IPOG

Fundamentos de Data Science

EDITORA IPOG

IP5p Instituto de Pós-Graduação e Graduação – IPOG

Fundamentos de Data Science 2. / Autor: Joelma de Moura Ferreira.

240f. :il.

ISBN:

1. Ciência de Dados 2. Inteligência Artificial 3. Aprendizado de Máquina 4. Análise de Dados 5. Python.

CDU: 005

Bibliotecário responsável: xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

IPOG

Instituto de Pós Graduação e Graduação

Sede

Av. T-1 esquina com Av. T-55 N. 2.390 - Setor Bueno - Goiânia-GO. Telefone (0xx62) 3945-5050

http://www.ipog.edu.br

SUMÁRIO

APRESENTAÇÃO

Seja bem-vindo ou bem-vinda à disciplina de Fundamentos de Data Science.

Nesta jornada de aprendizado, vamos explorar os pilares que formam o mundo dinâmico da Ciência de Dados. Desde entender a distinção entre Ciência de Dados, Análise de Dados e Engenharia de Dados até sobre as aplicações práticas dessa disciplina em diversos setores, iremos construir uma base sólida de conhecimento nessa área vital da tecnologia da informação.

Na primeira unidade, começaremos com uma apresentação da disciplina e sua importância na era da informação, compreendendo como a Ciência de Dados se tornou um componente essencial da vida moderna. Em seguida, exploraremos a relação entre inteligência artificial e ciência de dados, reconhecendo como essas áreas se influenciam mutuamente .

Na segunda unidade, adentraremos no mundo das tecnologias e ferramentas em Ciência de Dados, explorando linguagens como Python e R, além de frameworks como Scikit-learn e PyTorch, fundamentais para análise e modelagem de dados. Conheceremos também plataformas e ambientes de desenvolvimento e ferramentas de análise de dados.

Na terceira unidade, mergulharemos na exploração e análise de dados, entendendo o processo de análise exploratória, métodos de visualização de dados e técnicas de pré-processamento de dados, fundamentais para extrair insights valiosos de conjuntos de dados complexos.

Na quarta unidade, focaremos na modelagem e aprendizado de máquina, compreendendo conceitos básicos, tipos de algoritmos e desafios enfrentados nesse campo em constante evolução.

Por fim, na quinta unidade, exploraremos tópicos avançados e tendências em Ciência de Dados, como Redes Neurais e Deep Learning, Visão Computacional, Processamento de Linguagem Natural (NLP) e Modelos Generativos, que estão moldando o futuro dessa área empolgante.

Estou confiante de que esta jornada será enriquecedora e desafiadora, e estou animada para vê-los explorar cada tópico com curiosidade e entusiasmo. Esteja preparado para expandir seus horizontes e criar uma base sólida para o seu crescimento como profissional de Ciência de Dados.

Boa leitura e estudos!

Profa. Dra. Joelma de Moura Ferreira

OBJETIVOS

OBJETIVO GERAL

Ao final da disciplina os alunos compreenderão os fundamentos, técnicas e aplicações da Ciência de Dados, incluindo a importância dos processos de coleta, processamento, análise e interpretação de dados. Eles conhecerão as principais tecnologias, ferramentas e linguagens empregadas na área, bem como a compreenderão o papel ético e responsável na manipulação de dados.

OBJETIVOS ESPECÍFICOS

- Entender as diferenças entre as diversas profissões relacionadas a ciência de dados
- Conhecer as principais ferramentas utilizadas na análise de dados
- Compreender as etapas de exploração de dados, transformação de dados e criação de modelos de dados
- Familiarizar as tendências da inteligência artificial

Conheça como esse conteúdo foi organizado!

Unidade 1: Fundamentos da Ciência de Dados

Unidade 2: Tecnologias e Ferramentas em Ciência de Dados

Unidade 3: Exploração e Análise de Dados

Unidade 4: Modelagem e Aprendizado de Máquina

Unidade 5: Tópicos Avançados e Tendências em Ciência de Dados

UNIDADE 1 FUNDAMENTOS DA CIÊNCIA DE DADOS

OBJETIVOS DA UNIDADE 1

Ao final dos estudos, você deverá ser capaz de: Entender a importância da Ciência de Dados na era da informação e sua aplicabilidade em diversos contextos. Diferenciar os conceitos de Ciência de Dados, Análise de Dados e Engenharia de Dados. E refletir sobre as questões éticas e a responsabilidade associadas à manipulação de dados, desenvolvendo uma consciência crítica sobre seu uso.

1.1 A importância da Ciência de Dados na era da informação

Videoaula do tópico disponível no AVA:

Videoaula 1: A importância da Ciência de Informação na era da informação.

Com a explosão de dados na era digital advindo de fontes diversas, como transações comercias, redes sociais, **dispositivos IoT**, etc. surgiu uma nova área chamada Ciência de Dados.

IoT, ou Internet das Coisas (em inglês, Internet of Things), refere-se a um conceito onde objetos físicos cotidianos estão conectados à internet e podem coletar e trocar dados entre si e com outros sistemas. Dispositivos IoT são os componentes físicos dessa rede interconectada.

A Ciência de Dados é um campo que combina habilidades de programação, estatística, e conhecimento de domínio para analisar conjuntos de dados complexos e tomar decisões.

Em um mundo onde os dados são abundantes, extrair significado e insights valiosos tornou-se crucial para empresas e organizações em todos os setores.

Por exemplo, empresas de comércio eletrônico usam análises de dados para entender o comportamento do cliente e otimizar suas estratégias de marketing, enquanto instituições financeiras empregam modelos de previsão para mitigar riscos e maximizar retornos.

Aprenda Mais: O trabalho será transformado pela ciência de dados. Disponível em: https://www.youtube.com/watch?v=E9vqtzILM9U

1.2 Ciência de Dados x Análise de Dados x Engenharia de Dados

Videoaula do tópico disponível no AVA:

Videoaula 2: Ciência de Dados x Análise de Dados x Engenharia de Dados.

Embora frequentemente interligadas, Ciência de Dados, Análise de Dados e Engenharia de Dados têm papéis distintos no ciclo de vida dos dados.

A Ciência de Dados concentra-se na descoberta de padrões, insights e na construção de modelos **preditivos ou prescritivos** a partir dos dados. Um cientista de dados geralmente lida com problemas mais complexos e exploratórios, usando uma combinação de habilidades em programação e

estatística. Eles estão envolvidos em todas as etapas do ciclo de vida dos dados, desde a coleta e limpeza até a análise e interpretação, muitas vezes utilizando técnicas avançadas de machine learning e inteligência artificial para resolver problemas.

A análise descritiva envolve a descrição e resumo de dados de forma clara e concisa, geralmente por meio de estatísticas básicas e técnicas de visualização de dados. Seu principal objetivo é descrever o que aconteceu (passado) nos dados, sem tentar extrair conclusões além dessa descrição. Exemplo: identificar quais produtos mais vendem.

A análise diagnóstica refere-se ao processo de investigar e compreender os dados na busca identificar e compreender as causas ou origens de um determinado problema, fenômeno ou situação. É uma etapa fundamental que visa identificar padrões, tendências, anomalias e características dos dados que podem fornecer insights sobre o problema em questão. Exemplo: investigar causas de flutuações de vendas.

A análise preditiva envolve a aplicação de técnicas estatísticas e algoritmos de aprendizado de máquina para fazer previsões sobre eventos futuros com base em dados históricos. Em vez de apenas descrever o que aconteceu no passado, a análise preditiva busca entender relações entre variáveis e utilizar essas relações para prever resultados futuros. Exemplo: antecipar quantidade futura de vendas de determinado produto.

A análise prescritiva vai além da descrição do passado e da previsão do futuro, ela busca prescrever recomendações ou ações específicas para otimizar os resultados. Em vez de apenas prever o que acontecerá, a análise prescritiva procura responder à pergunta "o que devemos fazer sobre isso?". Isso geralmente envolve a utilização de técnicas avançadas de otimização e simulação para identificar a melhor estratégia a ser adotada em uma determinada situação. Exemplo: recomendar estratégia de preços.

A Análise de Dados, por outro lado, se concentra na interpretação e comunicação desses insights para apoiar a tomada de decisões. Um analista de dados tende a se concentrar em tarefas mais específicas e operacionais, como a geração de relatórios regulares, análises **descritivas ou diagnósticas** e investigações ad hoc dos dados. Eles são habilidosos em manipulação e visualização de dados, utilizando ferramentas como SQL, Excel, Tableau e Power BI para apresentar informações de forma clara e acessível. Os analistas de dados geralmente trabalham mais próximos das necessidades imediatas do negócio, respondendo a perguntas específicas e fornecendo suporte analítico para tomada de decisões.

Já a Engenharia de Dados lida com a infraestrutura e processos necessários para coletar, armazenar e gerenciar grandes volumes de dados. Por exemplo, enquanto um cientista de dados desenvolve um modelo de aprendizado de

máquina para prever vendas futuras, um engenheiro de dados pode garantir que os dados necessários estejam disponíveis e acessíveis em um ambiente de armazenamento adequado.

Observação: Tanto cientistas de dados quanto analistas de dados podem realizar diferentes tipos de análise (descritiva, diagnóstica, preditiva ou prescritiva), dependendo de suas habilidades, experiência e do contexto do projeto em que estão trabalhando.

Aprenda Mais: Data Analytics vs Data Science. Disponível em: https://www.youtube.com/watch?v=dcXqhMqhZUo (use as legendas em Português)

1.3 Relação entre inteligência artificial e ciência de dados

A Inteligência Artificial, um campo amplo que refere-se à capacidade das máquinas de realizar tarefas que normalmente requerem inteligência humana. Envolve o desenvolvimento de algoritmos e sistemas que podem simular processos cognitivos, como aprendizado, raciocínio, resolução de problemas e percepção.

O Aprendizado de Máquina é um subcampo da Inteligência Artificial que se concentra no desenvolvimento de algoritmos que permitem aos computadores aprenderem a partir de dados.

Os algoritmos de aprendizado de máquina permitem que os sistemas melhorem automaticamente sua performance em uma determinada tarefa à medida que são expostos a mais dados.

Dentro do Aprendizado de Máquina, as Redes Neurais surgem como modelos computacional inspirados no funcionamento do cérebro humano, composto por neurônios artificiais interconectados. É uma poderosa ferramenta que permite sistemas a aprenderem padrões complexos e realizar tarefas sofisticadas. O Aprendizado Profundo, uma vertente do Aprendizado de Máquina, leva essa ideia adiante, utilizando redes neurais profundas para analisar e compreender dados não estruturados, como imagens e áudio.

Em paralelo, a Ciência de Dados se utiliza dessas técnicas e outras metodologias para extrair insights valiosos de grandes volumes de dos dados. Por meio da análise estatística e da aplicação de algoritmos de Aprendizado de Máquina e Aprendizado Profundo, os cientistas de dados podem desvendar

padrões complexos, gerando conhecimentos acionáveis para uma variedade de aplicações.

Fonte: https://www.serpro.gov.br/menu/noticias/noticias-2019/democratizando-a-inteligencia-artificial

1.4 Aplicações da Ciência de Dados em diversos setores

Videoaula do tópico disponível no AVA:

Videoaula 3: Aplicações da Ciência de Dados em diversos setores.

A aplicação da Ciência de Dados é vasta e abrange uma ampla gama de setores. Na saúde, por exemplo, análises de dados podem ser usadas para identificar padrões em grandes conjuntos de dados de pacientes e melhorar o diagnóstico e tratamento de doenças.

