ESPECIFICAÇÃO DO SEGUNDO TRABALHO DE INTELIGÊNCIA ARTIFICIAL

Data de envio: 20/11/2023

O trabalho consistirá na aplicação das duas técnicas de aprendizagem de máquina vistas em sala de aula (Árvores de Decisão e KNN) na classificação de uma base de dados de acesso público.

A base a ser investigada será a **Íris**, disponibilizada do repositório de dados UCI (Iris - UCI Machine Learning Repository). O conjunto de dados Iris é um conjunto de dados introduzido pelo estatístico e biólogo britânico Ronald Fisher em seu artigo de 1936. Posteriormente, Edgar Anderson coletou os dados para quantificar a variação morfológica das flores da íris de três espécies relacionadas. O conjunto de dados consiste em 50 amostras de cada uma das três espécies de Iris (Iris setosa, Iris virginica e Iris versicolor). Quatro variáveis foram medidas em cada amostra: o comprimento e a largura das sépalas e pétalas, em centímetros. Com base na combinação dessas quatro características, Fisher desenvolveu um modelo para distinguir as espécies umas das outras.

O objetivo é avaliar os métodos Árvores de Decisão (geração da árvore com a métrica entropia) e KNN para fazer a classificação desses dados. A qualidade de classificação cada método deve ser obtida, calculando-se as seguintes métricas:

- Acurácia

- Sensitividade

- Especificidade

- Precisão

Essas métricas devem ser calculadas em um procedimento com divisão Treinamento/Teste, explicado a seguir. Deve ser feita a divisão em treinamento e teste da seguinte maneira: dividam a base Iris em aproximadamente três partições com 1 terço dos registros: A, B e C. É importante que cada partição tenha aproximadamente a mesma proporção das 3 classes de flores (setosa, versicolor e virginica) que existe na base completa, para evitar tendências nos

dados. Posteriormente, as partições devem ser agrupadas em treinamento e teste resultando em 3 experimentos:

Primeiro: Treinamento (A+B) e Teste (C)

Segundo: Treinamento (A+C) e Teste (B)

Terceiro: Treinamento (C+B) e Teste (A)

1- Para cada experimento, as métricas (Acurácia, Sensitividade, Especificidade, Precisão) devem ser calculadas para a base de Teste. Isso deve ser feito para cada método de classificação investigado (KNN e Árvore).
2- Para cada experimento, deve ser apresentada a estrutura da árvore de decisão obtida durante o treinamento.
3- Ao final, deve ser apresentada o valor médio das métricas (considerando os 3 experimentos) para cada método de classificação.

Para a implementação do trabalho, vocês podem implementar os métodos (KNN e Árvores) em uma linguagem de programação qualquer (ex: C, Java, Python, etc), mas não é o mais recomendado. Existem várias ferramentas disponíveis para essas técnicas clássicas e algumas sugestões seguem abaixo:

Python – Scikit-learn: http://scikit-learn.org/

R Project: http://www.r-project.org

Weka: http://www.cs.waikato.ac.nz/ml/weka

Rapid Miner: http://rapidminer.com

Os trabalhos serão apresentados pelo grupo (duplas out trios) a partir do dia 21/11 em datas a serem disponibilizadas posteriormente. A princípio, irei considerar os mesmos grupos formados para o 1º trabalho. Caso alguma mudança queira ser feita pelos alunos (reagrupamentos novos) peço que me enviem por email até no máximo segunda (13/11) e aguardem meu retorno (por email) para a nova formação.