A blue and black logo

Description automatically generatedAprendizagem 2024/2025

**Homework I – Group 034**

(ist199417, ist1106481)

**I. Pen-and-paper**

**1)**

Como os cálculos para este tipo de exercícios são algo extensivos e podem dar aso a erro, utilizámos python para computar a entropia e o ganho de informação para um dado conjunto de dados.

Carregamos os dados desta forma. Se quisermos tirar ou adicionar instâncias é só comentar/descomentar as linhas onde figuram.

A classe (yout) é a primeira (índice 0).

**A screenshot of a computer code

Description automatically generated**

Calculamos a entropia da variável como se segue:

A screenshot of a computer program

Description automatically generated

Calculamos a entropia da variável em relação à variável da seguinte forma:

**A screenshot of a computer code

Description automatically generated**

Já temos tudo o que é preciso para calcular o ganho de informação da variável .

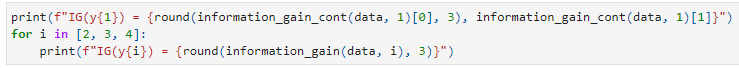
Como variáveis discretas e contínuas se calculam de maneiras diferentes, temos duas versões.

Em information\_gain\_cont tem-se em conta todos os modos de dividir a variável contínua em dois grupos e escolhe-se o modo com maior ganho de informação.

**A computer code with many numbers

Description automatically generated with medium confidence**

Agora basta chamar as funções e observar os resultados.

****

Comotem ganho de informação maior, escolhe-se a variável .  
Observemos o subconjunto de dados depois da escolha .

**A screenshot of a computer code

Description automatically generated**

Como e têm ambos ganho de informação igual, escolhe-se por ordem alfabética , com corte em .

Resposta Final:

A screenshot of a computer

Description automatically generated

**2)**

True

|  |  |  |  |
| --- | --- | --- | --- |
| Predicted | A | B | C |
| A | 2 | 0 | 0 |
| B | 0 | 4 | 0 |
| C | 1 | 0 | 5 |

**3)**

R: A classe com um F1 score mais baixo é a classe A.

**4)**

**A graph of mathematical equations

Description automatically generated with medium confidence**

**II. Programming**

**1)**

Característica menos discriminativa: Pressão Arterial.  
Característica mais discriminativa: Glicose.

**A graph of different colored lines

Description automatically generated with medium confidence**

**2)**

**A graph with blue and orange lines

Description automatically generated**

**3)**

Em primeiro lugar, quando o minimum\_sample\_split é 2, a training accuracy é obviamente 1 porque todas as instâncias são classificadas nas suas verdadeiras classes. Não há nenhuma folha que contenha duas instâncias de classes diferentes. Mas isto tem um problema: o overfitting. Como vemos no gráfico, para o grupo de teste, a accuracy pode ser melhorada. O modelo tem um desempenho excelente no grupo de treino, mas tem pouca capacidade de generalização.

Para mitigar este problema, uma ideia é aumentar o minimum\_sample\_split. Podemos então olhar para o outro lado do espetro. Dado minimum\_sample\_split igual a 100, vemos uma diminuição na accuracy a nível do grupo de treino, embora a accuracy no grupo de teste tenha aumentado. Esta perda de accuracy pode ser evidência de underfitting.

Por isso, há que encontrar um balanço da accuracy nos dois grupos. Analisando o gráfico, observamos que ter o minimum\_sample\_split entre 30 e 50 é o que nos dá esse tal balanço. A testing accuracy estabiliza daí para a frente e a training accuracy desce.

**4)**

**A diagram of a computer flowchart

Description automatically generatedi.**

**ii.**