DADOS, PYTHON E ESTATÍSTICA: CAMINHOS PARA STEM

Ana Júlia Amaro Pereira Rocha

OBJETIVOS

O projeto visa apresentar conceitos introdutórios de STEM com o intuito de descobrir novos talentos, principalmente femininos, para a área.
Também deseja-se apoiar jovens de Santa Maria do Suaçuí na escolha de sua carreira com um conhecimento geral de profissões que são menos comuns na região, mostrando para eles que há mais oportunidades do que as que já conhecem.

CRONOGRAMA GERAL SEMANAL

S	T	Q	Q	S
Dados Papel e caneta	**Python** Computador	**Python** Computador	**Estatística** Papel e caneta	**Apresentação** Apresentar a tarefa da semana

SOBRE MIM

- Graduanda do 7º período em Ciência de Dados e Inteligência Artificial na Escola de Matemática Aplicada da Fundação Getúlio Vargas (FGV-EMAp);
- ex-bolsista do Centro para o Desenvolvimento da Matemática e Ciências (FGV-CDMC);
- trimedalhista da Olimpíada Brasileira de Matemática das Escolas Públicas (OBMEP) com 2 pratas, 1 bronze e 2 menções honrosas;
- Bolsista de Iniciação Científica do projeto “Interdisciplinaridade: um superpoder para meninas e mulheres combaterem desigualdades nas ciências exatas” desenvolvido por pesquisadores da Fiocruz e financiado pela FAPERJ;
- Dev e cientista de dados do projeto de pesquisa Mosqlimate.

QUEBRANDO O GELO

- Qual é o seu nome?
- Quantos anos você tem?
- Alguma coisa legal que viveu na última semana;
- Algum desafio que precisa ser superado;
- Por que você quis participar desse projeto?

SEMANA 1

DIA 1

DADOS

Definição formal: Dados são informações brutas, não interpretadas, que podem ser representadas por números, textos, imagens, sons ou qualquer outro tipo de registro.

Os dados, por si só, podem não ter um significado claro ou contexto, mas quando organizados, processados e analisados, se tornam informações valiosas, sendo a base para análises, conclusões e tomada de decisões.

Exemplos: Dados podem ser a quantidade de horas que você dorme, a temperatura de um ambiente, a quantidade de pessoas com dengue em uma região, o nome de um produto, entre outros.

TIPOS DE DADOS

Dados estruturados: obedecem a um esquema rígido, com modelo bem definido de armazenamento (ex: tabelas com colunas fixas).

Dados semi-estruturados: não seguem totalmente a rigidez dos dados estruturados, mas têm metadados ou marcações internas que permitem identificar estrutura. Ex: e-mails (cabeçalho estruturado + corpo livre).

Dados não estruturados: Dados sem esquema predefinido ou estrutura rígida de organização. Dificilmente "cabem" em linhas/colunas de forma imediata. Ex: PDFs, imagens, áudios, vídeos, posts de redes sociais etc.

UM POUCO DE HISTÓRIA

- Até os anos 1980-1990, o maior desafio era coletar dados;
- A maioria dos registros era manual (em papel, formulários, fichas), com armazenamento físico limitado.
- Computadores estavam presentes, mas caros e restritos a empresas grandes e órgãos governamentais.
- A análise de dados era feita em pequenos bancos relacionais ou planilhas, o que limitava o volume e a variedade da informação.
- **Em resumo:** faltavam dados digitalizados e havia poucos recursos para coleta em massa.