No setor de varejo, a análise de dados de transações pode ser usada para personalizar recomendações de produtos, segmentar clientes, detectar tendências, personalizar produtos e otimizar a cadeia de suprimentos, resultando em economia de recursos e maior reconhecimento da marca.

No setor financeiro, a ciência de dados permite prever comportamentos de risco, reduzindo taxas de inadimplência e automatizando o gerenciamento de crédito, o que é crucial para políticas de concessão de crédito mais seguras.

Na manufatura, a análise de dados orientada por insights tem melhorado a eficiência, reduzindo custos e tempo de produção, além de otimizar processos de inventário e logística.

A logística também se beneficia da Ciência de Dados, com empresas utilizando algoritmos para traçar rotas mais precisas e eficientes, resultando em redução de custos e aumento da qualidade do serviço.

Até mesmo em áreas como esportes e entretenimento, a Ciência de Dados desempenha um papel significativo na análise de desempenho de atletas e na recomendação de conteúdo personalizado para os usuários.

Esses exemplos destacam como a Ciência de Dados está moldando e impulsionando a inovação em diversos setores, transformando a maneira como as organizações operam e interagem com seus clientes.

Segundo o McKinsey Global Institute, no artigo “Connected world: An evolution in connectivity beyond the 5G Revolution”, o mundo dever consumir até 20 vezes mais dados em 2030 do que consumia em 2020, isso demonstra que a demanda por profissionais da área de dados só tende a crescer.

McKinsey Global Institute**, Connected world: An evolution in connectivity beyond the 5G Revolution, 2020.** Disponível em: https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/connected-world-an-evolution-in-connectivity-beyond-the-5g-revolution

Hoje, parte das empresas já se preocupa em estabelecer um Centro de Excelência focado em dados e análises. A cultura Data Driven é crucial para transformar dados em conhecimento utilizável e para capacitar colaboradores a tomar decisões informadas.

1.5 Ética e responsabilidade na manipulação de dados

A manipulação de dados na era da Ciência de Dados levanta questões éticas e responsabilidades significativas. À medida que mais dados são coletados e analisados, surgem preocupações sobre privacidade, discriminação e viés algorítmico.

A garantia da segurança na coleta de informações é primordial, assim como a obtenção do consentimento dos indivíduos e a transparência no uso dos dados. A chegada da Lei Geral de Proteção de Dados (LGPD) estabeleceu parâmetros para o uso dos dados, exigindo respeito à privacidade e autonomia dos clientes,

enquanto ainda permite o acesso aos dados para estratégias organizacionais, exigindo medidas de segurança contra invasões e vazamentos.

A ética de dados envolve o estudo e a aplicação de princípios para coletar, usar e proteger os dados de forma responsável. Isso é crucial porque os dados podem ser utilizados para diversos fins, desde ajudar na tomada de decisões até serem explorados por criminosos. Portanto, é necessário considerar questões como o uso de tecnologias como o Big Data, garantindo que informações sensíveis não sejam expostas.

A importância da ética de dados está em garantir que as pessoas mantenham o controle sobre suas informações pessoais, reduzindo os riscos cibernéticos e protegendo a privacidade. Além disso, permite aos cidadãos o acesso e a correção de suas informações, enquanto possibilita às empresas a inovação de produtos e serviços que atendam aos interesses dos clientes, evitando responsabilidades legais.

Na atualidade, aspectos como consentimento informado, anonimato, confidencialidade, segurança, privacidade, exatidão, propriedade, honestidade e responsabilidade são fundamentais na ética de dados. É necessário garantir que os consumidores tenham transparência sobre como seus dados são compartilhados e acessados, além de investir em tecnologias e políticas de segurança para proteger as informações.

A regulação da Inteligência Artificial (IA) no Brasil está ganhando forma através de iniciativas legais e regulamentares que visam garantir a conformidade ética e legal no desenvolvimento e uso dessas tecnologias. Um exemplo é o Projeto de Lei (PL 2338/23), que estabelece diretrizes para a regulamentação da IA no país, buscando equilibrar inovação e segurança. Essa lei propõe mecanismos para assegurar que as soluções de IA sejam desenvolvidas e utilizadas de maneira responsável, considerando aspectos como transparência, segurança e direitos fundamentais.

Além disso, a Lei Geral de Proteção de Dados (LGPD) impõe exigências rigorosas sobre o tratamento de dados pessoais, que são muitas vezes o combustível para algoritmos de IA. A conformidade com a LGPD é a asseguração de que as práticas de IA respeitam a privacidade e os direitos dos indivíduos, prevenindo abusos e garantindo a proteção dos dados.

A implementação da ética de dados requer um compromisso compartilhado entre profissionais, organizações e sociedade. Além disso, é essencial compreender as leis e regulamentações, como a LGPD, e adotar práticas que promovam a segurança e a privacidade dos dados. Por fim, a ética de dados deve ser parte

integrante das políticas e procedimentos das empresas, com treinamento adequado e constante avaliação e atualização das medidas de segurança.

Aprenda Mais: De DPO para DPO: como tratar dados de forma ética em data Science. Disponível em: https://www.serpro.gov.br/lgpd/noticias/2019/tratamento-etico-dados-pessoais-ciencia-data-science

Exercício proposto: Para ampliar seu aprendizado e auxiliar a fixação dos conceitos, monte um mapa mental das palavras chaves do artigo de DPO para DPO: como tratar dados de forma ética em Data Science. Você pode utilizar um programa para construção de mapas mentais, por exemplo o https://miro.com/, ou fazer em uma folha de papel.

Tem dúvidas de como estudar com mapas mentais? Assista o vídeo Como fazer um MAPA MENTAL de forma simples e rápida: https://www.youtube.com/watch?v=rl23Ao4ccIE

1.5.1. Viés algorítmico

Um outro ponto de destaque do uso de dados na ciência de dados é o viés algorítmico, que se refere à tendência dos algoritmos de tomarem decisões enviesadas ou discriminatórias com base nos dados com os quais foram treinados. Por exemplo, algoritmos de recrutamento baseados em dados podem inadvertidamente perpetuar preconceitos existentes se os conjuntos de dados utilizados refletirem desigualdades históricas. Isso ocorre porque os algoritmos aprendem com os dados disponíveis, e se esses dados contiverem preconceitos, o algoritmo pode reproduzi-los em suas decisões.

Portanto, os profissionais de Ciência de Dados devem estar cientes desse viés e tomar medidas para mitigá-lo. Isso inclui não apenas garantir a precisão e eficácia dos modelos, mas também considerar os impactos sociais e éticos de suas decisões. Adotar práticas transparentes, garantir a proteção da privacidade dos dados e promover a equidade são elementos essenciais para o uso responsável da Ciência de Dados.

Ao lidar com o viés algorítmico, os profissionais de Ciência de Dados devem examinar criticamente os conjuntos de dados utilizados em seus modelos,

identificar e corrigir quaisquer tendências discriminatórias e buscar maneiras de aumentar a diversidade e representatividade dos dados. Além disso, é importante implementar medidas de transparência, permitindo que os usuários entendam como as decisões são tomadas e tenham a oportunidade de contestá-las ou solicitar revisões.

A equidade deve ser uma consideração central em todas as etapas do processo de ciência de dados, desde a coleta e preparação dos dados até o desenvolvimento e implementação dos modelos. Isso envolve não apenas a eliminação de preconceitos existentes, mas também a criação de sistemas que promovam a igualdade de oportunidades e tratamento justo para todos os envolvidos.

Aprenda Mais: Discriminação algorítmica. Disponível em: https://pt.wikipedia.org/wiki/Discrimina%C3%A7%C3%A3o_algor%C3%ADtmica

Exercício proposto:

1) Qual é o principal objetivo da Ciência de Dados na era da informação?

a) Desenvolver aplicativos móveis

b) Analisar conjuntos de dados complexos e extrair insights valiosos

c) Criar redes sociais

d) Implementar sistemas de segurança cibernética

2) Qual é a principal responsabilidade da Engenharia de Dados no ciclo de vida dos dados?

a) Construir modelos preditivos

b) Interpretar insights e comunicar resultados

c) Desenvolver aplicativos de visualização de dados

d) Coletar, armazenar e gerenciar grandes volumes de dados

3) O que significa viés algorítmico na Ciência de Dados?

a) Tendência dos algoritmos de tomarem decisões baseadas em dados históricos

b) Capacidade dos algoritmos de aprender com grandes conjuntos de dados

c) Inclusão de preconceitos nos algoritmos devido aos dados utilizados

d) Adoção de práticas transparentes na análise de dados

4) Uma empresa de tecnologia está planejando uma nova estratégia para lidar com seus dados e melhorar suas operações. Para isso, está considerando a contratação de profissionais especializados em Ciência de Dados, Análise de Dados e Engenharia de Dados. Abaixo estão breves descrições sobre cada uma dessas áreas:

I) Ciência de Dados: Esta área se concentra na descoberta de padrões, insights e na construção de modelos preditivos ou prescritivos a partir dos dados. Os profissionais de Ciência de Dados geralmente lidam com problemas mais complexos e exploratórios, utilizando uma combinação de habilidades em programação e estatística.

II) Análise de Dados: Por outro lado, a Análise de Dados foca na interpretação e comunicação desses insights para apoiar a tomada de decisões. Os analistas de dados tendem a se concentrar em tarefas mais específicas e operacionais, como a geração de relatórios regulares e análises descritivas dos dados.

III) Engenharia de Dados: Já a Engenharia de Dados lida com a infraestrutura e processos necessários para coletar, armazenar e gerenciar grandes volumes de dados. Os engenheiros de dados garantem que os dados estejam disponíveis e acessíveis em um ambiente de armazenamento adequado, além de cuidar da integridade e segurança desses dados.

Considerando as descrições fornecidas sobre Ciência de Dados, Análise de Dados e Engenharia de Dados, qual dessas áreas seria mais adequada para lidar com a análise de padrões de comportamento dos clientes e a construção de modelos de previsão de vendas para a empresa de tecnologia?

a) Ciência de Dado

b) Análise de Dados

c) Engenharia de Dados

d) Todas as áreas mencionadas

Resposta correta 1: b) Analisar conjuntos de dados complexos e extrair insights valiosos

Resposta correta 2: d) Coletar, armazenar e gerenciar grandes volumes de dados

Resposta correta 3: c) Inclusão de preconceitos nos algoritmos devido aos dados utilizados

Resposta correta 4: a) Ciência de Dados

UNIDADE 2 TECNOLOGIAS E FERRAMENTAS EM CIÊNCIA DE DADOS

OBJETIVOS DA UNIDADE 2

Ao final dos estudos, você deverá ser capaz de: Identificar principais tecnologias e linguagens empregadas em Ciência de Dados, como Python, Linguagem R, Tableau, PowerBI, e outros frameworks, compreendendo suas capacidades e aplicações práticas na interpretação de conjuntos de dados complexos.

