Etapas Principais Para implementar um projeto De ciência de dados

1. Definição do problema

- Entender o objetivo do negócio;
- Definição dos objetivos do projeto;
- Identificação das Perguntas De pesquisa;
- Análise dos DADOS Disponíveis;
- Definição das métricas de sucesso;
- Planejamento do Projeto;

2. Coleta e armazenamento De DADOS

- Dados Podem ser coletados Por: Pesquisas, Entrevistas, observações, sensores, registros de transações, APIs, Bancos de dados públicos.
- Garantir a qualidade dos dados, Técnicas De limpeza e validação dos Dados ajudam a manter a integridade dos dados.
- O armazenamento dos Dados deve ser eficiente e seguro. Bancos de Dados relacionais, NoSql, Data lakes, DATA Lakehouses e Data warehouses são opções.

Data Lakes = É um repositório centralizado que Permite armazenar DADOS estruturados e não estruturados. Podemos realizar qualquer Tipo de análise sem a necessidade De estruturar os Dados primeiramente, permite armazenar um grande Volume de Dados.
→ estão relacionados

Data warehouses = É um repositório centralizado que só armazena Dados estruturados e anteriormente pré-processados, armazena menos Tipos de Dados, porisso possui uma performance de pesquisa mais rápida.

Data Lakehouses = É um novo conceito que une a flexibilidade, economia e escalabilidade de um DataLake, e o gerenciamento de dados e os recursos de Transações ACID de uma Data warehouse

Transações
- Atomicity
- Consistency
- Isolation
- Durability

} Um sistema que aplica essas operações é chamado sistema Transacional.

- Proteção dos dados coletados. Implementar medidas de segurança, como criptografia, controle de Acesso e políticas de privacidade.

3. Preparação e limpeza dos Dados

- A preparação De dados é o processo de transformar dados brutos em formato Adequado para análise e modelagem;
- Etapas principais:

Limpeza: Envolve a remoção De valores nulos duplicados e inconsistentes;

Normalização: Padronização dos dados para manter a consistência;

Transformação: Conversão De dados em formatos apropriados para análise. Isso pode envolver agregação, codificação e criação De novas variáveis;

Integração: combina Dados De diferentes fontes para oferecer uma visão unificada. Técnicas de ETL são frequentemente utilizadas para consolidar Dados de diversos sistemas em um repositório central;
↳ ETL = Extração, Transformação e carga.

Codificação: Processo de transformar variáveis de Tipo Texto em representação numérica. modificamos os Dados sem modificar a informação.
Ex. Variável indicando se um cliente vai ou não fazer uma compra com valores sim/não. codificamos para 1/0 a fim De Treinar um modelo de Machine Learn.

Redução: simplifica conjuntos De Dados grandes e complexos, mantendo as informações mais relevantes. Metodos como seleção De características e PCA ajudam a reduzir a complexidade e melhorar a eficiência da análise.
↳ PCA: Análise De componentes Principais.

4. Análise Exploratória Dos Dados

A exploração e visualização De dados, conhecida como análise exploratória Dos Dados (EDA), é uma etapa crítica no fluxo de Trabalho de ciência de dados;

EDA: envolve análise inicial dos dados para entender suas principais características, padrões e anomalias.

- Esse processo utiliza técnicas estatísticas descritivas e ferramentas de visualização para resumir as distribuições dos dados, identificar relações entre variáveis e detectar valores discrepantes (outliers);

- O objetivo é explorar os dados para compreender seus padrões e detectar eventuais problemas. Também podemos aplicar engenharia de atributos durante ou logo após a EDA.

5. Modelagem preditiva/estatística

Modelagem Preditiva != Modelagem estatística

- Interesse em utilizar as variáveis para fazer previsões;

Abordagem:
- Inclui uma ampla gama de técnicas, como aprendizado de máquina supervisionado, não supervisionado, etc. Cada algoritmo pode ter suas próprias suposições que devem ser validadas. Focamos na performance preditiva;

Interpretação:
- Resultados mais difíceis de interpretar, especialmente modelos vindos de redes neurais. Foca na precisão das previsões;

Exemplos de uso:
- Usada em negócios para prever vendas futuras, compras de clientes e detecção de fraude. Aplicações em engenharia para prever falha de máquinas.

Ferramentas e técnicas:

Ferramentas: Python com scikit-learn, pytorch, tensorFlow, Linguagem R com caret, randomForest, julia, Rust, c++, java, java script.
Técnicas: Árvores de decisão, Floresta aleatória, Redes neurais, Deep learning, Svm, Boosting, Bagging, Regressão linear, Regressão logística;

- Interesse na análise e explicação da relação entre variáveis;

Abordagem:
- Utiliza técnicas como regressão linear, ANOVA, testes de hipóteses, etc. Geralmente assume que os dados seguem certas distribuições e que há relações lineares entre variáveis. Métodos: Paramétricos (suposições fortes), Não Paramétricos (suposições fracas).

Interpretação:
- Resultados interpretáveis e que podem ser usados para inferir relações entre as variáveis. Coeficientes com significados claros, intervalos de confiança, valores p (Resultados amplamente usados em modelagem estatística);

Exemplos de uso:
- Usada para testar hipóteses, como análise de pesquisa de mercado;

Ferramentas e técnicas:

Ferramentas: Linguagem R, stata, SAS, SPSS, Python com statsmodels.
Técnicas: Regressão linear, Regressão logística, ANOVA, Análise de sobrevivência, análise Fatorial, métodos probabilísticos, etc;

6. Avaliação e Teste

Etapa responsável por avaliar e testar as soluções criadas durante o projeto de data science.

- Métricas de avaliação;
- Validação cruzada;
- Overfitting e Underfitting;
- Curvas de Aprendizado;
- Análise de erros (Resíduos);
- Benchmarking;

7. Entrega de Resultados

A entrega de um projeto de ciência de dados dependendo dos objetivos, público-alvo e contexto do negócio:

- Relatório técnico ou científico;
- Relatório executivo;
- Dashboard interativo ou infográfico;
- Jupyter Notebook;
- código fonte e documentação;
- Api;
- Aplicação web para Deploy do modelo ML;
- Previsões do modelo ML em arquivo csv;
- Previsões do modelo ML em um banco de dados;
- Podemos também só entregar o arquivo do modelo ML.