U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Data exploration and enrichment for supervised classification

Elementos de Inteligência Artificial e Ciência de Dados

Cristiana Silva-up202305464
Filipa Marinha-up202304935
Filipa Ferreira-up202305895

Data exploration and enrichment for supervised learning

A análise exploratória de dados e a aplicação de modelos de 'supervised learning' de classificação desempenham um papel crucial no campo da ciência de dados, permitindo a extração de insights valiosos e a tomada de decisões baseadas em conjuntos de dados complexos.

Neste trabalho, ambicionamos realizar uma análise aprofundada dos dados fornecidos, referentes a um dataset da doença carcinoma hepatocelular, com o propósito de classificá-los mediante a utilização dos modelos de ‘supervised learning’.

Os principais passos a seguir para o desenvolvimento deste trabalho são: exploração, pré-processamento, modelação e avaliação de dados e interpretação dos resultados.

Formulação do problema

Objetivo Prever se o paciente, com base nos seus sintomas, 'lives' ou 'dies'.

Dataset Base de dados com um total de 165 inputs (número total de pacientes)

Etapas Análise de Dados; Pré-processamento de Dados; Análise Exploratória; Classificação; Comparação Resultados

Formulação do problema

Os dados fornecidos no dataset podem ser agrupados da seguinte forma:

- **Categorical**: gender (male or female), class (lives or dies), PS (active, ambulatory, restricted, selfcare, disable), Encephalophathy (None, Grade I/II, Grade III/IV), Ascites (None, Mild, Moderate/Severe
- **Numerical**: age, grams_day, packs_year, INR, AFP, hemoglobin, MCV, leucocytes, platelets, albumin, total_bil, ALT, AST, GGT, ALP, TP, creatinine, nodules, major_dim, dir_bil, iron, sat, ferritin
- **Boolean**: symptoms, alcohol, HBsAg, HBeAg, HBcAb, HCVAb, Cirrhosis, Endemic, Smoking, Diabetes, Obesity, Hemochro, AHT, CRI, HIV, NASH, Varices, Spleno, PHT, PVT, Metastasis, Hallmark

Os dados apresentados como 'Boolean' são na verdade expressos como 'Yes' ou 'No' invés de 'True' ou 'False'. Para além disso, todas as categorias apresentadas estão sujeitas a alguns valores em falta.

Por fim, visto que estamos a realizar uma 'classificação', podemos comprovar o quão precisos foram os resultados de modo a avaliar a eficácia dos métodos usados.

Pré-processamento dos dados

Eliminação de dados desconhecidos

Substituição de "?" por médias

Guardar os dados pré-processados

Para facilitar o acesso aos dados pré-processados criamos um novo arquivo com os dados modificados

Modelação de dados

Para melhor compreender a distribuição dos dados, transformou-se as ‘strings’ em ‘integers’ , utilizando, posteriormente, os dados para desenvolver gráficos.

Esta etapa envolve a aplicação de uma série de técnicas e transformações aos dados brutos com o objetivo de prepará-los para a análise e modelação de dados.

A este processo antecedeu-se uma ‘Exploração de Dados’ para uma melhor identificação de erros ou da relevância dos dados.

Classificação

Target 'lives' ou ' dies'

Assumindo que o teste 3 apresenta melhores resultados, verificou-se o seguinte:

Decision Tree

~ 0,62
'accuracy score': 0,82

K-NN

~ 0,76
'accuracy score': 0,76

Comparação de resultados

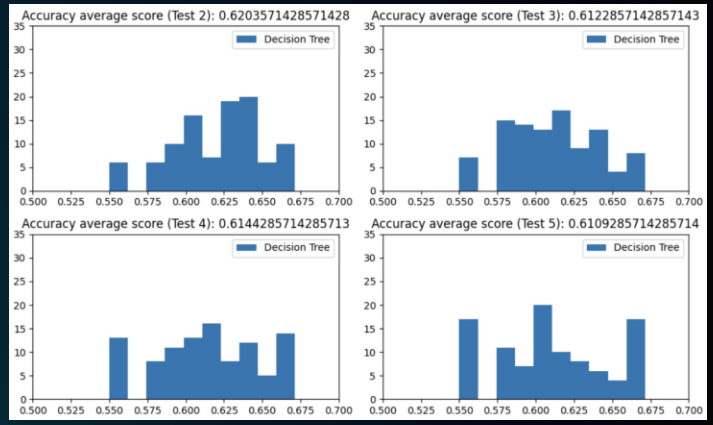

Figura 1: comparação dos gráficos acerca da 'accuracy average score' para a Decision Tree nos quatro testes executados.

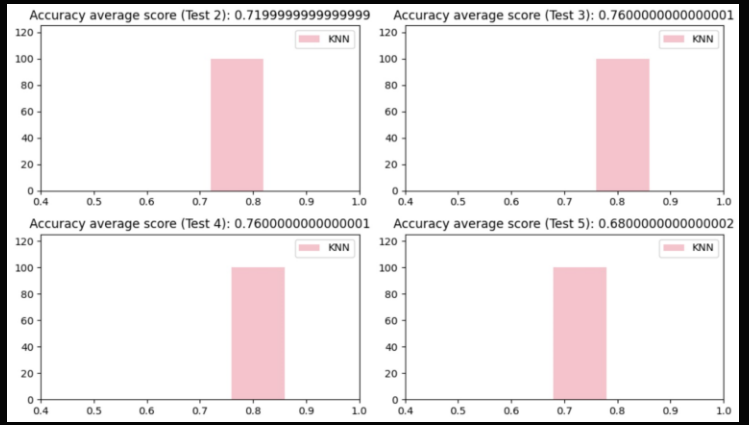

Figura 2: comparação dos gráficos acerca da 'accuracy average score' para o K-NN nos quatro testes executados.

Conclusão

No processo de classificação realizado neste estudo, foram conduzidos 4 testes abrangendo diferentes aspetos, desde a configuração dos dados até à aplicação de algoritmos de machine learning, como a 'Decision Tree' e o K-NN. Ao longo destes testes, observou-se que não houve grandes discrepâncias nos valores de pontuação média de precisão da árvore de decisão, que se manteve em torno de 0,62. Quanto ao K-NN, o teste 3 apresentou o valor mais alto de precisão média, atingindo 0,76, após a remoção de certos parâmetros (nomeadamente, a remoção das colunas que possuiam mais '?')

Em relação a ' accuracy', verificamos que o teste 3 revelou-se o melhor teste em que se obteve 0,82 para a Decision Tree e 0,76 para o K-NN. Desta forma, utilizando o teste 3, empregou-se a técnica grid search, alcançando um score de 0,70 para a Decision Tree e 0.67 para o K-NN.

Estas conclusões destacam a importância de realizar uma análise cuidadosa dos dados, incluindo a exclusão de atributos redundantes, além de ressaltar a necessidade de ajuste de parâmetros para obter um desempenho otimizado dos modelos de classificação usados.

Bibliografia

-Slides das aulas
-ChatGPT
-https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.replace.html
-https://saturncloud.io/blog/python-pandas-how-to-remove-nan-and-inf-values/
-https://scikit--learn-org.translate.goog/stable/modules/generated/sklearn.metrics.accuracy_score.html?_x_tr_sl=en&_x_tr_tl=pt&_x_tr_hl=pt-PT&_x_tr_pto=sc
-https://stackoverflow.com/questions/72876505/what-is-the-difference-between-df-dropinplace-true-and-df-df-drop
-https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
-https://scikit-learn.org/stable/modules/grid_search.html
-https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/
-https://www.youtube.com/watch?v=PHxYNGo8NcI