2.1 Plataformas e ambientes de desenvolvimento em Ciência de Dados

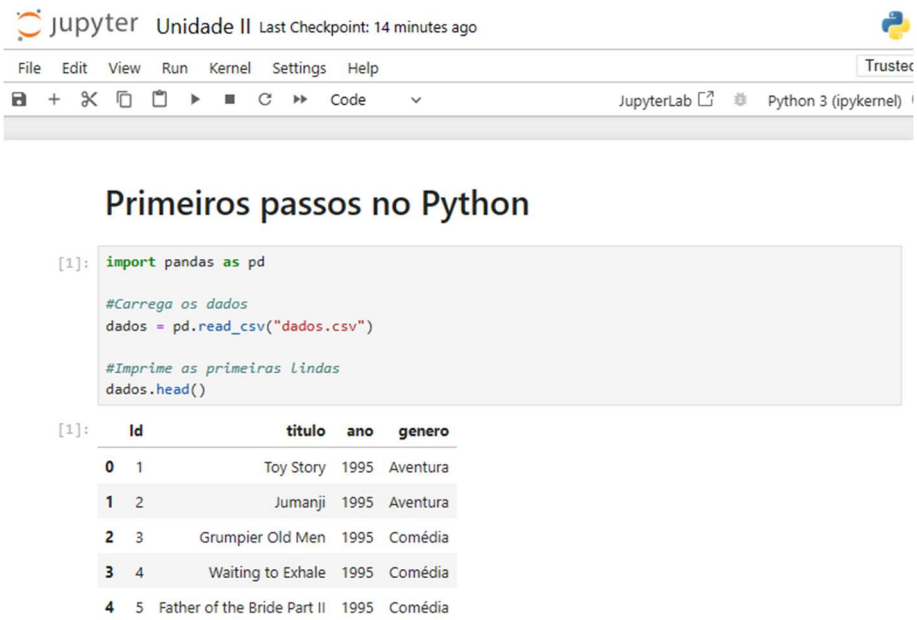
Videoaula do tópico disponível no AVA:

Videoaula 5: Plataformas e ambientes de desenvolvimento em Ciência de Dados.

2.1.1 Jupyter Notebooks e Google Colab

Jupyter Notebooks e Google Colab são ambientes de desenvolvimento interativos que permitem escrever e executar código em blocos, facilitando a exploração de dados e a criação de modelos.

O Jupyter Notebook é uma ferramenta de computação interativa que tem ganhado popularidade na comunidade de ciência de dados e programação em geral. Ele oferece um ambiente flexível onde os usuários podem escrever e executar código em várias linguagens de programação, incluindo Python, R, Julia e muitas outras. Uma das características distintivas do Jupyter Notebook é sua capacidade de integrar código, texto formatado e visualizações em um único documento, facilitando a criação de relatórios, demonstrações e análises interativas.

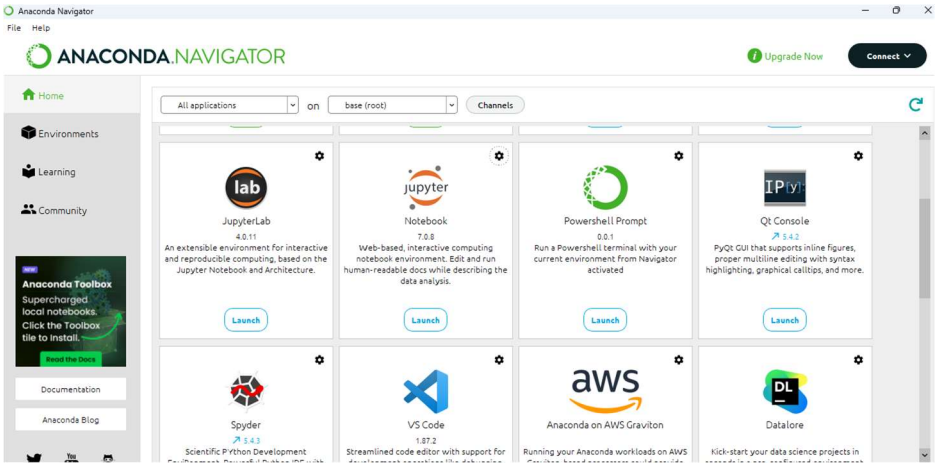


	Id	titulo	ano	genero
0	1	Toy Story	1995	Aventura
1	2	Jumanji	1995	Aventura
2	3	Grumpier Old Men	1995	Comédia
3	4	Waiting to Exhale	1995	Comédia
4	5	Father of the Bride Part II	1995	Comédia

No Jupyter Notebook, o código é organizado em células que podem ser executadas individualmente. Isso permite uma abordagem iterativa no desenvolvimento de código, onde os usuários podem testar pequenos trechos de código e ver imediatamente os resultados. Além disso, as células de texto formatado, escritas em Markdown, permitem a inclusão de explicações,

equações, imagens e links, tornando o notebook uma ferramenta ideal para documentar o processo de análise de dados e compartilhar resultados.

Para usar o Jupyter Notebook pode ser instalado o Anaconda Navigator que é uma distribuição popular de Python e R, utilizada especialmente em ambientes de ciência de dados, que inclui um conjunto de pacotes pré-instalados, incluindo o Python, bibliotecas científicas como NumPy, pandas, Matplotlib, scikit-learn, entre outras, além de ambientes de desenvolvimento como o Jupyter Notebook e o JupyterLab.



O Google Colab, abreviação de Google Colaboratory, é uma plataforma gratuita baseada na nuvem oferecida pelo Google que permite aos usuários escrever e executar código Python em notebooks interativos. Essa ferramenta foi desenvolvida para facilitar o desenvolvimento e a colaboração em projetos de ciência de dados, machine learning, pesquisa e educação. Uma das principais vantagens do Google Colab é que ele elimina a necessidade de configurar um ambiente de desenvolvimento local, pois todo o processamento é feito nos servidores do Google.

Os notebooks do Google Colab fornecem um ambiente semelhante ao Jupyter Notebook, onde os usuários podem criar e executar células de código, além de adicionar texto formatado, gráficos e visualizações. Além disso, o Colab oferece integração com várias bibliotecas populares de ciência de dados e machine learning, como TensorFlow, PyTorch, scikit-learn e muitas outras. Isso permite que os usuários aproveitem as poderosas capacidades dessas bibliotecas sem a necessidade de configurar seus próprios ambientes de desenvolvimento.

Uma das características mais atraentes do Google Colab é sua capacidade de utilizar GPUs (Graphics Processing Units) e TPUs (*Tensor Processing Units*) gratuitamente. Isso é especialmente útil para tarefas que exigem grande poder computacional, como treinamento de modelos de deep learning em conjuntos de dados volumosos. Além disso, o Colab oferece integração com serviços do

Google, como Google Drive, facilitando o compartilhamento de notebooks e dados entre colaboradores.

Aprenda Mais:

Jupyter Notebook ou Colab | Qual o melhor? Disponível em: https://www.youtube.com/watch?v=AhucYVEMAzw

2.2 Principais tecnologias e linguagens utilizadas em Ciência de Dados

Videoaula do tópico disponível no AVA:

Videoaula 6: Principais tecnologias e linguagens utilizadas em Ciência de Dados.

A Ciência de Dados é um campo interdisciplinar utiliza uma variedade de tecnologias e linguagens para coletar, processar, analisar e interpretar dados. Desde linguagens de programação até ferramentas especializadas em visualização e machine learning, o arsenal de tecnologias disponíveis para os cientistas de dados é vasto e em constante evolução. Podemos destacar algumas:

- No âmbito das linguagens de programação, Python e R são as principais escolhas, destacando-se por sua versatilidade, vasta comunidade de desenvolvedores e extensivas bibliotecas especializadas em análise de dados.
- O aprendizado de máquina (machine learning) é uma faceta essencial da Ciência de Dados, e ferramentas como TensorFlow, scikit-learn e PyTorch oferecem uma gama de algoritmos e técnicas para modelagem preditiva, classificação, regressão e agrupamento de dados.
- Na área de visualização de dados, ferramentas como Matplotlib, Seaborn, Power BI e Tableau são utilizadas para criar visualizações claras e informativas

2.2.1 Python

Python é uma das linguagens mais populares na ciência de dados devido à sua simplicidade, versatilidade e uma grande quantidade de bibliotecas especializadas. Com bibliotecas como NumPy, Pandas, Matplotlib e Scikit-learn,

os cientistas de dados podem realizar desde tarefas básicas de manipulação de dados até algoritmos complexos de aprendizado de máquinas.

Por exemplo, abaixo está um código simples que carrega um conjunto de dados CSV usando Pandas e mostra o dataframe.

O arquivo csv teria a seguinte estrutura:

```
Id,titulo,ano,genero

1,Toy Story,1995,Aventura

2,Jumanji,1995,Aventura

3,Grumpier Old Men,1995,Comédia

4,Waiting to Exhale,1995,Comédia

5,Father of the Bride Part II,1995,Comédia

6,Heat,1995,Ação

7,Sabrina,1995,Comédia

8,Tom and Huck,1995,Aventura

9,Sudden Death,1995,Ação

10,GoldenEye,1995,Ação
```

O código faz a importação do Pandas e a leitura do arquivo, gerando a partir dos dados um Dataframe que é impresso.

```
import pandas as pd
#Carrega os dados
dados = pd.read_csv("dados.csv")

#Imprime as primeiras lindas
dados.head()
```

O resultado da impressão do dataframe é:

	Id	titulo	ano	genero
0	1	Toy Story	1995	Aventura
1	2	Jumanji	1995	Aventura
2	3	Grumpier Old Men	1995	Comédia
3	4	Waiting to Exhale	1995	Comédia
4	5	Father of the Bride Part II	1995	Comédia

O Pandas é uma biblioteca poderosa para manipulação dados cuja principal estrutura de dados é chamada de DataFrame que é fácil de usar e adequada para áreas que dependem muito de dados.

Aprenda Mais:

Análise de Dados com Python e Pandas. Disponível em: https://www.kaggle.com/code/joaoavf/introducao-a-analise-de-dados-python-e-pandas

Pandas User Guide. Disponível em: https://pandas.pydata.org/docs/user_guide/index.html

2.2.2 Linguagem R

R é outra linguagem popular entre os cientistas de dados, especialmente na comunidade acadêmica. Ele oferece uma ampla gama de pacotes estatísticos e ferramentas de visualização que facilitam a análise exploratória de dados e a criação de gráficos informativos. Abaixo está um exemplo de como usar R para criar um gráfico de dispersão simples:

```r
# Carregar o conjunto de dados
dados <- read.csv("dados.csv")

# Criar um gráfico de dispersão
plot(dados$X, dados$Y, main="Gráfico de Dispersão", xlab="Eixo X", ylab="Eixo Y")
```

2.3 Exemplos de Frameworks: Scikit-learn, Matplotlib, Seaborn e PyTorch

Videoaula do tópico disponível no AVA:

Videoaula 7 Exemplos de Frameworks: Scikit-learn, Matplolib, Seaborn e PyTorch.

2.2.3.1 scikit-learn

O scikit-learn, muitas vezes abreviado como sklearn, é uma biblioteca em Python amplamente utilizada para aprendizado de máquina. Ela oferece diversos algoritmos de aprendizado de máquina, bem como ferramentas para pré-processamento de dados, seleção de modelos, validação e avaliação de desempenho. A biblioteca é construída sobre outras bibliotecas populares, como NumPy, SciPy e matplotlib.

Vamos a um exemplo simples. Quando vamos construir um modelo de machine learning, por exemplo, para calcular a previsão de vendas do próximo semestre, é importante antes padronizar os dados. Padronizar os dados é um processo comum no pré-processamento de dados cujo objetivo é colocar as variáveis em uma escala comum, tornando-as comparáveis. Um método comum de padronização é a padronização Z-score, que transforma os dados para que eles tenham uma média de zero e um desvio padrão de um.

