Relatório - Trabalho Prático 2
Universidade de Évora
Aprendizagem Automática 2022/2023
Bernardo Vitorino (l48463), Daniel Barreiros (l48452) e Tomás Antunes (l48511)

I – Introdução

O relatório que se segue é alusivo ao trabalho prático da disciplina de aprendizagem automática, lecionado pelo docente Luís Rato e tem como objetivo indicar os requisitos em termos de bibliotecas python necessárias para o funcionamento do trabalho proposto, analisar e descrever o conjunto de dados disponibilizados, descrever o conjunto de experiências realizadas que consideramos mais relevantes e indicar o desempenho que obtivemos com estas experiências.

Por último iremos apresentar uma breve discussão sobre os resultados bem como as conclusões obtidas através da análise dos mesmos.

II – Requisitos

Para a realização deste trabalho foi-nos permitido utilizar bibliotecas do sklearn e pandas, entre as quais utilizamos:

- sklearn.metrics
  - para utilização da função classification_report;
  - para utilização da função recall_score;
  - para utilização da função precision_score;
- sklearn.tree para utilização do classificar DecisionTreeClassifier;
- sklearn.model_selection para utilização do algoritmo GridSearchCV para selecionar os valores que maximizam a cobertura.
- biblioteca pandas para manipulação do ficheiro .csv.

III – Analise do Conjunto de Dados

Através da análise do ficheiro “dropout-trabalho2.csv” observamos todos os atributos e raciocinamos quais não seriam importantes para a construção do modelo. Decidimos assim retirar o atributo Id pois este difere em todas as instâncias e não tem qualquer correlação com o resultado pretendido, visto que, o nº de aluno não tem impacto no desempenho académico.
Analisando o mesmo ficheiro, na concretização do objetivo secundário, observamos todos os atributos e decidimos escolher os que nos pareceram mais relevantes para o objetivo em questão. Escolhemos (Créditos concluídos, Créditos Inscritos e Notas Obtidas), contudo, visto que um dos critérios estabelecido seria a utilização de apenas dois, tentamos encontrar a solução que nos pareceu mais logica.

O primeiro raciocínio consistiu na utilização dos créditos inscritos e dos créditos concluídos considerando que um aluno que tenha uma boa quantidade de créditos concluídos relativamente aos créditos inscritos mais dificilmente iria desistir que um aluno que obtivesse poucos créditos concluídos, relativamente ao número de créditos inscrito.

No segundo raciocínio pensamos que a primeira opção poderia não ser tão logica visto que o aluno poderia não querer quantidade, mas sim qualidade, no sentido em que, iria preferir obter menos créditos concluídos mas todos eles com uma nota mais alta (de forma a obter melhor classificação global de curso). A obtenção de tais resultados não iria refletir que o aluno estaria propenso a desistir, mas sim a querer obter boas classificações.

Após ponderarmos qual o caminho a seguir, decidimos optar pela segunda opção visto que, para nós, faz mais sentido um aluno preferir ter menos créditos feitos com boa nota, ao invés de muitos créditos feitos e com baixa qualificação.

Posteriormente decidimos utilizar os valores da média aritmética de cada aluno para cada um dos atributos visto que através da média aritmética é possível obter um resultado equilibrado dos parâmetros.

IV – Descrição do Conjunto de Experiências e Considerações

No desenvolvimento do código começamos por criar o modelo de maneira a que usasse um classificador de forma direta (Referindo os parâmetros utilizados e os seus valores explicitamente na criação de uma instancia deste classificador).

Após analisar os resultados obtidos nesta abordagem reparamos que esta não seria a forma mais correta de obter os melhores valores aplicados aos parâmetros de forma que a cobertura fosse maximizada e a precisão tivesse o valor mínimo de 70%.
Assim, decidimos optar por outra abordagem. Esta abordagem tem como base o uso do algoritmo GridSearchCV que permite o uso de fits sucessivos variando a combinação dos valores de cada parâmetro de forma a encontrar a melhor combinação. Este algoritmo faz uso de cross-validation de forma a que a construção do modelo tenha menos risco de perder padrões importantes sem risco de incrementar o erro, induzido por bias.
Deste modo utilizando o parâmetro do GridSearchCV scoring com o valor ‘recall’ e dispondo de uma gama de valores para diferentes parâmetros obtivemos a combinação de valores desejada (De forma a maximizar a cobertura).

Juntamente com o GridSearchCV tivemos de escolher um classificador. Optámos assim por escolher um dos apresentados nas aulas da disciplina.

O classificador escolhido foi as Arvores de Decisão (DecisionTreeClassifier) juntamente com os parâmetros max_depth e min_samples_split. A gama de valores escolhidos para o parâmetro max_depth foram [1, 2, 4, 6, 8, 10, 12, … , 20] e a gama de valores escolhidos para o parâmetro min_samples_split foram [2, 4, 6, 8, 10, 12, 13, 14, 15].

Para o objetivo principal, após testarmos várias vezes a execução do programa verificamos que no parâmetro max_depth eram escolhidos maioritariamente valores altos mas sem nunca alcançar o valor 20. Já no parâmetro min_samples_split o valor maioritariamente escolhido é o 12.

Para o objetivo secundário, após testarmos várias vezes a execução do programa verificamos que no parâmetro max_depth eram escolhidos maioritariamente o valor 4. Já no parâmetro min_samples_split o valor maioritariamente escolhido é o 6.

V – Desempenho

Como era pretendido no trabalho, a níveis de desempenho obtivemos na precisão valores superiores a 70% e na cobertura valores superiores a 80% mas nunca superiores a 92%.

Assim sendo, consideramos que obtivemos um bom desempenho dado que este era o pedido no trabalho.

VI – Discussão e Conclusão

A parte do trabalho que consideramos mais desafiante foi a escolha dos atributos do dataset a usar no treino e no teste do objetivo secundário. Achamos interessante visto que surgiu discussão entre os elementos do grupo levando a que houvesse uma melhor analise e um melhor entendimento dos dados em estudo.

Relativamente à programação, tivemos de analisar a API do sklearn de modo que encontrássemos as melhores abordagens para a realização do trabalho.

Aprofundámos todos os temas deste trabalho de modo a que pudéssemos enriquece-lo e de maneira a que ficássemos a compreender melhor a matéria lecionada nesta disciplina.

Na parte principal do trabalho, do nosso ponto de vista, conseguimos eleger um classificador bom. Já na parte secundária conseguimos eleger dois atributos que fizeram o trabalho ter um bom desempenho.