Data Science

TechGuide - Alura, FIAP e PM3

Data Science

Nivel 1

- [ ] **Ciência de Dados - Fundamentos**:
    - A Ciência de Dados é o ato de utilizar algoritmos e sistemas para extrair, organizar e analisar dados a partir de diversas fontes, a fim de detectar padrões e tomar decisões de negócios. As áreas de aplicação são infindáveis, como em negócios, biologia, medicina, engenharia, etc.
    - Conhecer o conceito de Exploração de Dados
    - Conhecer as principais funções, como 'describe', 'info', etc
    - Entender o papel de visualizações como histogramas e boxplots
    - Saber o que são variáveis categóricas nominais e ordinais
    - Explorar os dados em Python com as bibliotecas Pandas, Matplotlib, Seaborn e Numpy
- [ ] **Feature Engineering**:
    - Feature Engineering (Engenharia de atributos/características) refere-se ao processo de usar o conhecimento de domínio para selecionar e transformar as variáveis mais relevantes dos dados brutos ao criar um modelo preditivo usando aprendizado de máquina ou modelagem estatística, a fim de melhorar o desempenho dos algoritmos de aprendizado de máquina.
    - Gerar novas variáveis a partir dos dados disponíveis

- Transformar dados brutos em características
- Realçar os problemas dos modelos preditivos
- Melhorar a precisão do modelo para novos dados
- Selecionar e criar features utilizando as bibliotecas Pandas e Scikit-learn

☐ **Extração e Tratamento de Dados**:

- A extração de dados é o processo de coleta ou recuperação de tipos diferentes de dados de uma variedade de fontes, muitos dos quais podem estar mal organizados ou completamente desestruturados.
- Obter os dados que serão analisados
- Tratar os dados obtidos, transformando-os, alterando sua estrutura e valores a fim de deixar a base de dados mais coerente e garantir que os dados que serão trabalhados estejam nas melhores condições para serem analisados
- Utilizar as bibliotecas Pandas e Scikit-learn para tratar os dados

☐ **Python para Ciência de Dados**:

- Python é uma linguagem de programação interpretada de alto nível e que suporta múltiplos paradigmas de programação, como imperativo, orientado a objetos e funcional. É uma linguagem com tipagem dinâmica e gerenciamento automático de memória.
- Aprender lógica de programação em Python
- Aprenda os fundamentos da linguagem como variáveis, funções, listas, condicionais e imports
- Criar análises de dados
- Utilizar o Matplotlib para gerar gráficos
- Usar e manipular listas para agrupar dados
- Conhecer a biblioteca NumPy
- Conhecer a biblioteca Pandas

☐ **Jupyter & Colab notebooks**:

- Jupyter Notebook e Google Colaboratory são Notebooks que permitem a criação de blocos de texto e blocos de código
- Os Notebooks facilitam a elaboração de projetos de Data Science por ser possível visualizar o resultado da execução logo após o trecho de código
- O Google Colaboratory permite escrever e executar códigos Python diretamente no navegador, sem nenhuma ou poucas configurações necessárias
- Essas ferramentas facilitam o compartilhamento de projetos entre o time

☐ **R para Ciência de Dados**:

- R é uma linguagem de programação comumente usada em estatística computacional e análise de dados.
- Aprender a analisar, limpar e visualizar dados
- Elaborar gráficos
- Juntar bancos de dados
- Tirar sumários estatísticos
- Aprender modelos preditivos no R

☐ **Estatística e Matemática - Fundamentos**:

- Equações, funções e limites
- Logaritmos
- Matrizes, determinantes, vetores e espaço vetorial
- Derivadas e integrais
- Diferença entre média, mediana e moda
- Distribuição de frequência
- Variância e desvio padrão
- Distribuição binomial, poisson e normal
- Nível e intervalo de confiança
- Técnicas de amostragem
- Introdução à Regressão linear

- Séries temporais

- [ ] **Visualização de Dados**:

- A visualização de dados é uma expressão contemporânea da comunicação visual que consiste na representação visual de dados.
- Mapear dados abstratos em representações visuais
- Representar visualmente os dados que estão presentes no nosso mundo real
- Usar Python, Matplotlib and Seaborn para gerar visualizações de dados

Nivel 2

- [ ] **Machine Learning - Fundamentos**:

- O Aprendizado de Máquina ou Machine Learning é um subcampo da Engenharia e da Ciência da Computação que evoluiu do estudo de reconhecimento de padrões e da teoria do aprendizado computacional em inteligência artificial.
- Aprendizado supervisionado
- Utilizar algoritmos de classificação
- Utilizar algoritmos de regressão
- Utilizar o Scikit-learn para criar modelos de machine learning