O código abaixo é um exemplo simples de como poderíamos usar a biblioteca scikit-learn para nos ajudar nisso. Neste código, estamos usando a classe StandardScaler do módulo sklearn.preprocessing para padronizar os dados = [[3, 56], [12, 45], [34, 45], [56, 8]]

```
from sklearn.preprocessing import StandardScaler

dados = [[3, 56], [12, 45], [34, 45], [56, 8]]
scaler = StandardScaler()

dados_alterados = scaler.fit_transform(dados)
print(dados)
print(dados_alterados)
```

```
[[3, 56], [12, 45], [34, 45], [56, 8]]
[[-1.13154094  0.9629786 ]
 [-0.69352509  0.35767777]
 [ 0.37718031  0.35767777]
 [ 1.44788572 -1.67833413]]
```

Assim, os dados que realmente seriam usados para treinar o modelo de machine learning seriam os dados_alterados = [[-1.13154094, 0.9629786], [-0.69352509, 0.35767777], [0.37718031, 0.35767777], [1.44788572, -1.67833413]]

Padronizar os dados antes de alimentá-los a um modelo de machine learning é importante por várias razões, entre elas, manter a uniformidade de escala, pois algoritmos de machine learning podem ter desempenho inferior quando os dados têm escalas muito diferentes; facilitar a interpretação dos coeficientes, uma vez que, modelos lineares, como regressão linear ou regressão logística, a interpretação dos coeficientes é mais fácil quando os dados estão padronizados; além de, assegurar o desempenho do modelo, em geral, o desempenho de muitos algoritmos de machine learning pode ser melhorado ao padronizar os dados.

A padronização é apenas umas das funcionalidades presentes no scikit-learn, mais informações estão no site oficial: https://scikit-learn.org/stable/index.html. Utilizaremos muito essa biblioteca no curso.

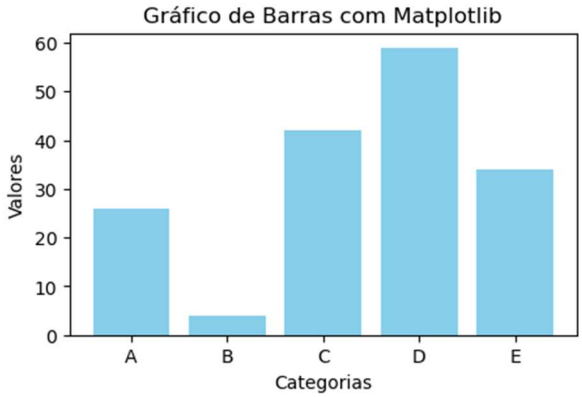
2.2.3.2 Matplotlib e Seaborn

Imagine que você quer desenhar um gráfico ou plotar dados em um papel. Você pegaria um lápis ou caneta e começaria a desenhar pontos, linhas ou barras para representar seus dados. Matplotlib ou o Seaborn são como essa caneta ou lápis, mas para o mundo digital. Matplotlib e Seaborn são bibliotecas de visualização de dados que permitem criar gráficos e plots estilizados.

Matplotlib é uma biblioteca de visualização de dados em Python. Ela permite que você crie gráficos e plots de maneira fácil e flexível. Com o Matplotlib, você pode fazer todos os tipos de visualizações, desde simples gráficos de linha até gráficos 3D complexos. Ele é muito útil para explorar dados, comunicar resultados e entender padrões nos dados.

Abaixo está um exemplo simples de como usar Matplotlib para criar um gráfico de barras a partir de dados gerados randomicamente:

```python
import matplotlib.pyplot as plt
import numpy as np

# Gerar dados aleatórios
categorias = ['A', 'B', 'C', 'D', 'E']
valores = np.random.randint(1, 100, size=len(categorias))

# Criar o gráfico de barras usando Matplotlib
plt.figure(figsize=(5, 3))
plt.bar(categorias, valores, color='skyblue')
plt.xlabel('Categorias')
plt.ylabel('Valores')
plt.title('Gráfico de Barras com Matplotlib')
plt.show()
```

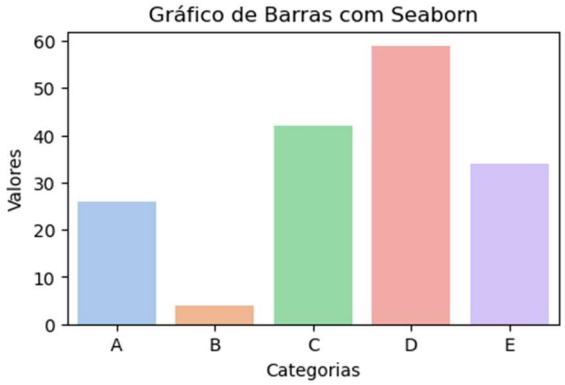


Agora, imagine que você quer tornar esses gráficos e plots ainda mais bonitos e atraentes, mas sem muita complicação. É aí que o Seaborn entra em cena.

Seaborn é uma biblioteca de visualização de dados construída sobre o Matplotlib. Ele oferece uma interface de alto nível para criar gráficos estilizados com apenas algumas linhas de código. O Seaborn vem com uma variedade de temas e paletas de cores predefinidos que tornam fácil criar visualizações atraentes sem precisar se preocupar muito com os detalhes de design. Além disso, o Seaborn também possui funções especializadas para visualizar relacionamentos estatísticos complexos entre variáveis.

Agora para o mesmo conjunto de dados um gráfico usando a biblioteca Seaborn.

```
import seaborn as sns

# Criar o mesmo gráfico de barras usando Seaborn
plt.figure(figsize=(5, 3))
sns.barplot(x=categorias, y=valores, palette='pastel')
plt.xlabel('Categorias')
plt.ylabel('Valores')
plt.title('Gráfico de Barras com Seaborn')
plt.show();
```



Exercício proposto:

Utilize os dados de um conjunto de vendas de uma loja online para criar visualizações utilizando as bibliotecas Matplotlib e Seaborn. O objetivo é entender melhor o desempenho de vendas ao longo do tempo e explorar possíveis tendências.

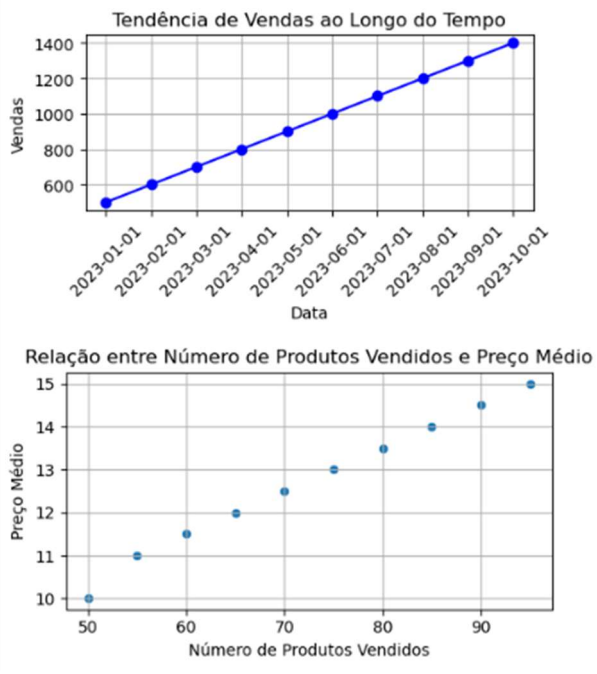
- Monte o CSV sales_data.csv , utilizando algum editor te textos, com os dados abaixo. No Python, carregue os dados de vendas.

```
Data,Vendas,Num_Produtos,Preco_Medio
2023-01-01,500,50,10
2023-02-01,600,55,11
2023-03-01,700,60,11.5
2023-04-01,800,65,12
2023-05-01,900,70,12.5
2023-06-01,1000,75,13
2023-07-01,1100,80,13.5
2023-08-01,1200,85,14
2023-09-01,1300,90,14.5
2023-10-01,1400,95,15
```

- Pesquise como utilizar o Matplotlib para criar um gráfico de linha mostrando a tendência de vendas ao longo do tempo.
- Pesquise como utilizar o Seaborn para criar um gráfico de dispersão mostrando a relação entre o número de produtos vendidos e o preço médio dos produtos.
- Interprete os insights obtidos a partir das visualizações.

Possível resposta:

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# 1. Carregar os dados
sales_data = pd.read_csv('sales_data.csv')

# 2. Gráfico de linha com Matplotlib
plt.figure(figsize=(5, 3))
plt.plot(sales_data['Data'], sales_data['Vendas'], marker='o', color='b')
plt.title('Tendência de Vendas ao Longo do Tempo')
plt.xlabel('Data')
plt.ylabel('Vendas')
plt.xticks(rotation=45)
plt.grid(True)
plt.tight_layout()
plt.show()