UM POUCO DE HISTÓRIA

- A popularização da internet, de dispositivos móveis, sensores, câmeras, redes sociais e sistemas corporativos aumentou drasticamente a geração de dados.
- Empresas passaram a registrar tudo digitalmente: transações, logs de sistemas, comportamento de usuários.
- O problema deixou de ser “não ter dados” para se tornar **como armazenar e processar o excesso de dados.**
- Vivemos em um cenário de **superabundância de dados.**

UM POUCO DE HISTÓRIA

- O conceito de **Big Data** surgiu para caracterizar esse fenômeno, geralmente explicado pelos 3 Vs (e depois ampliado para 5 ou mais):
  - **Volume:** quantidades massivas de dados gerados por segundo.
  - **Velocidade:** dados chegando em tempo real (ex.: redes sociais).
  - **Variedade:** dados estruturados, semi-estruturados e não estruturados (texto, vídeo, sensores etc.).
  - (+ **Veracidade** e **Valor**, em algumas definições).
- Agora, o desafio não é mais ter dados, mas **filtrar, organizar e extrair valor** deles – daí o crescimento de áreas como ciência de dados, machine learning e inteligência artificial.

COLETANDO DADOS

A coleta de dados é o momento em que registramos informações para depois analisar. **Ex:** Ficha em loja, cadastro em rede social, lista de chamada na escola etc.

Por que é tão importante essa etapa?

- As lojas usam dados do cadastro para entender seus clientes, direcionando de forma eficaz suas promoções, bem como as redes sociais e seus anúncios;
- Lista de chamada na escola ajuda a entender a frequência dos alunos;
- Pesquisas do IBGE coletam dados de milhões de pessoas e servem para decidir políticas públicas como onde construir escolas e hospitais;
- **Se os dados não forem verdadeiros não retratarão a realidade.**

ARMAZENANDO DADOS

Banco de dados: Conjunto estruturado de dados armazenados em tabelas com linhas (registros) e colunas (atributos). Usado para armazenar e gerenciar dados do dia a dia, geralmente em tempo real.

Analogia: É como uma agenda telefônica ou uma geladeira organizada onde cada item tem um lugar definido, fácil de acessar e consultar.

ARMAZENANDO DADOS

Data Warehouse: Repositório centralizado que integra dados de várias fontes, organizados e tratados para análises históricas e estratégicas. É otimizado para consultas complexas e relatórios.

Analogia: É como um arquivo de escola ou uma despensa organizada em que tudo já está limpo, catalogado e separado, pronto para usar em planejamentos maiores como comparar anos diferentes.

ARMAZENANDO DADOS

Data Lake: Repositório que armazena dados em grande volume e em diferentes formatos (estruturados, semiestruturados e não estruturados), sem necessidade de pré-processamento.

Analogia: É como uma caixa gigante onde você joga todos os cadernos, provas, bilhetes e até vídeos da sala (está tudo lá, mas ainda sem organização; pode ser útil no futuro).

FORMATOS DE DATASETS

Dataset é um conjunto de dados organizados que podemos analisar.
Ex: uma planilha com notas dos alunos (linhas representam os alunos e colunas as matérias). Os formatos mias comuns de datasets são CSV, XLSX, JSON e SQL:

CSV: Arquivo de texto simples onde os valores são separados por vírgula (ou ponto e vírgula). Muito usado por ser leve e fácil de abrir até no Excel.

EX:
NOME,IDADE,CIDADE
ANA,16,BELO HORIZONTE
MARIA,17,OURO PRETO
JÚLIA,15,ITABIRA

FORMATOS DE DATASETS

XLSX: Arquivo de planilha do Excel. Mais “rico” que o CSV porque pode ter cores, fórmulas e várias abas.

SQL: (Structured Query Language) são dados estruturados dentro de um banco de dados relacional, acessados via consultas.

JSON: (JavaScript Object Notation) é uma estrutura em formato de dicionário/chave-valor. Muito usado em dados da web.

EX: {"NOME": "ANA", "IDADE": 16, "CIDADE": "BH"}

DATASETS PÚBLICOS

Existem milhares de conjuntos de dados disponíveis gratuitamente. Os datasets públicos são como bibliotecas abertas, qualquer pessoa pode entrar, escolher um tema que gosta e começar a estudar.

- **Kaggle:** Plataforma mundial com datasets de todo tipo: saúde, esportes, filmes, economia etc.
- **IBGE:** Dados brasileiros sobre população, renda, educação, trabalho...
- **DATASUS:** Dados do sistema público de saúde brasileiro (internações vacinação, mortalidade etc.).