

DESAFIO: ANALISANDO DADOS DE PREÇOS DE CASAS NOS ESTADOS UNIDOS

Prof. Howard Roatti

O desafio proposto tem como objetivo que alunos tenham uma experiência prática com diferentes etapas do processo de análise de dados e construção de modelos de aprendizagem de máquina. Para isso, eles deverão seguir as seguintes tarefas:

1. **Análise exploratória de dados:** O aluno deverá realizar uma análise inicial dos dados do dataset escolhido, identificando quais variáveis estão presentes, quais são numéricas/categóricas e se existem valores faltantes ou outliers.
2. **Feature engineering:** Com base na análise exploratória, o aluno deverá realizar transformações nas variáveis do dataset, como normalização, codificação de variáveis categóricas, criação de novas features, dentre outras técnicas.
3. **Aprendizagem supervisionada:** O aluno deverá escolher um dos modelos de aprendizagem supervisionada (Regressão Linear, Naive Bayes, Regressão Logística, KNN, Árvore de Decisão, Random Forest ou XGBoost), realizar o treinamento do modelo com o dataset e avaliar seu desempenho com métricas adequadas.
4. **Aprendizagem não supervisionada:** O aluno deverá escolher uma técnica de aprendizagem não supervisionada (Clusterização, Redução de Dimensionalidade, Análise de Associação, Análise de Outlier ou Visualização de Dados), aplicá-la ao dataset e interpretar seus resultados.
5. **Métricas de avaliação e comparação:** O aluno deverá utilizar as métricas adequadas para avaliar o desempenho dos modelos de aprendizagem supervisionada, comparando-os entre si e com o modelo não supervisionado escolhido.
6. **Dataset público do Kaggle ou UCI:** O aluno deverá utilizar o dataset público selecionado para realizar todas as tarefas propostas acima.

Ao final do desafio, o aluno terá desenvolvido habilidades em análise exploratória de dados, feature engineering, construção e avaliação de modelos de aprendizagem supervisionada e não supervisionada, além de ter tido contato com um dataset real.

Objetivo: O objetivo deste desafio é explorar um conjunto de dados de preços de casas nos Estados Unidos e criar um modelo de regressão para prever o preço de venda de uma casa.

Dados: O conjunto de dados está disponível no Kaggle e contém informações sobre preços de casas em diferentes bairros dos Estados Unidos, bem como características das casas, como número de quartos, banheiros, tamanho do terreno, etc.

O conjunto de dados de preços de casas nos Estados Unidos mencionado está disponível no Kaggle. Você pode encontrá-lo no seguinte link: <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>

Tarefas:

1. **Análise exploratória de dados:** Faça uma análise exploratória dos dados para entender a distribuição das variáveis e identificar possíveis correlações entre elas.
2. **Featuring Engineering:** Realize a engenharia de características para selecionar as variáveis mais importantes para o modelo de regressão.
3. **Aprendizagem Supervisionada:**
 - a. Regressão Linear: Crie um modelo de regressão linear simples ou múltipla para prever o preço de venda de uma casa.
 - b. Classificação: Converta a variável de saída em uma variável binária e crie um modelo de classificação para prever se uma casa será vendida por um preço alto ou baixo.
4. **Aprendizagem não supervisionada:**
 - a. Clusterização: Use um algoritmo de clusterização para identificar grupos de casas com características semelhantes.
 - b. Redução de dimensionalidade: Use uma técnica de redução de dimensionalidade para visualizar os dados em um espaço de menor dimensão.
 - c. Análise de associação: Use o algoritmo Apriori para identificar associações entre as características das casas.
 - d. Análise de outlier: Use o algoritmo Local Outlier Factor para identificar casas que podem ser consideradas outliers.
5. **Métricas de avaliação e comparação:** Use diferentes métricas para avaliar o desempenho dos modelos criados e compará-los.
6. **Dataset:** Use o conjunto de dados de preços de casas nos Estados Unidos disponível no Kaggle ou UCI.

Distribuição dos Pontos - 8,0 (oito) pontos:

1. **Análise exploratória de dados e feature engineering (1 ponto)**
2. **Aprendizagem supervisionada:** treinamento e avaliação de modelos de regressão (1 ponto)
3. **Aprendizagem supervisionada:** treinamento e avaliação de modelos de classificação (1 ponto)
4. **Aprendizagem não supervisionada:** aplicação de técnicas de clusterização (1 ponto)
5. **Aprendizagem não supervisionada:** aplicação de técnicas de redução de dimensionalidade (1 ponto)
6. **Aprendizagem não supervisionada:** aplicação de técnicas de análise de associação e outlier (1 ponto)
7. **Visualização de dados:** utilização de técnicas de visualização para apresentação dos resultados obtidos (1 ponto)
8. **Organização do Repositório do Github: 0,5 ponto**
9. **Apresentação do Trabalho: 0,5 ponto**

Entrega: A entrega deverá ser feita através do Github, apenas o envio do link do repositório por um dos membros do grupo deverá ser feito via AVA.

Apresentação: A apresentação do trabalho deverá ser feita no dia combinado (**26/03** 5HC/5SC e **27/03** 4HC/4SC) através do Jupyter Notebook.