# 3. Gráfico de dispersão com Seaborn
plt.figure(figsize=(5, 3))
sns.scatterplot(x='Num_Produtos', y='Preco_Medio', data=sales_data)
plt.title('Relação entre Número de Produtos Vendidos e Preço Médio')
plt.xlabel('Número de Produtos Vendidos')
plt.ylabel('Preço Médio')
plt.grid(True)
plt.tight_layout()
plt.show()
```



4. Interpretação dos insights

O gráfico de linha mostra uma tendência crescente nas vendas ao longo do tempo. O gráfico de dispersão mostra uma correlação positiva entre o número de produtos vendidos e o preço médio dos produtos. Isso sugere que, à medida que o número de produtos vendidos aumenta, o preço médio dos produtos também tende a aumentar.

2.2.3.3 PyTorch

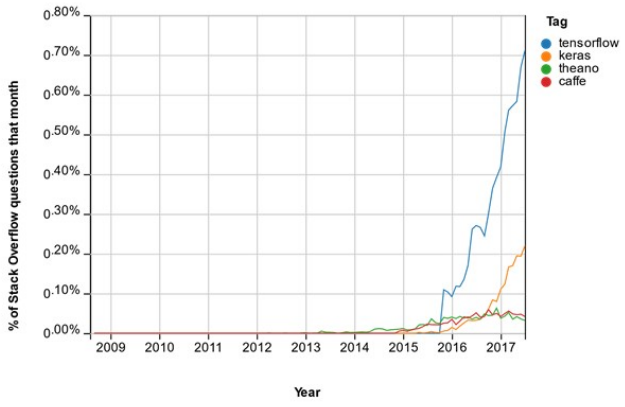
PyTorch é uma biblioteca de aprendizado de máquina de código aberto que facilita a criação e treinamento de modelos de *deep learning*.

Imagine que você está ensinando um computador a realizar uma tarefa específica, como reconhecer gatos em fotos. Você precisa fornecer ao computador muitas fotos de gatos e dizer a ele quais são gatos e quais não são. Mas como o computador realmente aprende a reconhecer os gatos? Aqui é onde entra o PyTorch.

PyTorch é como uma caixa de ferramentas que torna mais fácil ensinar computadores a aprenderem sozinhos. Ele é uma biblioteca de código aberto para aprendizado de máquina, especialmente para deep learning. Deep learning é uma forma avançada de aprendizado de máquina inspirada no funcionamento do cérebro humano. Com o PyTorch, você pode construir e treinar modelos de deep learning. Esses modelos podem aprender a reconhecer padrões complexos nos dados, como rostos em fotos ou padrões em séries temporais. O PyTorch simplifica muito o processo de construção e treinamento desses modelos.

Em resumo, o PyTorch é uma ferramenta poderosa que permite aos pesquisadores e desenvolvedores explorar e experimentar com técnicas de deep learning de uma maneira acessível e eficiente, ajudando a impulsionar avanços significativos em áreas como visão computacional, processamento de linguagem natural e muito mais.

É importante ressaltar que o PyTorch não é a única biblioteca disponível para deep learning. Existem outras bibliotecas como o TensorFlow, Keras ou MXNet. Teremos uma disciplina apenas para ensinar redes neurais.



Fonte: https://www.guru99.com/pt/deep-learning-libraries.html

Aprenda Mais:

Melhores bibliotecas de Machine e Deep Learning. Disponível em: https://itforum.com.br/noticias/melhores-bibliotecas-de-machine-e-deep-learning/

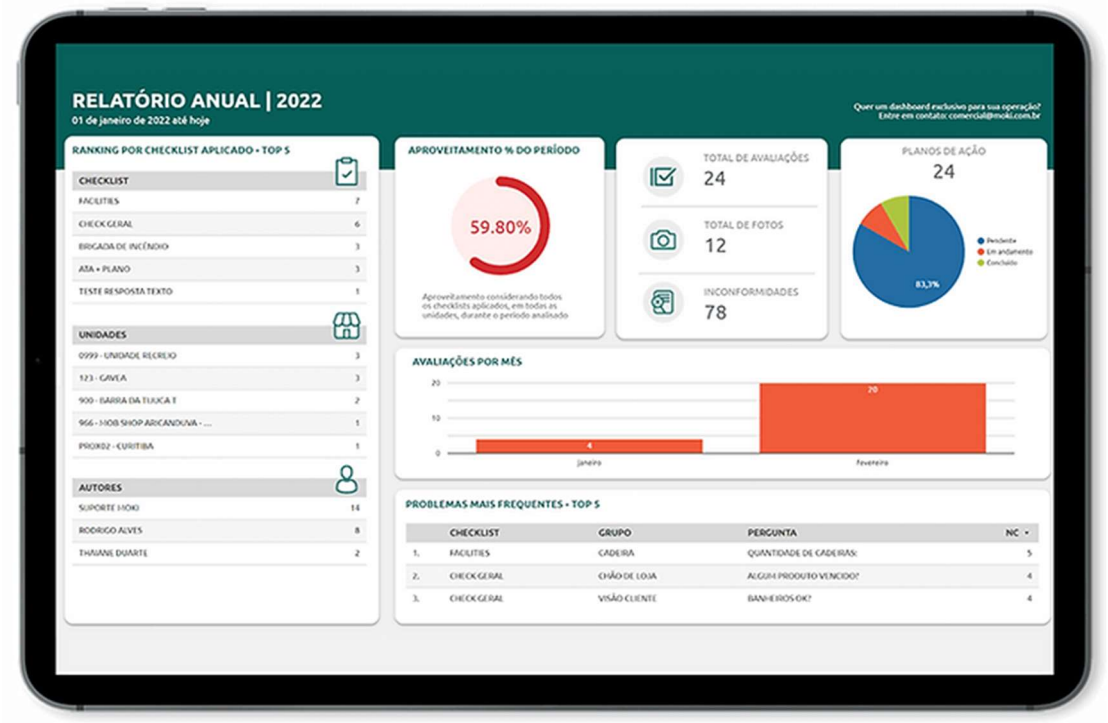
2.4 Ferramentas de Análise de Dados

O sonho de todo gestor é parar de acessar diferentes ferramentas para gerar relatórios, exportar planilhas e compilar dados manualmente. Com a evolução das tecnologias de **Business Intelligence (BI)**, esse sonho se tornou realidade. A implementação de soluções de BI em estruturas de dados complexas permite integrar, analisar e visualizar informações de maneira eficiente, simplificando a gestão.

Um dos principais componentes dessas soluções é o **dashboard**, um painel visual que reúne informações, métricas e indicadores essenciais para o negócio. Com um dashboard bem estruturado, os gestores podem:

- Visualizar **indicadores e métricas** de forma objetiva e clara, sem a necessidade de manipular múltiplas fontes de dados.
- **Embasar a tomada de decisões** com base em dados atualizados e precisos, aumentando a assertividade estratégica.
- **Acompanhar o desempenho** da empresa em tempo real, permitindo ajustes rápidos e informados.
- **Facilitar o monitoramento de dados**, garantindo que todas as áreas da empresa estejam alinhadas com as metas e objetivos estratégicos.

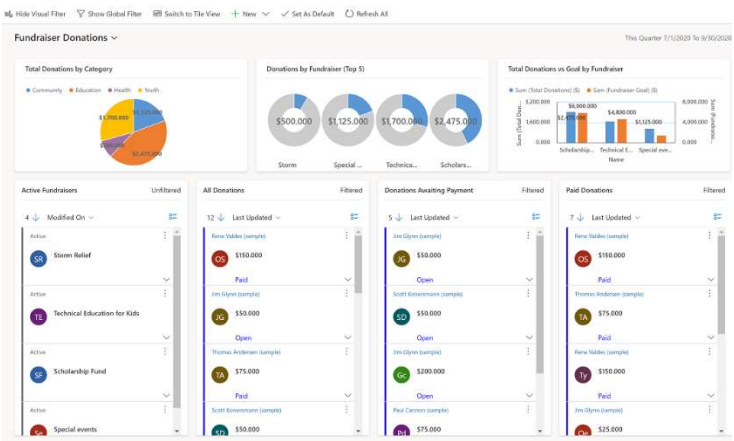
Com essas funcionalidades, o BI se torna um aliado na condução dos negócios, permitindo que os gestores se concentrem na estratégia, ao invés de perder tempo com tarefas operacionais repetitivas.



Fonte: https://www.site.moki.com.br/post/dashboard-o-que-e

2.4.1 Tableau, PowerBI e outros frameworks

Tableau e PowerBI são ferramentas de visualização de dados que permitem criar dashboards interativos e relatórios dinâmicos a partir de diversas fontes de dados. Eles oferecem uma ampla gama de opções de gráficos e filtros para explorar e comunicar insights de forma eficaz.



Fonte: https://learn.microsoft.com/pt-br/power-apps/user/interactive-dashboards

Com o Power BI, os usuários podem criar painéis personalizados e relatórios dinâmicos. A ferramenta oferece várias opções de visualização, como gráficos

de barras, linhas, pizza e mapas, além de recursos avançados como filtros interativos, segmentações e *drill-downs*. Isso permite que os usuários explorem os dados de diferentes perspectivas e identifiquem padrões e tendências importantes.

Outra vantagem do Power BI é sua facilidade de uso e acessibilidade. Com uma interface intuitiva e recursos de arrastar e soltar, os usuários podem criar relatórios e painéis sem a necessidade de habilidades de programação avançadas. Além disso, o Power BI oferece recursos de compartilhamento e colaboração que permitem aos usuários compartilhar relatórios e *insights* com colegas de equipe e partes interessadas, tanto dentro quanto fora da organização.

Aprenda Mais:

Visita Guiada Power BI Disponível em: https://powerbi.microsoft.com/pt-br/guidedtour/power-platform/power-bi/1/1

Tableau é uma plataforma líder em visualização de dados e análise visual, projetada para ajudar indivíduos e organizações a entender e comunicar insights complexos de dados de forma eficaz. Com uma interface intuitiva e poderosas capacidades de visualização, o Tableau permite que os usuários criem painéis interativos, dashboards e relatórios dinâmicos a partir de uma ampla variedade de fontes de dados.

Uma das características distintivas do Tableau é sua abordagem baseada em arrastar e soltar, que simplifica o processo de criação de visualizações complexas sem a necessidade de habilidades de programação avançadas. Os usuários podem escolher entre uma variedade de tipos de gráficos, incluindo barras, linhas, dispersões, mapas e muito mais, e personalizá-los de acordo com suas necessidades específicas. Além disso, o Tableau oferece recursos avançados, como filtros interativos, ações, cálculos personalizados e previsões, que permitem uma análise mais aprofundada dos dados.

UNIDADE 3 EXPLORAÇÂO E ANÁLISE DE DADOS

OBJETIVOS DA UNIDADE 3

Ao final dos estudos, você deverá ser capaz de: Compreender o processo de exploração e análise de dados, desde a coleta inicial até a interpretação dos resultados finais, reconhecendo a importância de cada etapa no desenvolvimento de insights significativos.

3.1 Processo de exploração e análise de dados

Videoaula do tópico disponível no AVA:

Videoaula 9: Processo de exploração e análise de dados.

O processo de exploração e análise de dados é um estágio fundamental em data science, no qual os profissionais buscam compreender os conjuntos de dados que estão sendo analisados. Este processo geralmente começa com a coleta de dados brutos de diversas fontes, seguido pela sua limpeza e transformação. Esta fase é iterativa, onde os dados são explorados repetidamente à medida que novas informações são descobertas.

Durante a exploração dos dados, os cientistas de dados utilizam uma variedade de técnicas estatísticas e algoritmos para identificar padrões, tendências e possíveis anomalias nos dados. Os cientistas de dados têm à sua disposição várias técnicas para explorar e analisar dados, entre elas destacamos a Análise Descritiva e a Análise Exploratória de Dados (EDA), que são pontos de partida do trabalho.

- **Análise Descritiva:** A análise descritiva de dados é uma etapa inicial e introdutória da análise de dados. Esta técnica envolve a descrição básica dos dados. Para entender e descrever os dados o cientista ou analista de dados calcula médias, medianas, desvios padrão, percentis, produz visualizações básicas, como histogramas e gráficos de dispersão. Isso tudo na tentativa de resumir e visualizar as características básicas dos dados, procurando insights iniciais que podem orientar análises mais avançadas.
- **Análise Exploratória de Dados (EDA):** EDA procura examinar um problema, procurando por diferentes possibilidades e pontos de vista. Além das técnicas da própria análise descritiva, a EDA pode incluir, por exemplo, traçar correlações primárias entre diferentes conjuntos e variáveis. Envolve a criação de Matrizes de Correlação, Análise de Outliers, Visualizações Interativas ou Técnicas de Redução de Dimensionalidade.

A análise descritiva e exploratória de dados são partes integrantes de um mesmo processo de compreensão inicial e investigação. Embora frequentemente empregadas em conjunto, especialmente em análises quantitativas, essas abordagens não são idênticas. Enquanto a análise descritiva busca resumir fatos e identificar tendências, padrões e anomalias nos dados, a análise exploratória vai além, estabelecendo conexões, agrupamentos e correlações entre os dados, permitindo a geração de novos insights e entendimentos sobre o cenário

retratado. Mas, é importante destacar que uma compreensão sólida do processo de exploração de dados é essencial para garantir que as análises subsequentes sejam precisas e significativas.

Vamos fazer um exercício. O exemplo de código abaixo mostra uma análise descritiva e uma análise exploratório de um conjunto fictício de dados. Na sequencia apresentamos um possível relatório que seria apresentado ao gestor.

```
import pandas as pd

# Criando o dataframe de exemplo
data = {
    'Idade': [25, 35, 35, 40, 45, 45, 55, 55, 65, 200],  # Introduzindo uma idade anômala
    'Salario': [3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500]
}
df = pd.DataFrame(data)

# Análise Descritiva
print("Análise Descritiva:")
print(df.describe())

# Análise Exploratória
print("Análise Exploratória:")
print("Correlação entre Idade e Salário:")
print(df.corr())

print("Distribuição de Idade:")
print(df['Idade'].value_counts())

print("Média de Salário por Faixa Etária:")
bins = [20, 30, 40, 50, 60, 70]
labels = ['20-29', '30-39', '40-49', '50-59', '60-69']
df['Faixa Etária'] = pd.cut(df['Idade'], bins=bins, labels=labels, right=False)
print(df.groupby('Faixa Etária')['Salario'].mean())
```

Depois de carregar os dados, podemos explorá-los usando métodos como describe() para obter estatísticas descritivas e info() para informações sobre os tipos de dados e valores ausentes. No exemplo usamos o describe().

