UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Disciplina: **Ciência de Dados**
Curso: **Sistemas de Informação**
Professora: **Elaine Ribeiro Faria**

Trabalho Final

Tema: Comparação de Classificadores

Instruções:
1- Crie um relatório a partir das solicitações descritas a seguir.
2- Adicione um cabeçalho com a seguinte informação
Nome e nro de matrícula dos estudantes
3- Gere um arquivo pdf
4- Envie pelo Microsoft Teams

Objetivo:
Aplicar, avaliar e interpretar o desempenho de algoritmos de classificação supervisionada em duas bases de dados reais distintas, considerando:

a. Pré-processamento detalhado dos dados
b. Escolha e justificativa dos algoritmos
c. Avaliação crítica dos resultados

1- Para esta atividade avaliativa as seguintes etapas deverão ser realizadas:

a. Escolha da base de dados

Escolher duas bases de dados públicas para usar na atividade. É importante citar a fonte da base de dados, incluindo o link para download. Também é possível usar uma base de dados dos próprios estudantes, desde que eles apresentem essa base antes para o professor validar.

b. Exploração dos dados

Descrever o entendimento obtido sobre os dados. Aqui é possível também desenhar gráficos que auxiliem no entendimento e exploração dos dados. Comente sobre quais são os atributos, tipo dos dados, se há valores ausentes, qual é o atributo classe, se é um problema desbalanceado ou não, etc.

c. Pré-processamento

Escolher e justificar quais etapas de pré-processamento serão feitas nos dados. Ex: normalização ou re-escala dos dados, substituição/remoção de valores ausentes, agregação de dados, amostragem, etc. Lembre-se as bases de dados deverão ser pré-processadas adequadamente de forma a permitir a execução de cada algoritmo.

d. Algoritmo de classificação

Usar três algoritmos de classificação nas bases de dados escolhidas, sendo que pelo menos um deles deve ser um algoritmo que não foi visto em sala de aula. Para esta etapa , os estudantes poderão utilizar uma ferramenta de mineração de dados (como por exemplo, o *Weka*), poderão utilizar implementações prontas dos algoritmos disponíveis na Internet, usar implementações disponíveis em pacotes nas diferentes linguagens de programação (ex: *scikit learn* no *Python*, pacotes da linguagem R, etc.), buscar por implementações dos algoritmos em repositórios de código, implementar os seus próprios algoritmos.

e. Avaliação dos algoritmos

Uma estratégia de divisão da base de dados em treino e teste deve ser escolhida. Além disso, pelo menos duas medidas de avaliação deverão ser usadas. A tarefa consiste em avaliar o desempenho dos três algoritmos de classificação em cada uma das bases de dados usando as medidas de avaliação escolhidas. Tente descrever qual algoritmo ficou melhor em cada base de dados. Veja também se teve um algoritmo que se comportou melhor nas duas bases. Aqui é importante testar diferentes parametrizações dos algoritmos. Pelos menos duas parametrizações para cada algoritmos devem ser testadas. Ex: dois valores diferentes de K para o KNN. Tente interpretar os resultados produzidos.

2- Um relatório deverá ser produzido contendo:
   a. Introdução e objetivos

   b. Descrição detalhada das bases de dados incluindo: nome da base de dados usada, o link de onde a base foi obtida, uma descrição da base que inclui: nro de instâncias, nro de atributos, tipo de dado de cada atributo e se possui valores ausentes.

   c. Metodologia, ou seja, quais foram os passos executados. Aqui é importante destacar:
      i. pré-processamentos que foram aplicados na base antes de executar cada um dos algoritmos de classificação;
      ii. nome da ferramenta usada para os experimentos, ou o nome da linguagem e pacote de dado usado ou o link da implementação obtida na Internet;
      iii. nome dos algoritmos utilizados e uma descrição do algoritmo, que não foi visto em sala de aula. Tente descrever os passos desse algoritmo;
      iv. nome da estratégia de divisão da base de dados em treino e teste, que foi adotada;
      v. nome e a fórmula das medidas de avaliação utilizadas;
      vi. parâmetros usados na configuração do algoritmo.

   d. Discussão de Resultados
   Uma discussão dos resultados obtidos realçando qual foi o melhor algoritmo para cada base de dados. É importante ter uma tabela que sumarize os resultados obtidos.

   e. Uso de LLMs
   Os estudantes deverão indicar se fizeram uso de LLMs no trabalho. Em caso afirmativo, deverão indicar qual(is) LLMs foram usadas, em qual parte do trabalho elas foram usadas e qual o prompt usado.

3- Regras
   a. Não serão aceitos trabalhos atrasados. Se o grupo não entregar o trabalho no dia combinado, ele receberá nota zero.
   b. Em caso de projetos copiados de colegas todos os envolvidos recebem nota zero. Lembre-se é muito improvável que haja trabalhos iguais, afinal há várias bases de dados e diferentes algoritmos de classificação. Se os alunos usarem códigos disponíveis na Internet, é preciso entendê-lo antes da apresentação. É preciso também citar a referência a eles.
   c. A professora em hipótese alguma verificará ou ajudará na construção do código fonte.
   d. A professora poderá tirar dúvidas conceituais sobre o trabalho em horário de aula ou horário extra-classe.
   e. A interpretação dos resultados e o entendimento dos algoritmos fazem parte da avaliação e devem ser realizados pelos estudantes.
   f. Caso a professora identifique que o grupo usou LLMs sem a devida citação, todos os envolvidos receberão nota zero.
   g. A nota da prova sobre o trabalho será um valor entre 0 e 1 que pondera a nota do trabalho. A prova sobre o trabalho será feita com consulta ao trabalho. A prova é individual e cada integrante do grupo deve ter a sua cópia impressa do trabalho.