- [ ] **Machine Learning - Aprendizado Não Supervisionado**:

- O Aprendizado Não Supervisionado utiliza algoritmos de Aprendizado de Máquina para analisar e agrupar conjuntos de dados não rotulados. Estes algoritmos descobrem padrões ocultos ou agrupamentos de dados sem a necessidade de intervenção humana.
- O clustering ou análise de agrupamento de dados é o conjunto de técnicas de prospecção de dados (data mining) que visa fazer agrupamentos automáticos de dados segundo o seu grau de semelhança.
- Conhecer a análise exploratória

- Utilizar os métodos K-means, DBSCAN e Mean shift para agrupar dados sem classificação
- Avaliar a qualidade de uma clusterização
- Parametrizar métodos de clusterização através do máximo coeficiente de silhueta
- Entender a matemática por trás das métricas de validação Silhouette, Davies Bouldin e Calinski Harabasz
- Conhecer técnicas de redução de dimensionalidade

☐ **Machine Learning - Avaliação de Modelos**:

- A Avaliação de Modelos é o processo que utiliza métricas para nos ajudar a analisar se um modelo treinado terá um bom desempenho de previsão quando exposto a novos conjuntos de dados.
- Conhecer diferentes estratégias de avaliação e otimização de modelos
- Utilizar um pipeline para treino e validação
- Conhecer as principais métricas de avaliação de modelos de machine learning

☐ **SQL - Fundamentos**:

- SQL (Structured Query Language, traduzindo, Linguagem de Consulta Estruturada) é uma linguagem de programação padronizada que é usada para gerenciar bancos de dados relacionais e realizar várias operações sobre os dados neles contidos.
- Conhecer os comandos mais comuns do SQL
- Usar SELECT para consultar uma tabela
- Usar INSERT para inserir dados em uma tabela
- Usar UPDATE para atualizar uma tabela
- Usar DELETE para remover dados de uma tabela
- Usar JOIN para conectar os dados de múltiplas tabelas
- Conhecer as cláusulas (FROM, ORDER BY, etc)

- [ ] **Testes Estatísticos**:
  - Testes estatísticos são usados para examinar as relações entre as variáveis e as hipóteses de teste.
  - Criar intervalos de confiança para amostras
  - Comparar grupos de amostras
  - Executar testes estatísticos
  - Planejar experimentos para a coleta de dados
  - Propor modelos matemáticos para entender um dado problema
  - Construir mapas de cores para auxiliar a interpretação dos dados
- [ ] **Regressão Linear e Logística**:
  - As regressões são os métodos mais simples de aprendizado supervisionado, porém encontram diversas aplicações.
  - A regressão linear é usada para relacionar uma variável dependente contínua a uma ou mais variáveis independentes contínuas. O objetivo é encontrar uma relação linear que melhor se ajuste aos dados.
  - A regressão logística, por outro lado, é usada para problemas de classificação binária, onde a variável de saída é categórica e possui apenas duas categorias. Ela estima a probabilidade de um evento ocorrer com base nas variáveis independentes.
- [ ] **Web Scraping**:
  - Web scraping ou raspagem de dados na web é a extração de dados de websites.
  - Usar o BeautifulSoup e Python para coletar dados
  - Pesquisar e navegar no HTML
  - Acessar o conteúdo e atributos das tags HTML
  - Construir datasets com os resultados das raspagens

Nivel 3

- [ ] **Deep Learning**:
  - Deep Learning é um ramo de Machine Learning baseado em um conjunto de algoritmos que tentam modelar abstrações de alto nível de dados usando um grafo profundo com várias camadas de processamento, compostas de várias transformações lineares e não lineares.
  - Construir e treinar modelos com Keras
  - Construir e treinar modelos com Tensorflow
  - Selecionar as camadas de um modelo
  - Classificar imagens
  - Entender os conceitos de pesos e vieses
  - Redes neurais para regressão
  - Entender o conceito de Redes recorrentes
- [ ] **Aprendizado por Reforço**:
  - O Aprendizado por Reforço é uma área de Machine Learning que se preocupa com a forma como agentes inteligentes devem tomar medidas num ambiente, a fim de maximizar a noção de recompensa cumulativa.
  - Entender os conceitos de agente e recompensa
  - Entender a diferença entre reforço positivo e negativo
  - Conhecer o modelo Markov Decision Process
  - Entender o conceito de Retorno
  - Utilizar o algoritmo Q-learning
- [ ] **Visão Computacional**:
  - Visão Computacional é um campo científico interdisciplinar que lida com a forma como os computadores podem ganhar conhecimentos de alto nível a partir de imagens ou vídeos digitais. Da perspectiva da engenharia, procura compreender e automatizar tarefas que o sistema visual humano pode fazer.
  - Extrair regiões de interesse de uma imagem