```
print(dados.describe())
print(dados.info())
```

```
Análise Descritiva:
            Idade      Salario
count   10.000000    10.000000
mean    60.000000  5250.000000
std     50.552503  1513.825177
min     25.000000  3000.000000
25%     36.250000  4125.000000
50%     45.000000  5250.000000
75%     55.000000  6375.000000
max    200.000000  7500.000000
```

A análise descritiva revelou que a média de idade dos funcionários é de aproximadamente 60 anos, com um desvio padrão de 50 anos, indicando uma dispersão significativa dos dados.

Em relação aos salários, observamos uma média de R$ 5250, com um desvio padrão de R$ 1513, evidenciando certa variabilidade nos ganhos dos colaboradores.

No entanto, identificamos uma anomalia na idade de um dos funcionários, que apresentou uma idade máxima registrada como 200 anos. Esta é uma discrepância significativa em relação às demais idades e pode indicar um erro de digitação ou coleta dos dados.

```
Análise Exploratória:
Correlação entre Idade e Salário:
            Idade   Salario
Idade    1.000000  0.700546
Salario  0.700546  1.000000
Distribuição de Idade:
Idade
35     2
45     2
55     2
25     1
40     1
65     1
200    1
Name: count, dtype: int64
Média de Salário por Faixa Etária:
Faixa Etária
20-29    3000.0
30-39    3750.0
40-49    5000.0
50-59    6250.0
60-69    7000.0
Name: Salario, dtype: float64
```

A correlação entre idade e salário foi calculada, revelando uma relação forte entre as duas variáveis (correlação de aproximadamente 0.70).

Ao analisarmos a distribuição de idade dos funcionários, notamos que a maioria está concentrada em faixas etárias mais baixas, com poucos funcionários acima dos 60 anos.

Para uma análise mais detalhada, agrupamos os funcionários em faixas etárias e calculamos a média de salário para cada faixa. Os resultados mostram uma tendência de aumento salarial conforme a faixa etária avança, indicando uma possível valorização da experiência e senioridade.

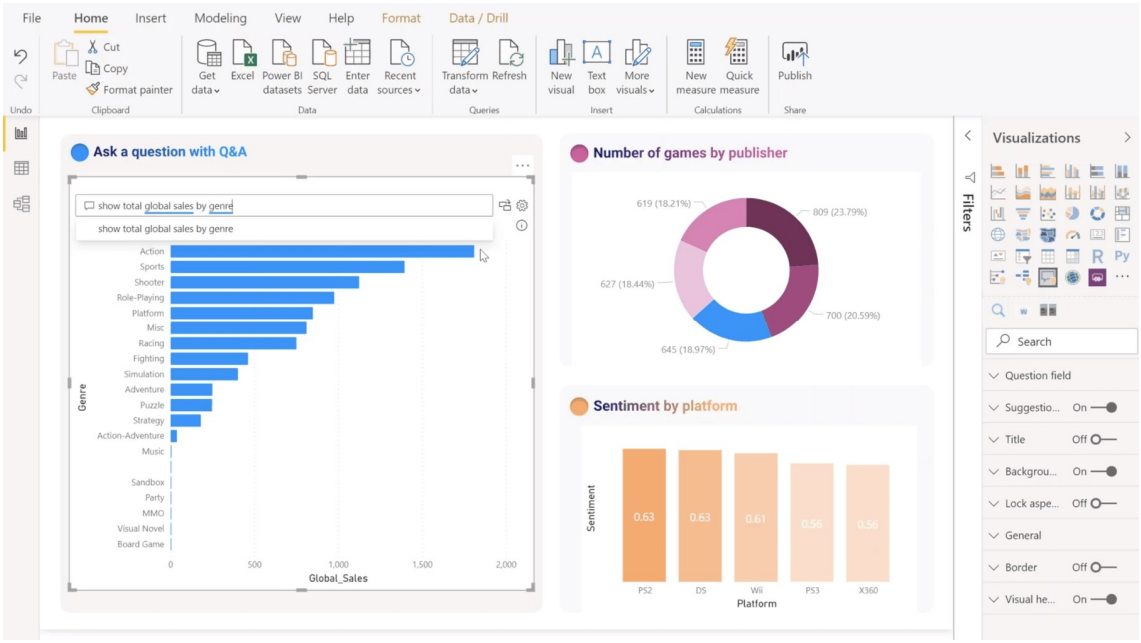
Claro que não é apenas isso que é feito. Existe um conjunto de técnicas de exploração de dados que podem e devem ser usadas para garantir que se entenda muito bem as características dos dados e quais procedimentos de limpeza são necessários. Cabe ao cientista ou analista de dados escolher as técnicas apropriadas.

Aprenda Mais:

Quais são os tipos de análise de dados e como aplicá-los para obter respostas mais precisas? Disponível em: https://niteolearning.com/blog/tipos-de-analise-de-dados-quais-sao-quando-como-aplicar/

3.2 Métodos de visualização de dados e sua importância na análise exploratória

Os métodos de visualização de dados desempenham um papel importante na análise exploratória, pois permitem que os cientistas de dados representem graficamente os padrões e relações presentes nos dados. Gráficos como histogramas, gráficos de dispersão e box plots são ferramentas comuns utilizadas para explorar a distribuição dos dados e identificar possíveis outliers. Por exemplo, ao utilizar um histograma, podemos visualizar a distribuição dos dados em intervalos, o que nos ajuda a entender a frequência de ocorrência de diferentes valores em uma variável. Da mesma forma, um gráfico de dispersão nos permite identificar relações entre duas variáveis, como correlações positivas ou negativas.



Fonte: https://blog.somostera.com/data-science/visualizacao-de-dados

Além dos gráficos básicos, existem visualizações mais avançadas que podem revelar insights complexos em conjuntos de dados multidimensionais. Por exemplo, mapas de calor são úteis para representar a densidade ou intensidade de uma variável em diferentes regiões de um espaço bidimensional. Já os gráficos de rede podem ser empregados para visualizar conexões e relacionamentos entre entidades em um conjunto de dados, como redes sociais ou redes de coautoria em artigos acadêmicos.

A visualização de dados não só facilita a compreensão dos dados, mas também ajuda na identificação de padrões ocultos que podem não ser evidentes apenas olhando para os números. Por exemplo, ao plotar um gráfico de dispersão e aplicar uma linha de tendência, podemos identificar padrões de crescimento ou decrescimento em uma série temporal. Esses padrões podem fornecer insights valiosos para tomar decisões informadas em diferentes domínios, como finanças, marketing ou saúde.

Além disso, as visualizações de dados são uma ferramenta poderosa para comunicar resultados para stakeholders não técnicos. Gráficos claros e intuitivos podem ajudar a transmitir informações complexas de forma compreensível, facilitando a tomada de decisões estratégicas. Portanto, investir em técnicas de visualização de dados é fundamental para uma análise exploratória eficaz, pois ajuda os cientistas de dados a extrair insights significativos e a comunicá-los de maneira eficaz para diferentes audiências.

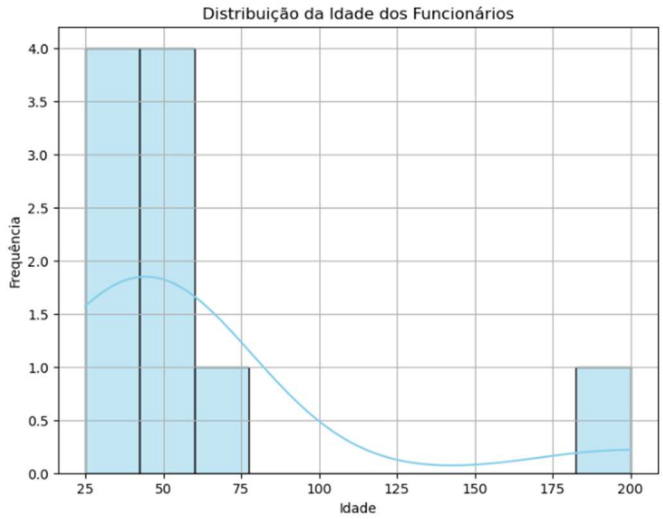
Abaixo temos alguns exemplos simples de visualizações de dados em Python. Existem muitas outras opções e personalizações disponíveis em bibliotecas como Matplotlib, Seaborn, Plotly, entre outras, para explorar e representar graficamente os dados de maneira eficaz e informativa.

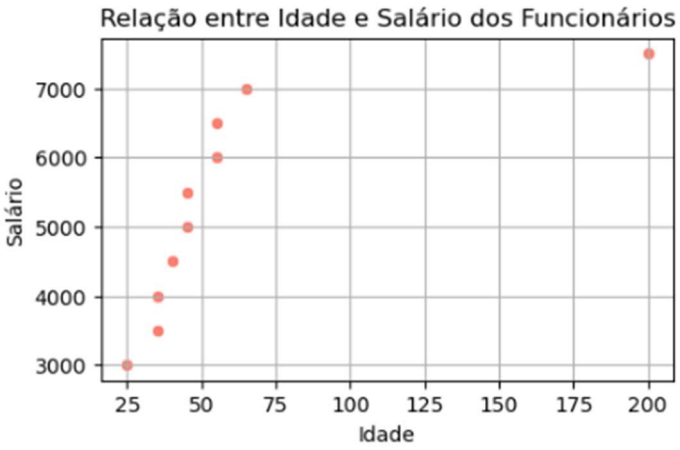
```python
import matplotlib.pyplot as plt
import numpy as np

# Gerando dados aleatórios
dados = np.random.normal(loc=0, scale=1, size=1000)

# Plotando histograma
plt.hist(dados, bins=30, color='blue', edgecolor='black')
plt.xlabel('Valores')
plt.ylabel('Frequência')
plt.title('Histograma dos Dados')
plt.show()
```

```python
import seaborn as sns
import numpy as np

# Gerando dados aleatórios
dados = np.random.normal(loc=0, scale=1, size=100)

# Plotando boxplot
sns.boxplot(data=dados, color='green')
plt.title('Boxplot dos Dados')
plt.show()
```

A análise que fizemos na seção anterior poderia ser melhorada acrescentando gráficos para identificar padrões e tendências nos dados.

```python
# Histograma da distribuição da idade
plt.figure(figsize=(8, 6))
sns.histplot(data=df, x='Idade', bins=10, kde=True, color='skyblue')
plt.title('Distribuição da Idade dos Funcionários')
plt.xlabel('Idade')
plt.ylabel('Frequência')
plt.grid(True)
plt.show()

# Gráfico de dispersão da relação entre idade e salário
plt.figure(figsize=(5, 3))
sns.scatterplot(data=df, x='Idade', y='Salario', color='salmon')
plt.title('Relação entre Idade e Salário dos Funcionários')
plt.xlabel('Idade')
plt.ylabel('Salário')
plt.grid(True)
plt.show()
```





3.3 Técnicas de pré-processamento de dados

Videoaula do tópico disponível no AVA:

Videoaula 11: Técnicas de pré-processamento de dados.

O pré-processamento de dados é uma etapa crítica na preparação dos dados para análise. Envolve uma série de técnicas, incluindo limpeza de dados para remover valores ausentes ou inconsistentes, normalização para garantir que todas as variáveis tenham a mesma escala e transformações para tornar os dados mais adequados para modelos de análise específicos.

Um pré-processamento de dados eficaz é fundamental para garantir que os resultados da análise sejam confiáveis e significativos, evitando conclusões errôneas ou enviesadas devido a dados mal preparados.

Uma das técnicas mais comuns é a limpeza de dados, que visa remover valores ausentes, inconsistentes ou duplicados que podem distorcer os resultados da análise. Por exemplo, podemos usar o método dropna() em Pandas para remover linhas ou colunas com valores ausentes em um DataFrame do Python.

No próximo exemplo, criamos um conjunto de dados aleatórios contendo 100 registros e 5 colunas. Para simular dados ausentes, introduzimos valores NaN (não numéricos) em algumas células aleatórias das colunas 'A' e 'B'. Em seguida, realizamos uma etapa de pré-processamento para lidar com esses dados ausentes. Utilizamos o método dropna() da biblioteca pandas para eliminar todas as linhas que contenham pelo menos um valor NaN. Isso resultou em um conjunto de dados limpo, sem valores ausentes, pronto para análise posterior.

