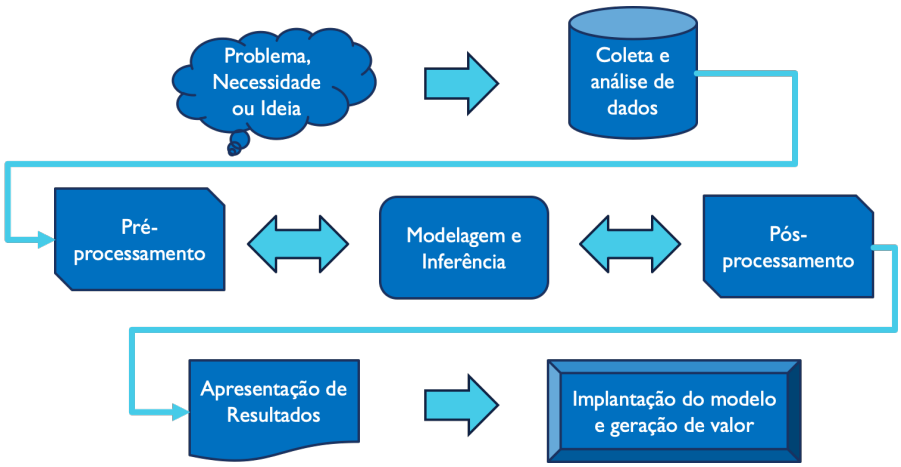
Especialização em Ciência de Dados – PUC-Rio
Disciplina: Introdução a Ciência de Dados
Professora: Tatiana Escovedo

Enunciado do Trabalho de Classificação no Weka

O trabalho deverá ser feito em grupos de 3 a 7 alunos e deverá ser entregue para a professora via Moodle e apresentado em aula, em data a combinar. Devem ser entregues na data estipulada 2 arquivos: a) um **relatório textual**, feito no Word, e b) uma **apresentação de slides**, feita no Power Point. O grupo deve escolher uma base de dados para um problema de classificação simples. Sugere-se usar uma das bases de dados do Weka (preferencialmente) ou uma base de dados disponibilizada no UCI Machine Learning Repository.

Você deverá trabalhar desde a definição do problema até a apresentação os resultados. Lembre-se das etapas para implementação de projetos de ciência de dados, apresentadas em sala de aula, e trabalhe com esta base de dados no ambiente Weka. Se desejar, use outras ferramentas auxiliares, como o Excel, para produzir gráficos e tratar dados. Finalize o relatório em formato PDF, descrevendo textualmente e analisando cada uma das etapas realizadas. Justifique todas as suas escolhas.

A apresentação deverá ser realizada até 15 minutos. A apresentação será avaliada pela professora e pelos alunos da turma nos quesitos i) Apresentador(es) (ritmo, cadência da apresentação, etc); ii) Apresentação (qualidade dos slides, conteúdo, fluidez, etc) e iii) Processo (etapas para levar do problema aos resultados, conceitos técnicos aplicados, etc).



Sugere-se a seguinte estrutura de tópicos para o **relatório**:

1. Apresentação do problema
2. Coleta e análise de dados
    a. Carga do dataset

    b. Análise do Dataset
        i. Estatísticas descritivas
        ii. Distribuições dos atributos
        iii. Análise de interações entre atributos
3. Pré-processamento
    a. Preparação de visões dos dados (apenas as operações realidas, tais como:)
        i. Normalização/Padronização
        ii. Tratamento de missings
        iii. Outras transformações e operaçoes realizadas
    b. Seleção de variáveis
4. Modelagem e inferência
    a. Aplicação dos algoritmos
    b. Variação dos hiperparâmetros
5. Pós-processamento
    a. Análise de resultados
6. Apresentação de resultados
    a. Resumo dos resultados
    b. Apresentação do modelo escolhido
7. Implantação do modelo
    a. Finalização do modelo
        i. Treinamento do modelo escolhido
        ii. Predição para novos dados
        iii. Conclusões finais

Sugere-se a seguinte estrutura de slides para a **apresentação**:

- Slide 1: Título e componentes do grupo
- Slide 2: Apresentação do problema
- Slide 3: Resumo das análises exploratórias realizadas
- Slide 4: Resumo das operações de tratamento de dados realizadas
- Slide 5: Modelos de classificação aplicados e variações de hiperparâmetros
- Slide 6: Análise dos resultados
- Slide 7: Implementação do modelo escolhido
- Slide 8: Conclusões

OBS: *O Checklist a seguir deverá servir apenas como* ***guia*** *para os experimentos. Espera-se que o trabalho seja escrito em formato de relatório formal, com capa, seções separadas, índice, figuras numeradas, etc.*

Checklist para o Trabalho

Definição do Problema

Objetivo: entender e descrever claramente o problema que está sendo resolvido.