- Normalizar e pré-processar dados de imagens
- Construir classificadores para reconhecimento de faces
- Extrair regiões do rosto humano baseado em marcos faciais
- Analisar diferentes condições de cada componente do rosto humano
- Conhecer Redes Neurais Convolucionais
- Usar OpenCV

- [ ] **Processamento de Linguagem Natural**:

- Processamento de língua natural (PLN) é uma subárea da inteligência artificial e da linguística que estuda os problemas da geração e compreensão automática de línguas humanas naturais.
- Análise de Sentimento
- Criar visualizações para facilitar a análise de dados textuais
- Conhecer as bibliotecas NLTK e Scikit-Learn
- Normalizar textos
- Usar TF-IDF e Ngrams para melhorar a classificação
- Conhecer o conceito de Transformers e como são aplicados para LLMs
- Utilizar o SKlearn
- Utilizar Regex em PLN
- Conhecer o Word2Vec
- Combinar vetores de palavras para representar textos e classificá-los

- [ ] **Previsão de Séries temporais**:

- A previsão de séries temporais (Time series Forecasting), também conhecida como previsão temporal, é uma técnica de análise estatística que envolve a previsão de valores futuros ou padrões com base em dados históricos ordenados no tempo. Em uma série temporal, os dados são coletados sequencialmente em intervalos regulares, como horários, diários, mensais ou anuais, e exibem dependências temporais.

- [ ] **MLOps**:

- O MLOps permite que os modelos de Machine Learning sejam implantados de maneira rápida e confiável, o que é especialmente importante em empresas que lidam com grandes volumes de dados.
- Ajuda a garantir a qualidade e confiabilidade dos modelos de Machine Learning em produção, além de facilitar a manutenção e atualização desses modelos.
- Permite que as pessoas engenheiras de dados trabalhem em conjunto com cientistas de dados e desenvolvedores de software para implementar soluções de Machine Learning em larga escala.
- Ajuda a garantir a governança e a conformidade dos modelos de Machine Learning com as políticas e regulamentações da empresa.

Habilidade Auxiliar: Cloud, Big Data e Sistemas

☐ **Engenharia de Dados - Fundamentos**:

- Uma pessoa Engenheira de Dados desempenha um papel crucial ao projetar, implementar e manter as infraestruturas de dados que permitem que uma organização trabalhe de maneira eficiente com suas informações. A pessoa engenheira de dados cria pipelines de dados para integrar, limpar e transformar dados provenientes de várias fontes e formatos. Isso possibilita a geração de insights valiosos para o negócio e melhora a tomada de decisões estratégicas.

☐ **Big Data - Fundamentos**:

- Big Data refere-se a conjuntos de dados extremamente grandes e complexos, que não podem ser facilmente processados ou gerenciados por métodos tradicionais. Esses conjuntos de dados são caracterizados por seu volume massivo, velocidade de geração e variedade de tipos e formatos. Com a explosão da quantidade de dados gerados diariamente, provenientes de diversas fontes como redes sociais, dispositivos IoT e transações comerciais, o Big Data apresenta desafios e oportunidades para extrair insights valiosos e tomar decisões informadas.
- A análise de Big Data envolve o uso de técnicas e tecnologias avançadas, como armazenamento distribuído, processamento em paralelo e

aprendizado de máquina, para explorar e transformar esses dados em informações significativas para empresas, organizações e pesquisadores.

- [ ] **Cloud - Fundamentos**:

  - Cloud, ou computação em nuvem é a distribuição de serviços de computação pela Internet usando um modelo de preço pago conforme o uso. Uma nuvem é composta de vários recursos de computação, que abrangem desde os próprios computadores (ou instâncias, na terminologia de nuvem) até redes, armazenamento, bancos de dados e o que estiver em torno deles. Ou seja, tudo o que normalmente é necessário para montar o equivalente a uma sala de servidores, ou mesmo um data center completo, estará pronto para ser utilizado, configurado e executado.
  - Conhecer a diferença entre IaaS, PaaS e SaaS
  - Conhecer os maiores provedores de cloud
  - Especializar-se em algum provedor