```python
import pandas as pd
import numpy as np

# Gerando um conjunto de dados aleatórios com 100 registros e 5 colunas
np.random.seed(0)  # Para garantir a reprodutibilidade dos resultados
data = pd.DataFrame(np.random.randint(0, 100, size=(100, 5)), columns=['A', 'B', 'C', 'D', 'E'])

# Introduzindo dados ausentes aleatoriamente em algumas células
missing_indices = np.random.choice(data.index, size=20, replace=False)
data.loc[missing_indices, 'A'] = np.nan
data.loc[missing_indices, 'B'] = np.nan

# Visualizando o conjunto de dados com dados ausentes
print("Conjunto de Dados com Dados Ausentes:")
print(data.head())

# Pré-processamento: Eliminando linhas com dados ausentes
data_cleaned = data.dropna()

# Visualizando o conjunto de dados após a eliminação dos dados ausentes
print("\nConjunto de Dados Após Eliminação de Dados Ausentes:")
print(data_cleaned.head())
```

```
Conjunto de Dados com Dados Ausentes:
      A     B   C   D   E
0  44.0  47.0  64  67  67
1   NaN   NaN  21  36  87
2  70.0  88.0  88  12  58
3  65.0  39.0  87  46  88
4  81.0  37.0  25  77  72

Conjunto de Dados Após Eliminação de Dados Ausentes:
      A     B   C   D   E
0  44.0  47.0  64  67  67
2  70.0  88.0  88  12  58
3  65.0  39.0  87  46  88
4  81.0  37.0  25  77  72
5   9.0  20.0  80  69  79
```

Outra técnica importante é a normalização, que é utilizada para garantir que todas as variáveis tenham a mesma escala. Isso é especialmente importante em algoritmos de machine learning que são sensíveis à escala das variáveis. O scikit-learn, por exemplo, oferece uma classe chamada MinMaxScaler para realizar a normalização de dados.

```python
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
dados_normalizados = scaler.fit_transform(dados)
```

Além disso, as técnicas de redução de dimensionalidade, como a Análise de Componentes Principais (PCA), são frequentemente aplicadas para lidar com conjuntos de dados de alta dimensionalidade. Isso pode ajudar a reduzir a complexidade dos dados, preservando ao mesmo tempo as informações mais relevantes. Por exemplo, podemos usar o PCA para reduzir um conjunto de dados

de alta dimensionalidade para duas dimensões, facilitando a visualização e a análise.

