PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Instituto de Ciências Exatas e Informática

Bacharel em Sistemas de Informação - Unidade São Gabriel

TRABALHO PRÁTICO

Tecnologias para Descoberta de Conhecimento

Maria Luiza Moura Rocha (672283)

Matheus Santos Soares (641345)

OUTUBRO

2021

Relatório - Percepções do Dataset

Percepções importantes sobre o Dataset

Nosso Dataset escolhido apresenta um conjunto de dados contendo o panorama das taxas de suicídio, onde são levados em considerações as informações socieconômicas para calcular as taxas de suicídio por ano e país, com fins de encontrar sinais correlacionados ao aumento das taxas de suicídio por geração. É composto por uma grande quantidade de dados, sendo um total de 27.822 mil instâncias e possui 12 atributos, exemplificamos na entrega da primeira etapa do trabalho. Realizamos diversos testes e tratamos os dados no Dataset escolhido para eliminar dados faltantes, tratar os ruídos, remoção de outliers para aprimorar a taxa de precisão dos dois algoritmos implementados (J48 e IBK).

- Segue abaixo os tratamentos que realizamos no Dataset:

- No atributo ***country***, pensamos em classificá-los por continentes, mas ocasionaria ruídos nos dados. O próprio Weka selecionou e organizou os dados dos países que repetiam várias vezes.
- Nos atributos: ***year***, ***age*** e ***suicides_no***, quando rodamos o Dataset no Weka com o algoritmo J48, conseguimos analisar ao gerar a árvore de decisão, que esses atributos são “fortes” comparados aos demais.
- O atributo ***country-year***, nós analisamos que ele é basicamente a repetição dos dados dos atributos ***country*** e ***year***. Assim, resolvemos eliminá-lo, visto que não influenciou na porcentagem de classificação nos algoritmos J48 e IBK.
- O atributo ***hdi_for_year***, notamos diversos dados faltantes. Logo, cogitamos a em utilizar algumas das técnicas de ML, mas no nosso caso não seria muito viável pois temos uma grande quantidade de dados e cada país tem um HDI - Índice de Desenvolvimento Humano por ano. Dessa forma, resolvemos eliminar este atributo.
- Nos demais atributos: ***population***, ***suicides/100k pop***, ***gdp_for_year ($)***, ***gdp_per_capita ($)***, ***generation*** (Classe), não tratamos os dados, visto que com ou sem eles, a porcentagem de classificação nos algoritmos J48 e IBK foi alta.

Link do Dataset: *Suicide Rates Overview 1985 to 2016 | Kaggle*