Projeto - Preparação de Dados

CMC-13 Introdução a Ciência de Dados
(Trabalho Individual ou em Grupo de dois ou três alunos)
Prof. Paulo André Castro

1. Objetivo

Exercitar e fixar conhecimentos adquiridos sobre Ciência de Dados e preparação de dados utilizando uma base de dados fornecida.

2. Descrição do Trabalho

2.1. Base de dados (dataset)

Neste projeto, devem ser usadas as informações de classificações de filmes (1 a cinco estrelas) para os filmes e usuários fornecidos no pacote de dados (data_preparation.zip) disponibilizada no site da disciplina. A base de dados oferece informações sobre usuários, filmes e as classificações dadas, aproximadamente 1 milhão de classificações dadas por 6000 usuários para 4000 diferentes filmes. Veja o arquivo LEIAME.txt para mais detalhes.

2.2. Prepare os dados para criação de um árvore de decisão

Utilizando a base de dados fornecida, processe os dados de modo a deixa-lo pronto para ser apresentado a um algoritmo de criação de árvore de decisão. Incluindo limpeza de dados, remoção de atributos irrelevantes, dados faltantes,

Justifique sua decisão sobre: eliminação de atributos, transformação de atritbutos, resolução de dados falantes.

Faça pelo menos uma transformação de dados para tratar a data de nascimento, discretize tal atributo segundo a idade do usuário:

* 1: "Under 18" * 18: "18-24" * 25: "25-34"
* 35: "35-44" * 45: "45-49" * 50: "50-55" * 56: "56+"

2.3. Classificador baseado em árvore de decisão

Utilizando a base de dados fornecida, criar um **classificador baseado em árvore de decisão** usando o algoritmo ID3 visto em sala, e que classifique de 1 a 5 estrelas um determinado filme para um determinado usuário conhecido (informações em users.csv).

2.4. Classificador a priori

Crie um classificador *a priori,* isto é que não usa nenhuma informação além da própria identificação do filme. Faça a média (truncada) das classificações para cada filme.

2.5. Análise Comparativa

Crie um conjunto de dados de testes com dez filmes listados em movies.csv que você (ou algum colega) tenha visto e os classifique de 1 a 5 estrelas. Compare os dois classificadores utilizando o dataset de testes : taxa de acerto, matriz de confusão, erro quadrático médio e estatística kappa.

Para fazer a comparação, selecione pelo menos dez filmes que você assistiu dentro da base e dê sua classificação em estrelas. Discuta qual classificador entre os dois é melhor e como poderia ser criado um classificador ainda melhor.

3. Material a ser Entregue e Prazo

Material: Relatório e Código; **Prazo de Entrega: 25/abril/2022**; Entregar através do Google Classroom!

Relatório do Projeto (**arquivo em formato pdf até 4 páginas**) com:

Título: Projeto Aprendizado de Máquina e Nomes da Equipe

1. Resultados Obtidos

2. Conclusões: Comentários e sugestões sobre o trabalho (complexidade/facilidade, sugestões, etc.).

3. Descrição da Implementação: Linguagem e IDE utilizados, comentários necessários para a execução do projeto.

Código do Projeto Notebook com Código-fonte do Sistema (em Python , R,Java, C, C++ ou C#).

Bom Trabalho!
Prof. Paulo André Castro