**Gabriel Henrique Alves Spacki** **20/12/2023**

## Lidando com Dados do Mundo Real

O curso Machine Learning, Data Science and Generative AI with Python tem se mostrado cada vez mais intensivo em questões de teoria e código. Dessa vez, abordando as principais técnicas para lidar com dados do mundo real, enquanto visitamos famosos algoritmos de aprendizado de máquina.

## Sessão 6: More Data Mining and Machine Learning Techniques

### K-Nearest-Neighbors: Conceitos

O algoritmo K-Nearest-Neighbors (KNN) é uma técnica de aprendizado de máquina que classifica um ponto de dados com base na maioria dos k vizinhos mais próximos. A proximidade é medida por uma métrica como distância euclidiana. O KNN é simples de entender e implementar, mas sua eficácia depende da escolha adequada de k e da qualidade dos dados. É uma abordagem não paramétrica, o que significa que não faz suposições sobre a distribuição dos dados.

### Redução de Dimensionalidade: Análise de Componentes Principais (PCA)

A Redução de Dimensionalidade é crucial para lidar com conjuntos de dados de alta dimensionalidade. A Análise de Componentes Principais (PCA) é uma técnica comum que transforma variáveis correlacionadas em um conjunto menor de variáveis não correlacionadas chamadas de componentes principais. Isso facilita a visualização e o processamento de dados, preservando a maior parte da variabilidade original. PCA é amplamente utilizado em várias áreas, incluindo reconhecimento de padrões e compressão de dados.

### Visão Geral do Armazenamento de Dados: ETL e ELT

No contexto do armazenamento de dados, a extração, transformação e carga (ETL) e a extração, carga e transformação (ELT) são processos cruciais. Eles envolvem a coleta de dados de diversas fontes, a transformação para um formato adequado e a carga eficiente em um data warehouse. ETL é tradicionalmente usado para consolidar dados antes do armazenamento, enquanto o ELT envolve a transferência direta de dados para o armazém antes da transformação. Ambos são essenciais para a integridade e qualidade dos dados.

### Aprendizado por Reforço

O Aprendizado por Reforço é um paradigma em aprendizado de máquina onde um agente aprende a realizar ações em um ambiente para maximizar uma recompensa acumulada ao longo do tempo. Esse tipo de aprendizado é frequentemente aplicado em ambientes dinâmicos e interativos. O agente toma decisões exploratórias para aprender padrões e estratégias ótimas, sendo amplamente utilizado em jogos, robótica e sistemas de controle.

## Q-Learning

Q-Learning é uma técnica específica de aprendizado por reforço que utiliza uma tabela Q para avaliar a qualidade de ações em um determinado estado. O agente aprende iterativamente a melhor política através da atualização da tabela Q com base nas recompensas recebidas. Q-Learning é eficaz para resolver problemas complexos de decisão sequencial e tem aplicações em jogos, navegação autônoma e otimização de sistemas.

## Compreensão de uma Matriz de Confusão

A Matriz de Confusão é uma ferramenta fundamental na avaliação de modelos de classificação. Ela exibe a relação entre as previsões do modelo e as classes reais do conjunto de dados. Os elementos da matriz incluem verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. Essa análise é crucial para entender o desempenho do modelo e ajustar parâmetros para melhorar a precisão.

## Medindo Classificadores (Precisão, Recall, F1, ROC, AUC)

A avaliação de classificadores envolve diversas métricas. A precisão mede a proporção de instâncias corretamente classificadas. O recall avalia a capacidade do modelo de capturar todas as instâncias relevantes. A pontuação F1 combina precisão e recall. A curva ROC e a área sob a curva (AUC) são úteis para avaliar o desempenho em diferentes limiares de classificação, oferecendo uma visão abrangente do desempenho do modelo em diversas condições. Essas métricas são essenciais para selecionar e ajustar modelos de classificação de forma eficaz.

# Sessão 7: Dealing with Real-World Data

## Trade-off Viés/Variância

O trade-off entre viés e variância é um desafio central no desenvolvimento de modelos de machine learning. Um modelo com alto viés simplifica demais a representação do problema, resultando em subajuste. Em contrapartida, um modelo com alta variância se ajusta demais aos dados de treinamento, falhando em generalizar para novos dados. Encontrar o equilíbrio adequado é essencial para criar modelos que sejam precisos e generalizáveis.

## Limpeza e Normalização de Dados

A limpeza de dados envolve a identificação e correção de erros, valores ausentes e inconsistências em conjuntos de dados. A normalização é o processo de ajustar os valores das variáveis para uma escala padrão. Ambas as práticas são vitais para garantir a qualidade dos dados e melhorar o desempenho dos modelos de machine learning, evitando distorções causadas por dados ruidosos ou inconsistentes.

**Normalização de Dados Numéricos**

Normalizar dados numéricos é um passo importante no pré-processamento de dados. Esse processo ajusta as escalas das variáveis numéricas, garantindo que todas contribuam de maneira equitativa para o modelo. Isso é especialmente relevante em algoritmos sensíveis à escala, como k-Nearest Neighbors, onde variáveis com escalas diferentes podem impactar negativamente o desempenho do modelo.

**Engenharia de Recursos e o Problema da Dimensionalidade**

A engenharia de recursos é o processo de criar novas variáveis ou modificar as existentes para melhorar o desempenho do modelo. No entanto, o aumento no número de variáveis pode levar ao "problema da dimensionalidade", onde a complexidade do modelo aumenta significativamente. Gerenciar esse dilema é crucial, pois muitas variáveis podem resultar em sobreajuste, prejudicando a generalização do modelo para novos dados.

**Técnicas de Imputação para Dados Ausentes**

Dados ausentes são comuns em conjuntos de dados do mundo real e exigem técnicas de imputação para estimar valores faltantes. Métodos como imputação média, mediana, regressão e técnicas mais avançadas, como MICE (Multiple Imputation by Chained Equations), são empregados para preservar a integridade dos dados durante o processo de análise.

**Tratamento de Dados Desbalanceados: Oversampling, Undersampling e SMOTE**

Desbalanceamento de dados, onde uma classe é representada em menor quantidade, pode prejudicar a capacidade do modelo de aprender padrões. Estratégias como oversampling (aumentar a amostra da classe minoritária), undersampling (reduzir a amostra da classe majoritária) e SMOTE (Synthetic Minority Over-sampling Technique) são aplicadas para equilibrar as classes, melhorando a capacidade do modelo de lidar com dados desbalanceados.

**Binning, Transformação, Codificação, Escalonamento e Embaralhamento**

Essas práticas referem-se a diferentes técnicas de pré-processamento de dados. Binning agrupa valores contínuos em intervalos discretos, transformações alteram a distribuição dos dados, codificação converte variáveis categóricas em formatos adequados para modelos, escalonamento ajusta a escala das variáveis e o embaralhamento reorganiza a ordem dos dados. Cada técnica desempenha um papel único no preparo dos dados para análise e modelagem.

## Conclusão

Ao explorar desde fundamentos como o algoritmo K-Nearest-Neighbors até estratégias avançadas como o tratamento de dados desbalanceados, adquirimos habilidades essenciais para lidar com a complexidade e a diversidade de conjuntos de dados. O entendimento profundo de técnicas como Análise de Componentes Principais, Aprendizado por Reforço e Q-Learning, aliado à avaliação criteriosa por meio de métricas como Matriz de Confusão e Precisão, posiciona o desenvolvimento de modelos de machine learning robustos e eficazes.