Descrição do Problema:

- Qual é a descrição informal do problema?
- Qual é a descrição formal do problema?
- Que premissas ou hipóteses você tem sobre o problema?

Dados Disponíveis:

- Que restrições ou condições foram impostas para selecionar os dados?
- Defina cada um dos atributos conjunto de dados (dataset) disponibilizado.

Análise de Dados

Objetivo: entender a informação disponível que será usada para construir o modelo.

Estatísticas descritivas

- Quantos atributos e instâncias existem?
- Quais são os tipos de dados dos atributos?
- Verifique as primeiras linhas do dataset. Algo chama a atenção?
- Há valores faltantes ou inconsistentes?
- Faça um resumo estatístico dos atributos com valor numérico (mínimo, máximo, mediana, moda, média, desvio padrão e número de valores ausentes). O que você percebe?

Visualizações

- Verifique a distribuição de cada atributo (com histogramas). O que você percebe?
    - Pode dar ideias sobre a necessidade de transformações na etapa de preparação de dados (por exemplo, converter atributos de um tipo para outro, realizar operações de discretização, normalização, padronização, etc).
- Verifique a distribuição de frequência das classes. O que você percebe?
    - Pode indicar a possível necessidade de balanceamento de classes;
    - A distribuição de frequência do atributo de classe (ou média de uma variável de saída de regressão) é útil porque você pode usá-lo para definir a acurácia mínima de um modelo preditivo.
        i. Por exemplo, se houver um problema de classificação binária (2 classes) com a distribuição de 80% de maçãs e 20% de bananas, um modelo poderá prever "maçãs" para cada instância de teste e garantir uma precisão de 80%. Esse é o algoritmo de pior caso que todos os algoritmos no equipamento de teste devem vencer ao avaliar algoritmos.

- Verifique a distribuição dos atributos separados por classe em histogramas. O que você percebe?
- Verifique os atributos em pares, com scatter plots. O que você percebe?

Pré-Processamento de Dados

Objetivo: realizar operações de limpeza, tratamento e preparação dos dados para a etapa de modelagem.

- Verifique quais operações de pré-processamento podem ser interessantes para o seu problema e salve visões diferentes do seu dataset. Por exemplo:
  - Normalização
  - Padronização
  - Discretização
  - One-hot-encoding
- Trate (removendo ou substituindo) os valores faltantes (se existentes)
- Avalie os subconjuntos de atributos que podem ser mais interessantes para os modelos preditivos (use as técnicas demonstradas em aula).

→ Explique, passo a passo, as operações realizadas, justificando cada uma delas.
→ Volte na etapa de Análise Exploratória e verifique se surge algum *insight* diferente com as operações realizadas.

Modelagem e Inferência

Objetivo: realizar a avaliação de diversos algoritmos para o problema

- Experimente diversos algoritmos de classificação (Exemplos: regressão logística, Árvore de classificação, KNN, Naive Bayes e SVM) com a configuração padrão do Weka. Além do dataset original, utilize diferentes visões do dataset de acordo com as operações realizadas na etapa de pré-processamento de dados (normalização, padronização, seleção de variáveis, etc). Analise os resultados
- Experimente diferentes valores de parâmetros para os melhores algoritmos de classificação da etapa anterior. Analise os resultados.

Pós-Processamento

Objetivo: finalizar o modelo e prepara-lo para uso

- Escolha qual será o modelo final e os valores dos seus parâmetros, decidindo se irá utilizar dataset original ou uma das visões na etapa de pré-processamento. Justifique sua escolha.

- Separe 5 instâncias do dataset para serem usadas como dados novos, não vistos pelo modelo (sugestão: fazer no excel de forma aleatória, observando a proporção das classes).
- Treine o modelo escolhido com o dataset completo (excluindo as 5 linhas da etapa anterior).
- Execute o modelo para os dados não vistos. Analise o resultado.

Apresentação dos Resultados

Objetivo: realizar uma análise crítica do trabalho realizado

Escreva de forma resumida:

- Qual era o problema?
- Qual foi a solução proposta?
- Quais foram as principais descobertas?
- Quais foram as limitações e dificuldades encontradas?
- Quais são as principais conclusões?