Instituto Superior de Engenharia do Porto

Mestrado em Engenharia Informática

Mineração de Dados
Exame Época Normal
18 janeiro 2024
Duração 2:00h

---

1. Considere uma base de dados contendo informações sobre várias Instituições de Ensino Superior (IES), incluindo cursos oferecidos, unidades de curriculares, alunos matriculados, resultados das avaliações e informação sobre ex-alunos. Descreva modelos de mineração de dados que podem ser construídos tanto para obter informações sobre a situação atual das IES como para fazer análises preditivas.

2. Seja o seguinte conjunto sobre planetas habitáveis ou não.

Size	Orbit	Temperature	Habitable
Small	Far	High	No
Big	Far	Low	Yes
Small	Near	High	No
Small	Near	Low	Yes
Big	Near	High	No
Small	Near	Low	Yes
Small	Far	Low	Yes

Qual a previsão para a instância:

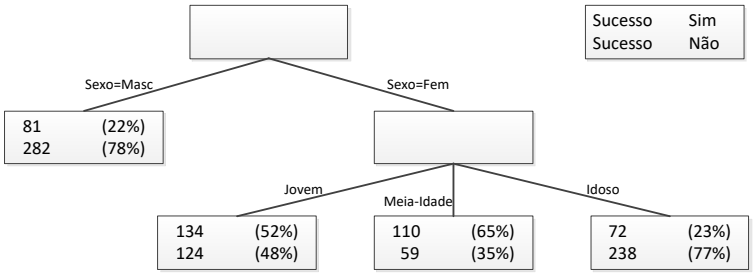
Size = Big; Orbit=Far; Temperature = High?

a) Usando o classificador Naive Bayes. Apresente os cálculos e se necessário considere como critério Laplace = 1.
b) Usando o algoritmo K-Vizinhos-mais-Próximos, considere K=3.

3. Defina aprendizagem em conjunto (*ensemble learning*). Explique o seu princípio e como difere este tipo de aprendizagem das abordagens tradicionais de aprendizagem (com apenas um único modelo).

4. Foi aplicado um algoritmo de construção de árvore de decisão sobre um amostra de dados relativa a **1650 pacientes** submetidos a um tratamento de obesidade. O modelo criado tem por objetivo prever o sucesso do tratamento para a obesidade e apresenta três atributos para cada um dos pacientes.

- Idade: Jovem, Meia-idade, Idoso
- Sexo: Fem, Masc
- Sucesso: Sim, Não



a) Assumindo que o método de amostragem usado para a criação/avaliação do modelo foi o método **holdout estratificado** (2/3 para treino; 1/3 para teste) e o modelo criado apresenta uma Taxa de Verdadeiros Positivos (TPR) de 35% e uma Taxa de Falsos Positivos (FPR) 45% apresente a matriz de confusão. Justifique apresentando todos os cálculos.

b) Diga qual o significado dos Falsos Positivos (FP) e dos Falsos Negativos (FN) neste modelo. Das métricas de avaliação de modelos de classificação estudadas, qual/quais as mais adequadas para avaliar este modelo em particular? Explique.

Instituto Superior de Engenharia do Porto

Mestrado em Engenharia Informática

Mineração de Dados
Exame Época Normal
18 janeiro 2024
Duração 2:00h

5. Considere a seguinte base de dados de compras:

tid	Items	A	B	C	D	E	F
1	ABCD	1	1	1	1		
2	BCEF		1	1		1	1
3	ADEF	1			1	1	1
4	AEF	1				1	1
5	BDF		1		1		1

a) Calcule as medidas de suporte, confiança e interesse das regras:

{E} => {F}

{A,B} => {C}

{D} => {F}

{D,E} => {A}

{F} => {C}

b) O modelo Suporte/Confiança apresenta algumas limitações. Nesse sentido das regras acima indique uma regra que apresente:

i. Um relacionamento ilusório entre os itens. Justifique a sua escolha.

ii. Uma regra rara mas interessante. Justifique a sua escolha.

iii. Uma regra descartável. Justifique a sua escolha.

6. Considere a seguinte matriz com o número de ocorrências de vários termos em vários documentos:

Documento/Termo	**t1**	**t2**	**t3**	**t4**	**t5**	**t6**	**t7**
d1	0	4	10	8	0	5	0
d2	5	19	7	16	0	0	32
d3	15	0	0	4	9	0	17
d4	22	3	12	0	5	15	0
d5	0	7	0	9	2	4	12

a) Calcule a ponderação tf-idf do termo t6 no documento d4.

b) Com base na informação disponível na matriz, explique como poderia calcular a similaridade entre os documentos.

7. O que é uma série temporal?

a) Explique o que significa a sigla ARIMA.

b) Descreva os passos para identificar e ajustar um modelo ARIMA a uma série temporal?

FIM