Trabalho Prático
Vinho e qualidade:
Uma análise utilizando Machine Learning

Iyan Lucas Duarte Marques[1], **Samir do Amorim Cabraia**[1], **Wesley Filemon Rocha Rodrigues**[1]

[1]Instituto de Ciências Exatas e Informática - Pontifícea Universidade Católica Minas Gerais (PUC-MG)

1. Introdução

2. Base de Dados

A base utilizada é a *Wine Quality Data Set*, disponível no *UCI Machine Learning Repository*. A mesma foi introduzida em um paper por Paulo Cortez [1] que propõe uma forma de predição das preferências de gosto e qualidade de vinho do homem. Desta forma, a base consiste em análise de críticos do vinho *"Vinho Verde"*[1] português. Há duas tabelas, uma de vinho branco com 4898 instâncias e de vinho tinto com 1599 instâncias, para a finalidade deste trabalho, se utilizou somente a base de vinhos brancos, pela quantidade e estado das instâncias. Eles são classificados de um score de 0 a 10, sendo 0 muito ruim e 10 extremamente excelente. Este score (atributo 12) é o resultado entre a média dos resultados dados por 3 sommeliers experts.

2.1. Atributos

São 12 atributos, cada um representando um teste objetivo:

- 1 - Acides fixa
- 2 - Acides volátil
- 3 - Acido cítrico
- 4 - Açúcar residual
- 5 - Clorídeos
- 6 - Dióxido de enxofre livre
- 7 - Total de dióxido de enxofre
- 8 - Densidade
- 9 - pH
- 10 - Sulfatos
- 11 - álcool
- 12 - qualidade (score entre 0 e 10)

3. Resultados Preliminares

Ao acessar e estudar a base, percebe-se que a mesma está relativamente desbalanceada. Há mais vinhos normais do que ruins ou excelentes, apesar de que as classes estão ordenadas. Desta forma, para medida de comparação, foi executado na plataforma WEKA[2] o algoritmo Random Forest, o qual apresentou os seguintes resultados:

Analisando os resultados, é evidente que a base está com algum conflito, decorrências do desbalanceamento. Conforme o resultado, o time se empenhou em atingir

[1]Companhia portuguesa de vinhos e adegas. Mais informações: https://www.vinhoverde.pt

[2]Waikato Environment for Knowledge Analysis

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.5027
Mean absolute error                      0.5861
Root mean squared error                  0.7205
Relative absolute error                 88.2976 %
Root relative squared error             84.8633 %
Total Number of Instances              100
```

Figura 1. Resultados com a base desbalanceada

um método de que se faça o balanceamento de forma que se mantenha a integridade da base e o numero médio de instâncias. Após reuniões, concluimos que o algoritmo que apresentou os melhores resultados foi o Resample[3], que apresentou os seguintes resultados a partir do Random Forest

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8873
Mean absolute error                      0.253
Root mean squared error                  0.4186
Relative absolute error                 36.809  %
Root relative squared error             46.8794 %
Total Number of Instances              4898
```

Figura 2. Resultados com a base desbalanceada

Analisando os resultados do balanceamento, é perceptível uma melhora de mais de 50% em todos os atributos gerados pelo *output* do método.

Referências

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009. Smart Business Networks: Concepts and Empirical Evidence.

[3]Produz uma subamostra aleatória de um conjunto de dados usando amostragem com substituição ou sem substituição.