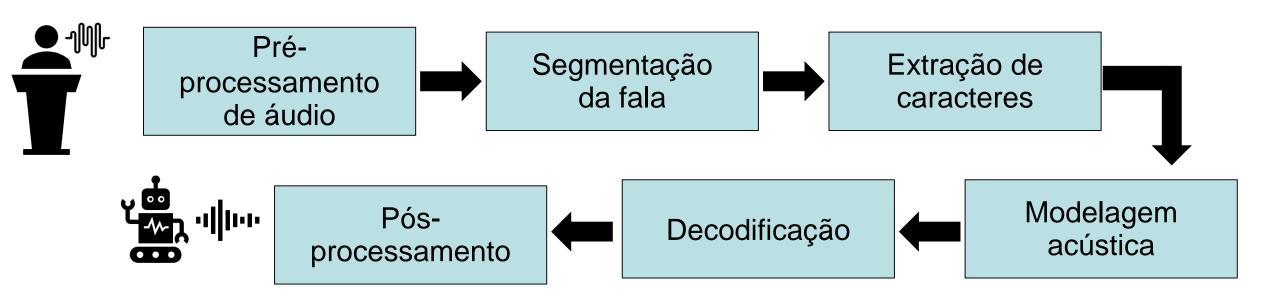
 Permitir que os computadores entendam e processem a fala humana, facilitando a interação homem-máquina.

Onde usar?

 Assistentes virtuais, sistemas de atendimento ao cliente, sistemas de transcrição automática de áudio, tradução de voz em tempo real, legendagem automática em vídeos, entre outros.

Reconhecimento de fala – ASR (Automatic speech recognition)

Etapas do PLN



APIs de Inteligência Artificial

O que são

 São conjuntos de ferramentas, bibliotecas e recursos que dão acesso a recursos de IA em aplicações e sistemas de software.

Objetivo

 Oferecer de forma conveniente recursos de IA avançados, sem que os desenvolvedores precisem se preocupar em desenvolver essas funcionalidades do zero.

Funcionalidades comuns às API de IA

 Reconhecimento de imagem, processamento de linguagem natural, reconhecimento de fala, detecção de sentimentos, entre outras.

Onde usar

 Chatbots, assistentes virtuais, sistemas de análise de dados, sistemas de automação de tarefas, de tomada de decisões, entre outros.

Interatividade

Considerando o que foi abordado sobre as tecnologias de aplicações avançadas relacionadas ao aprendizado de máquina, qual das alternativas abaixo <u>não</u> indica uma tecnologia citada em nosso material de estudos?

- a) Reconhecimento biométrico.
- b) Visão computacional.
- c) Processamento de linguagem natural.
- d) Reconhecimento de fala.
- e) APIs de IA.

Resposta

Considerando o que foi abordado sobre as tecnologias de aplicações avançadas relacionadas ao aprendizado de máquina, qual das alternativas abaixo <u>não</u> indica uma tecnologia citada em nosso material de estudos?

- a) Reconhecimento biométrico.
- b) Visão computacional.
- c) Processamento de linguagem natural.
- d) Reconhecimento de fala.
- e) APIs de IA.

Referências

- ALPAYDIN, E. Introduction to machine learning. Cambridge: MIT Press, 2004.
- CASTRO, L. N.; FERRARI, D. G. Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.
- FACELI, K.; LORENA, A. C.; GAMA, J.; DE CARVALHO, A. C. P. L. F. *Inteligência Artificial*: uma abordagem de aprendizado de máquina. 1. ed. Rio de Janeiro: LTC, 2011.
- MITCHELL, T. M. Machine learning. New York: McGraw-Hill, 1997.
- NORVIG, P. Inteligência Artificial. São Paulo: Grupo GEN, 2013.
- TAN, P. N.; STEINBACH, M.; KUMAR, T. *Introduction to data mining*. Redwood City: Addison Wesley, 2005.
 - ZIKOPOULOS, P.; Eaton, C. *Understanding Big Data*: analytics for enterprise class hadoop and streaming data. New York: McGraw-Hill / Osborne Media, 2011.

ATÉ A PRÓXIMA!