Trabalho 01 - Pré-processamento de Dados, Redução de Dimensionalidade e Aprendizado Supervisionado

Descrição do Trabalho: O trabalho poderá ser realizado em grupo de até 04 (quatro) pessoas e deverá ser desenvolvido em Python, utilizando a plataforma Google Colab. Cada grupo deverá elaborar um relatório detalhado explicando o processo de resolução do exercício, além de incluir o código-fonte utilizado.

Ambos os arquivos, o relatório e o código, deverão ser organizados e disponibilizados em um repositório no GitHub. O link para o repositório deverá ser enviado por e-mail para a professora, seguindo o formato de assunto: **IA - Trabalho 01 - NOME DO INTEGRANTES DO GRUPO.**

Instruções de envio:

- Relatório: Apresentar a explicação detalhada do desenvolvimento e da solução proposta.
- Código-fonte: O código deverá ser bem estruturado e comentado.
- GitHub: Criar um repositório público ou privado (com acesso concedido à professora) contendo os arquivos.

Prazo: O link contendo o repositório deve ser enviado até **03/11/2024**.

Desconto por atraso: Para cada dia de atraso no envio do trabalho, será descontado **1 ponto** da nota máxima.

Plágio e similaridade: Caso seja detectado **plágio** ou um **alto índice de similaridade** com outro material, o trabalho poderá ser **anulado**, resultando em nota zero.

Apresentação presencial: O grupo poderá ser convocado a apresentar o trabalho presencialmente caso a professora tenha dúvidas sobre o conteúdo ou sua autoria.

1. Para a base de dados Communities and Crime (disponível em https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime) faça:
    a. Faça as análises e alterações necessárias na base de dados para predizer a variável ViolentCrimesPerPop usando regressão linear. Observe que essa base de dados possui valores faltantes. Explique as considerações e mudanças propostas.
    b. Divida aleatoriamente a base de dados em duas partes: treino, com 70% das amostras, e teste, com 30%. Use a parte de treino para estimar um modelo linear que melhor se ajuste aos dados. Obtenha os valores de RMSE e MAE sobre o conjunto de treino e teste.
    c. Aplique PCA sobre os dados de treino para reduzir os dados para 5 atributos. Realize análise gráfica sobre as variáveis e proponha alterações para melhorar o modelo de regressão linear (que poderá ser um modelo polinomial). Com esses atributos, obtenha os valores de RMSE e MAE sobre o conjunto de treino e teste. Compare com os resultados da letra b).

2. Realize a classificação da base de dados HTRU2 (disponível em https://archive.ics.uci.edu/ml/datasets/HTRU2) usando o esquema de validação holdout. Para cada execução, use 6000 amostras de treino selecionadas aleatoriamente e o restante para teste (normalmente o conjunto de treinamento é maior do que de teste, mas para reduzir o

custo computacional ele foi reduzido aqui). Execute 5 vezes o treinamento e teste e retorne a acurácia, recall e precisão média para cada algoritmo. Faça a classificação usando:

   a. kNN com métrica de distância Euclidiana. Escolha 5 valores diferentes de k. Para selecionar o melhor valor de k divida a base de treinamento em duas partes iguais: uma para treinar e a outra para validar e encontrar o melhor valor de k;
   b. Compare os resultados, tempos de execução e número de protótipos usados por cada valor de k da letra a). Considerando a distribuição das classes, você considera o valor da acurácia média relevante? Por quê?

3. Para a base Nursery (disponível em https://archive.ics.uci.edu/dataset/76/nursery):
   a. Construa uma árvore de decisão com dois níveis de nó de decisão (isto é, o primeiro nó de decisão (primeiro nível), os nós de decisão abaixo dele (segundo nível) e em seguida os nós folha) usando a medida de Ganho de Informação. Selecione aleatoriamente 10000 amostras dos dados para treinamento que serão usados para construir a árvore. Retorne a estrutura da árvore construída.
   b. Use os restantes dos dados para avaliação. Retorne a acurácia obtida.
   c. Tente obter as regras de decisão a partir da árvore construída.

4. Explique o dilema entre bias e variância e o seu relacionamento com underfitting e overfitting.

5. Comente sobre a veracidade das afirmações:
   a. “Quanto mais variáveis de entrada forem usadas em um modelo de aprendizado de máquina, melhor será a qualidade do modelo”.
   b. “Independente da qualidade, quanto mais amostras forem obtidas para uma base de dados, maior a tendência de se obter modelos mais adequados”.
   c. “Às vezes com simples manipulações na base de dados (limpeza, conversão de valores, etc.) pode-se conseguir melhoras significativas nos resultados, sem fazer nenhuma alteração na técnica de aprendizado de máquina usada”.

6. Em uma empresa é adotado um método de Aprendizado de Máquina para detectar defeito de fabricação de peças mecânicas, sendo que raramente acontece este tipo de problema na fábrica. Um funcionário anuncia empolgado que o sistema alcançou uma acurácia de 99%, porém seu gerente não achou o resultado tão relevante. Responda:
   a. Por que o gerente não ficou empolgado com o resultado achado?
   b. O que o funcionário poderia fazer para confirmar se o método empregado é adequado para o problema?

7. Como pode ser usada uma árvore (de regressão ou de decisão) para avaliar uma amostra quando ela possui uma ou mais variáveis faltantes?