```
from sklearn.decomposition import PCA

pca = PCA(n_components=2)
dados_reduzidos = pca.fit_transform(dados)
```

Esses são apenas alguns exemplos de técnicas que poderiam ser utilizadas para realizar o pré-processamento e evitar conclusões errôneas ou enviesadas devido a dados mal preparados.

3.4 Importância da interpretação dos resultados na análise de dados

Videoaula do tópico disponível no AVA:

Videoaula 12: Treinamento do modelo

A interpretação dos resultados na análise de dados envolve compreender não apenas os resultados estatísticos, mas também o contexto em que foram obtidos e as possíveis limitações dos dados e das análises realizadas. Habilidades de comunicação são essenciais para transmitir esses resultados de forma clara e compreensível para stakeholders não técnicos. Uma interpretação cuidadosa dos resultados ajuda a evitar conclusões precipitadas ou interpretações errôneas, garantindo que as decisões baseadas em dados sejam sólidas e confiáveis.

UNIDADE 4 MODELAGEM E APRENDIZAGEM DE MÁQUINA

OBJETIVOS DA UNIDADE 4

Ao final dos estudos, você deverá ser capaz de: Compreender os conceitos fundamentais de modelagem e aprendizado de máquina, bem como a importância da escolha adequada do algoritmo para cada problema específico.

4.1 Conceitos básicos de modelagem e aprendizado de máquina

Videoaula do tópico disponível no AVA:

Videoaula 13: Conceitos básicos de modelagem e aprendizado de máquina.

Imagine que você tem uma grande pilha de dados sobre vendas de uma loja. Você deseja entender melhor esses dados para prever as vendas futuras ou identificar padrões de compra dos clientes. Aqui é onde entra a modelagem e o aprendizado de máquina.

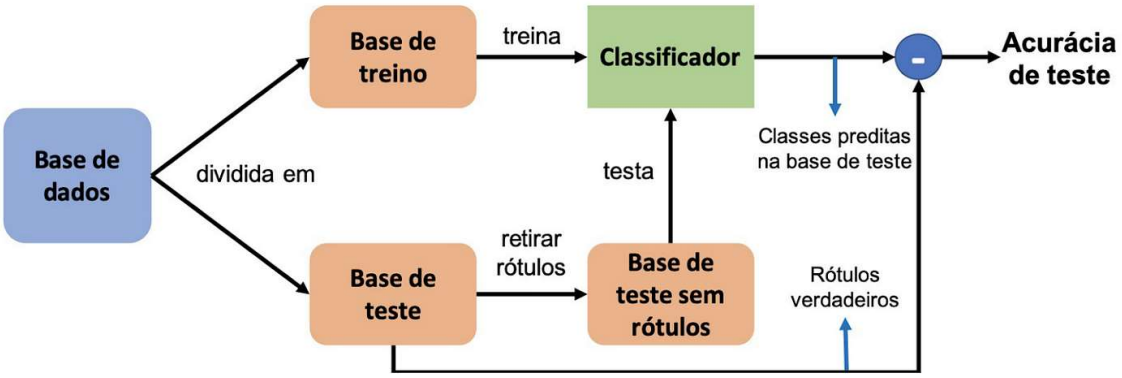
Modelagem e aprendizado de máquina são como a caixa de ferramentas que os cientistas de dados usam para extrair informações valiosas de dados complexos. Um modelo de aprendizado de máquina é a representação matemática ou computacional do nosso problema. É como uma fórmula matemática que aprende com os dados de treinamento para fazer previsões ou tomar decisões.

Por exemplo, um modelo de aprendizado de máquina pode aprender com os dados históricos de vendas para prever quantas vendas você pode esperar em um determinado dia da semana ou temporada do ano.

Para treinar um modelo de aprendizado de máquina são necessários muitos dados. Esses dados são divididos em dados de teste e dados de treinamento. Para serem usados os todos os dados precisam ter passado pela fase de pré-processamento.

Os dados de treinamento são a base fundamental para o desenvolvimento e treinamento de modelos de aprendizado de máquina. Eles consistem em um conjunto de exemplos que são usados para ensinar o modelo a fazer previsões ou tomar decisões com base nos padrões presentes nos dados.

O papel dos dados de treinamento é essencialmente ensinar o modelo a aprender a partir de exemplos passados, para que possa generalizar e fazer previsões precisas sobre novos dados no futuro. Imagine que você está treinando um modelo para reconhecer cães em fotos. Seus dados de treinamento seriam imagens de cães previamente rotuladas como "cães", junto com outras imagens de coisas que não são cães, como "gatos", "carros", "árvores", etc. Ao expor o modelo a esses exemplos variados, ele aprende a identificar padrões visuais que são comuns a imagens de cães e a distinguir cães de outras coisas.



Fonte: https://tatianaesc.medium.com/machine-learning-conceitos-e-modelos-f0373bf4f445

A figura acima criada por Escovedo mostra o esquema de treinamento de uma modelo de classificação a autora sugere que “a base de treino é submetida ao modelo (classificador) para que seus parâmetros sejam calibrados de acordo com os dados apresentados. Após esta etapa, ocorre a etapa de predição de classes: os exemplos da base de teste são apresentados para o modelo treinado para que este realize a predição de suas classes”. Podemos perceber que os dados de treinamento são usados para treinar e construir um modelo genérico e os dados de testes são usados para testar se a eficiência pretendida para o modelo foi atingida.

Sem dados de treinamento e testes suficientes e representativos, o modelo não terá informações suficientes para aprender e não será capaz de fazer previsões precisas sobre novos dados. Portanto, a qualidade e a quantidade dos dados de treinamento e testes desempenham um papel crucial no sucesso do modelo de aprendizado de máquina.

Aprenda Mais:

Machine Learning: Conceitos e Modelos — Parte I: Aprendizado Supervisionado. Disponível em: https://tatianaesc.medium.com/machine-learning-conceitos-e-modelos-f0373bf4f445

4.2 Tipos de algoritmos de aprendizado de máquina e suas aplicações

Videoaula do tópico disponível no AVA:

Videoaula 14: Tipos de algoritmos de aprendizado de máquina e suas aplicações.

No mundo do aprendizado de máquina, existem diferentes tipos de algoritmos, cada um com sua própria maneira de aprender com os dados e fazer previsões. São os clássicos:

- **Regressão**: É como uma linha de melhor ajuste que tenta prever valores contínuos com base em variáveis de entrada. Por exemplo, prever o preço de uma casa com base em características como tamanho, número de quartos, etc.
- **Classificação**: Este tipo de algoritmo é usado quando queremos classificar algo em categorias. Por exemplo, identificar se um e-mail é spam ou não, com base no texto do e-mail e em outros atributos.
- **Agrupamento**: Agrupamento é usado quando queremos encontrar padrões nos dados sem ter rótulos explícitos. Por exemplo, agrupar clientes com base em padrões de compra semelhantes, como fazemos em uma loja online.
- **Aprendizado Profundo:** Este tipo de aprendizado de máquina é inspirado no funcionamento do cérebro humano e é usado para lidar com dados complexos, como imagens, áudio e texto. Redes neurais profundas são um exemplo disso.

4.3 Avaliação de modelos de aprendizado de máquina

Videoaula do tópico disponível no AVA:

Videoaula 15: Avaliação de modelos de aprendizado de máquina.

Imagine que você treinou um modelo para identificar padrões de compra dos clientes, e sempre que você fornece dados de um novo cliente ele prevê o padrão de compra, mas como você sabe se ele está funcionando corretamente? É aqui que a avaliação de modelos entra em cena. Depois de treinar um modelo com os dados de treinamento, é importante entender como ele se comporta e quão confiável é em fazer previsões ou tomar decisões.

Isso é feito executando o modelo com os dados de testes, que ainda não são conhecidos do modelo. As respostas preditas pelo modelo são comparadas com

as respostas corretas dos dados de teste, e então calculado métricas de definem o quanto o modelo acertou.

A importância da avaliação do modelo reside no fato de que ela nos permite entender a eficácia e a capacidade de generalização do modelo. Em outras palavras, ela nos ajuda a responder à pergunta: "O quão bem o modelo está realizando suas previsões ou tomadas de decisão em dados que não foram vistos durante o treinamento?"

Ao avaliar um modelo, podemos determinar se ele está sofrendo de overfitting (ajuste excessivo aos dados de treinamento), underfitting (não sendo capaz de capturar a complexidade dos dados) ou se está funcionando de forma ideal para o problema em questão.

Além disso, a avaliação do modelo nos ajuda a comparar diferentes modelos e escolher o mais adequado para a tarefa em mãos. Por exemplo, podemos comparar o desempenho de diferentes algoritmos ou ajustar os parâmetros de um único algoritmo para melhorar seu desempenho.

4.4 Desafios e limitações do aprendizado de máquina

Embora o aprendizado de máquina seja incrivelmente poderoso, ele também enfrenta alguns desafios importantes:

- **Interpretabilidade**: Modelos complexos, como redes neurais, podem ser como "caixas-pretas", difíceis de entender. Isso pode ser problemático em situações em que precisamos explicar as decisões do modelo.
- **Viés e Discriminação:** Modelos de aprendizado de máquina podem refletir e até amplificar viés e discriminação presentes nos dados de treinamento. Isso pode levar a decisões injustas ou prejudiciais.
- **Escassez de Dados:** Algoritmos de aprendizado de máquina geralmente precisam de muitos dados para funcionar bem. A falta de dados pode levar a modelos que não generalizam bem para novas situações.
- **Overfitting e Underfitting**: Estes são problemas comuns em que nosso modelo se ajusta muito bem ou muito mal aos dados de treinamento, respectivamente, tornando-se incapaz de generalizar para novos dados.
- **Custo Computacional:** Algoritmos complexos, como redes neurais profundas, podem exigir muitos recursos computacionais para treinamento e inferência, tornando-os caros em termos de tempo e hardware.

Estes são apenas alguns dos desafios que os cientistas de dados enfrentam ao trabalhar com aprendizado de máquina. Navegar por esses desafios requer uma abordagem cuidadosa e ética, além de um entendimento sólido dos conceitos e técnicas subjacentes.

Compreender esses tópicos fundamentais é essencial para qualquer pessoa interessada em explorar o mundo emocionante da modelagem e do aprendizado de máquina. Ao dominar esses conceitos, você estará preparado para enfrentar uma variedade de problemas do mundo real e construir soluções inteligentes e eficazes.

UNIDADE 5 TÓPICOS AVANÇADOS E TENDÊNCIAS EM CIÊNCIA DE DADOS

OBJETIVOS DA UNIDADE 5

Ao final dos estudos, você deverá ser capaz de: Explorar Redes Neurais, Deep Learning, Visão Computacional, Processamento de Linguagem Natural (NLP) e Modelos Generativos compreendendo os princípios fundamentais por trás dessas técnicas e sua aplicação em problemas complexos de modelagem e predição.

5.1 Redes Neurais e Deep Learning

As redes neurais e o deep learning são tecnologias que têm revolucionado a forma como lidamos com dados complexos, especialmente em áreas como reconhecimento de padrões, processamento de linguagem natural e visão computacional. Uma rede neural é um modelo matemático inspirado no funcionamento do cérebro humano, composto por camadas de neurônios interconectados.

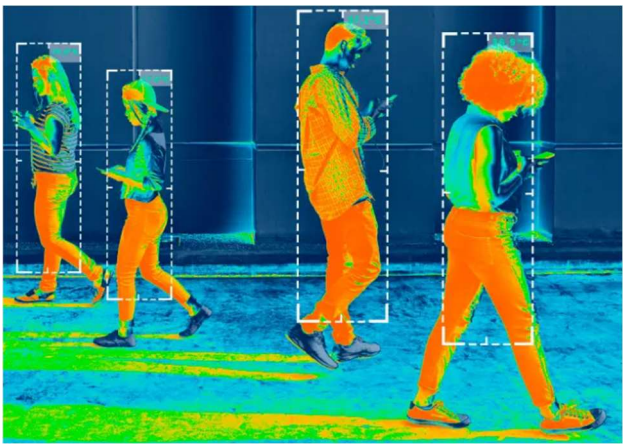
O deep learning, por sua vez, refere-se ao treinamento de redes neurais com muitas camadas, permitindo que elas aprendam representações de dados cada vez mais abstratas e complexas.

Exemplo: Um exemplo prático de aplicação de redes neurais e deep learning é o reconhecimento de voz por assistentes virtuais como o Siri da Apple ou o Google Assistant. Esses sistemas utilizam modelos de deep learning para interpretar e responder aos comandos dos usuários com precisão impressionante.

5.2 Visão Computacional

A visão computacional é uma área da ciência de dados que se concentra em capacitar os computadores a entenderem e interpretarem o mundo visual, de maneira semelhante ao cérebro humano. Isso envolve técnicas como detecção de objetos, reconhecimento facial, segmentação de imagens e muito mais.

Exemplo: Um exemplo de aplicação de visão computacional é a tecnologia de reconhecimento facial utilizada em sistemas de segurança e identificação em redes sociais, como o Facebook. Esses sistemas são capazes de identificar rostos em fotos, facilitando a marcação de amigos e a organização de álbuns.



Fonte: https://lyncas.net/visao-computacional-automatizar-processos/

Aprenda Mais:

Visão Computacional: 5 aplicações para aprimorar processos Disponível em: https://lyncas.net/visao-computacional-automatizar-processos/

5.3 Processamento de Linguagem Natural (NLP)

O processamento de linguagem natural é uma área da ciência de dados que se concentra na interação entre computadores e linguagem humana. Isso inclui tarefas como tradução automática, análise de sentimentos, geração de texto e muito mais.

Exemplo: Um exemplo de aplicação de processamento de linguagem natural é a assistência virtual em serviços de atendimento ao cliente, como os chatbots. Esses sistemas utilizam algoritmos de NLP para compreender e responder às perguntas dos usuários de forma natural e eficiente.

Fonte: https://lyncas.net/visao-computacional-automatizar-processos/

Aprenda Mais:

Tudo sobre NLP: o que é? Quais os desafios? Disponível em: https://www.blip.ai/blog/tecnologia/nlp-processamento-linguagem-natural/

5.4 Modelos Generativos

Um modelo generativo é uma abordagem na área de aprendizado de máquina que visa criar novos dados que se assemelham aos dados de treinamento. Ao contrário dos modelos discriminativos, que se concentram em prever a probabilidade de uma classe ou categoria específica para um dado conjunto de entrada, os modelos generativos têm como objetivo aprender a distribuição de probabilidade subjacente aos dados de treinamento. Isso permite que eles gerem novas amostras que são indistinguíveis das amostras originais.

As Large Language Models (LLMs), como o GPT (Generative Pre-trained Transformer), também podem ser consideradas modelos generativos. Elas absorvem uma vasta quantidade de texto durante o treinamento e aprendem a estrutura, a gramática e o estilo das linguagens humanas. Posteriormente, podem gerar texto coerente e contextualmente relevante com base em uma dada entrada inicial. Assim como outros modelos generativos, as LLMs utilizam uma abordagem de aprendizado de máquina para criar novos exemplos de texto que, idealmente, se parecem com os dados de treinamento.

No entanto, a diferença chave entre um modelo generativo tradicional, como uma GAN, e as LLMs reside no modo como geram novos dados. Enquanto as GANs geram dados diretamente no espaço de entrada, como imagens ou sons, as LLMs geram texto sequencialmente, palavra por palavra, com base na distribuição de probabilidade aprendida durante o treinamento. Isso permite que as LLMs capturem não apenas a estrutura estatística das palavras individuais, mas também a dependência de longo alcance entre elas, produzindo texto que mantém coerência e contexto.

Além disso, as LLMs têm uma vantagem adicional na capacidade de realizar tarefas específicas, como tradução automática, sumarização de texto e geração de código, além de simplesmente gerar texto. Essa flexibilidade decorre da capacidade das LLMs de aprender representações latentes complexas dos dados durante o treinamento, permitindo-lhes entender e manipular informações de maneiras que vão além da simples geração de texto. Em resumo, embora as LLMs compartilhem o objetivo fundamental de gerar dados, sua arquitetura e aplicação diferem dos modelos generativos tradicionais, refletindo a complexidade e a versatilidade das abordagens contemporâneas em inteligência artificial.

FINALIZAR

Chegamos ao fim desta jornada de exploração dos fundamentos da Ciência de Dados. Durante esta disciplina, mergulhamos em conceitos essenciais e práticas fundamentais para dominar esta área da tecnologia da informação.

Começamos compreendendo a importância da Ciência de Dados na era da informação, explorando sua relação com a análise e engenharia de dados. Discutimos também a interseção entre inteligência artificial e ciência de dados, reconhecendo como essas disciplinas se complementam e impulsionam avanços significativos em diversos setores.

Em seguida, falamos sobre o universo das tecnologias e ferramentas em Ciência de Dados, explorando linguagens como Python e R, bem como exemplos de frameworks como Scikit-learn, Matplotlib, Seaborn e PyTorch. Investigamos plataformas e ambientes de desenvolvimento como Jupyter Notebooks e Google Colab, além de ferramentas de análise de dados como Tableau, PowerBI e outros frameworks.

Na etapa seguinte, dedicamos tempo para compreender o processo de exploração e análise de dados, explorando métodos de visualização e técnicas de pré-processamento de dados. Reconhecemos a importância crucial da interpretação dos resultados na análise de dados, desenvolvendo habilidades críticas para avaliar e comunicar descobertas de forma clara e precisa.

Finalmente, exploramos a modelagem e o aprendizado de máquina, abordando desde conceitos básicos até tipos de algoritmos, avaliação de modelos e desafios enfrentados nesse campo.

Convido você a continuar explorando e aprofundando seus conhecimentos nesse fascinante campo da Ciência de Dados. Mantenha-se atualizado com as últimas tendências e tecnologias, e nunca deixe de praticar e aprimorar suas habilidades. Lembre-se de que a Ciência de Dados é um campo em constante evolução, e seu domínio pode abrir portas para oportunidades emocionantes em sua carreira profissional.

Profa. Dra. Joelma de Moura Ferreira

Sobre a autora

Joelma de Moura Ferreira é doutora em Ciência da Computação pela Universidade Federal de Goiás, com mestrado em Ciência da Computação pela Universidade Federal de Goiás, MBA em Gerenciamento de Projetos pela Fundação Getúlio Vargas, especialização em Redes de Computadores pela Universidade Salgado de Oliveira, MBA em Tecnologia para Negócios: AI, Data Science e Big Data pela Pontifícia Universidade Católica do Rio Grande do Sul e graduação em Ciência da Computação pela Universidade Católica de Goiás. Tendo atuado por mais de 20 anos como docente de graduação e pós-graduação em diversas instituições de ensino superior, incluindo Faculdade Sul-Americana, Universidade Paulista, Faculdade Estácio de Sá de Goiás, Pontifícia Universidade Católica, Centro Universitário Alves Farias. Desempenhou a função de coordenadora do curso de graduação de Sistemas de Informação e dos cursos de pós-graduação em Gestão de Projetos, Gestão de Tecnologia da Informação e Arquitetura e Engenharia de Software no Centro Universitário Alves Faria, onde também exerceu a atividade de pesquisadora no Mestrado em Desenvolvimento R Fora do domínio acadêmico, exerce a função de Cientista de Dados no Tribunal de Justiça do Distrito Federal e Territórios.

Referências Bibliográficas

VALDATI, A.B, **Inteligência artificial – IA**, Contentus, 2020.

GABRIEL, Martha. **Inteligência Artificial: Do Zero ao Metaverso**, Atlas, 2022.

GÉRON, Aurélien . **Mãos à Obra Aprendizado de Máquina com Scikit-Learn,**

GRUS, Joel. **Data Science do Zero,** Alta Books, 2019.

RUSSEL, Stuart J., NORVIG, P. In**teligência Artificial - Uma Abordagem Moderna**, LTC, 2022.

KAUFMAN, D., **Desmistificando a inteligência artificial**, Autentica, 2022.

EYSENCK, Michael W., EYSENCK, Christine. **Inteligência Artificial X Humanos: O que a Ciência Cognitiva nos Ensina ao Colocar Frente a Frente a Mente Humana e a IA.** Artmed, 2023

VILENKY, Renata. In**teligência Artificial - Uma oportunidade para você empreender**. Expressa, 2021

MILANI, A.M. et al. **Visualização de Dados,** SAGAH, 2020.