INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
SÃO PAULO
Campus Birigui

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
Campus Birigui
Bacharelado em Engenharia de Computação

Disciplina: Mineração de Dados

Professor: Murilo Varges da Silva

Orientações para trabalho final da disciplina

Objetivo:

Estimular o aluno a aplicar os conhecimentos apresentados no decorrer da disciplina em problemas reais de mineração de dados, utilizando as técnicas de seleção, pré-processamento e transformação de dados, técnicas de visualização de dados, análise descritiva, análise de grupos, classificação e estimação/regressão.

Entregas:

- Artigo com no máximo 15 páginas contendo a descrição da base de dados, detalhamentos das atividades executadas com descrição e imagens que permitam avaliar o resultado da atividade, entrega via MOODLE;
- Slides utilizados na apresentação, na data da apresentação via MOODLE;
- Código fonte disponibilizado no GITHUB.

Atividades desejadas:

1 – Seleção e pré-processamento de dados:

- Escolha uma base de dados em http://archive.ics.uci.edu/ml ou no kaggle.com
- Avalie as características da base de dados: problema a ser investigado, número de amostras, número de atributos, tipos de atributos, possui valores ausentes?
- Utilizando a linguagem **Python**, junto com as bibliotecas **pandas** e **numpy**, crie um código que efetue uma limpeza de dados aplicando as técnicas apresentadas na aula de hoje.
- Avaliar se os dados estão desbalanceados, ou seja, se existem mais amostras de determinada(s) classe(s). Se existir desbalanceamento aplicar alguma técnica de balanceamento.

2 – Normalização e redução de dados:

- Utilizando a base de dados escolhida;
- Primeiramente realize o procedimento de limpeza de dados Atividade 1;
- Utilize alguma técnica de normalização de dados;
- Utilize a técnica PCA e PLOT os dois principais componentes.
- Avaliar como se apresentam as classes na visualização dos principais componentes e se os dados são linearmente separáveis.

3 – Análise descritiva de dados - Visualização:

- Utilizando a base de dados escolhida;
- Primeiramente realize o procedimento de limpeza de dados Atividade 1;
- Realize a distribuição de frequência para algum(s) atributo(s) da base dados;
- Utiliza alguma técnica de visualização para analisar os dados com base na distribuição de frequência. (Histograma, Gráfico de setores, dispersão, etc)

4 – Análise descritiva de dados - Medidas:

- Utilizando a base de dados escolhida;
- Primeiramente realize o procedimento de limpeza de dados Atividade 1;
- Calcular medidas de resumo apresentadas na aula de hoje:
    - Medidas de tendencia central;
    - Medidas de dispersão;
    - Medidas de posição relativa;
    - Medidas de associação.

5 – Análise de grupos:

- Utilizando a base de dados escolhida;
- Primeiramente realize o procedimento de limpeza de dados Atividade 1;
- Utilizar os algoritmos K-means e GMM
    - Utilizando apenas 2 principais componentes (PCA);
    - Rodar KMeans variando o número de grupos (parâmetro k)
    - Plotar os resultados dos agrupamentos para diferentes valores de K;
    - Avaliar o k-means com diferentes medidas de distância;
- Utilize medidas para avaliar a qualidade dos agrupamentos: coeficiente de forma, homogeneidade, etc

6 – Classificação - KNN:

- Utilizando a base de dados escolhida;
- Primeiramente realize o procedimento de limpeza de dados Atividade 1;
- Fazer os procedimentos de normalização que achar necessário;
- Utilizar os seguintes classificadores:
    - Árvore de Decisão (Decision Tree)
    - K-NN (K-Mearest Neighbors)
        - Informar o parâmetro K que apresentou o melhor resultado e como ele foi encontrado.
    - SVM (Support Vectors Machine)
    - Rede Neural MLP (Multilayer Perceptron)
        - Informar arquitetura da rede: número de neurônios, camadas escondidas, tipo de função de ativação, etc.
- Para os experimentos com os classificadores utilizar os seguintes protocolos de divisão da base de dados e medidas:
    - Fazer a divisão da base utilizando:
        - Holdout (Treinamento 70% e Teste 30%)
        - Cross-Validation (k=10)
    - Classificar com K-NN e calcular as seguintes métricas
        - Matrix de confusão
        - Acurácia
        - F1 Score
- Apresentar uma comparação final dos classificadores destacando os melhores resultados.