- [ ] **Git e GitHub - Fundamentos**:

  - Git é um sistema de controle de versão distribuído gratuito e de código aberto projetado para lidar com tudo, desde projetos pequenos a muito grandes com velocidade e eficiência.
  - GitHub é um serviço de hospedagem para desenvolvimento de software e controle de versão usando Git.
  - Criar um repositório
  - Clonar um repositório
  - Fazer commit, push e pull de e para o repositório
  - Reverter um commit
  - Criar branches e pul requests
  - Lidar com merge e conflitos

- [ ] **Linux - Fundamentos**:

  - Linux é um termo popularmente empregado para se referir a sistemas operacionais que utilizam o Kernel Linux. As distribuições incluem o Kernel

Linux, além de softwares de sistema e bibliotecas.

- Conhecer o sistema de diretórios do Linux
- Compactar e descompactar arquivos
- Editar arquivos no console com o VI
- Gerenciar os processos rodando na máquina
- Conhecer as variáveis de ambiente e o PATH
- Gerenciar pacotes
- Realizar comunicação remota com o SSH e SCP

Habilidade Auxiliar: Business

- [ ] **Gestão de Processos de Negócios**:

- A Gestão de Processos de Negócios (BPM) é a um disciplina que utiliza vários métodos para descobrir, modelar, analisar, medir, melhorar, otimizar e automatizar processos de negócios.

- [ ] **Business Intelligence (BI) - Fundamentos**:

- Business Intelligence é um conjunto de teorias, metodologias, processos e tecnologias que possibilitam a transformação dos dados “crus” em informações extremamente relevantes para tomada de decisão de uma empresa.
- Conhecer o processo de ETL
- Realizar a modelagem e estruturação de tabelas em um Data Warehouse
- Criar visualizações que façam sentido
- Conhecer o PowerBI

- [ ] **Storytelling com dados**:

- Storytelling é uma forma de contar histórias que engajam e chamam a atenção da pessoa que está ouvindo. Dentro da análise de dados, é algo muito importante para passar as informações ao receptor de modo que o mesmo compreenda não apenas os dados, mas também todo o contexto.

- [ ] **Excel**:
    - O Microsoft Excel é um editor de planilhas produzido pela Microsoft com ferramentas de cálculo e de construção de tabelas.
    - Realizar as operações matemáticas básicas com seus operadores (soma, subtração, multiplicação e divisão)
    - Conhecer as principais fórmulas, como 'MÉDIA' (AVERAGE), 'ARRED' (ROUND), 'MÁXIMO' (MAX), 'MÍNIMO' (MIN), etc
    - Realizar buscas em colunas com a função 'PROCV'
    - Criar gráficos
- [ ] **Habilidades de comunicação**:
    - Um bom nível de comunicação facilita o atingimento de objetivos, resolução de problemas, além de aumentar a produtividade, porque cada profissional saberá exatamente o que se espera dele e transmitir com clareza suas ideias.
- [ ] **Inglês técnico**:
    - Um bom nível de inglês técnico facilita o atingimento de objetivos, resolução de problemas, além de aumentar a produtividade, e também permite o consumo de materiais de diferentes fontes, principalmente de documentações oficiais.
- [ ] **Governança de Dados**:
    - Governança de dados é um conjunto de políticas, processos e práticas que estabelecem a responsabilidade, a integridade, a qualidade e o uso adequado dos dados em uma organização. Ela envolve a definição de regras e padrões para a coleta, armazenamento, gerenciamento, compartilhamento e uso dos dados, garantindo que sejam confiáveis, consistentes e seguros ao longo de sua vida útil.
    - A governança de dados desempenha um papel fundamental na engenharia de dados, pois garante que os dados sejam gerenciados de forma eficiente, confiável e em conformidade com as regulamentações e políticas internas. Ela estabelece diretrizes para a gestão dos dados, incluindo a definição de

metadados, a identificação de proprietários de dados, a documentação de políticas de acesso e privacidade, a implementação de medidas de segurança e a garantia da qualidade dos dados.

- A governança de dados também promove a colaboração entre as equipes, estabelecendo processos de tomada de decisão baseados em dados confiáveis e padronizados. Isso resulta em uma melhor qualidade das análises, uma base sólida para tomada de decisões estratégicas e uma maior confiança nos dados utilizados para impulsionar as iniciativas de engenharia de dados.

☐ **Proteção de dados**:

- A proteção de dados é o processo de proteger informações importantes de forma que garanta a confidencialidade, integridade e a disponibilidade destes dados.
