Iniziato	Monday, 9 January 2023, 09:57
Stato	Completato
Terminato	Monday, 9 January 2023, 11:47
Tempo impiegato	1 ora 49 min.
Valutazione	Non ancora valutato

Domanda **1**

Completo

Punteggio max.: 20,00

The task is described in this document.

The data file is here. The file with the feature names is here

Upload only your notebook, not the data. Please name your notebook according to the directions given in the document linked above.

luigi_porcelli.ipynb

Domanda **2**

Risposta errata

Punteggio ottenuto 0,00 su 0,67

Which is the main reason for the *standardization* of numeric attributes?

Scegli un'alternativa:

- a. Map all the numeric attributes to a new range such that the mean is zero and the variance is one.
- b. Change the distribution of the numeric attributes, in order to obtain gaussian distributions
- c. Remove non-standard values
- d. Map all the nominal attributes to the same range, in order to prevent the values with higher frequency from having prevailing influence ✖

Your answer is incorrect.

La risposta corretta è: Map all the numeric attributes to a new range such that the mean is zero and the variance is one.

Domanda **3**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

Given the two binary vectors below, which is their similarity according to the Simple Matching Coefficient?

a b c d e f g h i j

1 0 0 0 1 0 1 1 0 1

1 0 1 1 1 0 1 0 1 0

Scegli un'alternativa:

- a. 0.5 ✔
- b. 0.3
- c. 0.2
- d. 0.1

Risposta corretta.

La risposta corretta è: 0.5

Domanda **4**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

Which of the following statements is *true*?

Scegli una o più alternative:

- a. Outliers can be due to noise ✔
- b. The noise can generate outliers ✔
- c. The noise always generate outliers
- d. The data which are similar to the majority are never noise

Your answer is correct.

Le risposte corrette sono: Outliers can be due to noise, The noise can generate outliers

Domanda **5**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

In which mining activity the *Information Gain can be useful?*

Scegli un'alternativa:

- a. Classification ✔
- b. Clustering
- c. Discovery of association rules
- d. Discretization

Your answer is correct.

La risposta corretta è: Classification

Domanda **6**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

What is the *cross validation*

Scegli un'alternativa:

- a. A technique to obtain a good estimation of the performance of a classifier when it will be used with data different from the training set ✔
- b. A technique to obtain a good estimation of the performance of a classifier with the training set
- c. A technique to improve the quality of a classifier
- d. A technique to improve the speed of a classifier

Risposta corretta.

La risposta corretta è: A technique to obtain a good estimation of the performance of a classifier when it will be used with data different from the training set

Domanda **7**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

A Decision Tree is...

Scegli un'alternativa:

- a. A tree-structured plan of tests on single attributes to forecast the target ✔
- b. A tree-structured plan of tests on multiple attributes to forecast the target
- c. A tree-structured plan of tests on single attributes to forecast the cluster
- d. A tree-structured plan of tests on single attributes to obtain the maximum purity of a node

Risposta corretta.

La risposta corretta è: A tree-structured plan of tests on single attributes to forecast the target

Domanda **8**

Risposta errata

Punteggio ottenuto 0,00 su 0,67

Which of the following preprocessing activities is useful to build a Naive Bayes classifier if the independence hypothesis is violated

Scegli un'alternativa:

- a. Feature selection
- b. Normalisation
- c. Standardisation
- d. Discretisation ✖

Risposta errata.

La risposta corretta è: Feature selection

Domanda **9**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

Which of the statements below about *Hierarchical Agglomerative Clustering* is true?

- a. Requires the definition of *distance between sets of objects* ✔
- b. Requires the definition of *Inertia* of clusters
- c. Is based on a well founded statistical model
- d. Is very efficient, also with large datasets

Your answer is correct.

La risposta corretta è:
Requires the definition of *distance between sets of objects*

Domanda **10**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

What does K-means try to minimise?

Scegli un'alternativa:

- a. The *distortion*, that is the sum of the squared distances of each point with respect to its centroid ✔
- b. The *separation*, that is the sum of the squared distances of each cluster centroid with respect tho the global centroid of the dataset
- c. The *distortion*, that is the sum of the squared distances of each point with respect to the points of the other clusters
- d. The *separation*, that is the sum of the squared distances of each point with respect to its centroid

Risposta corretta.

La risposta corretta è: The *distortion*, that is the sum of the squared distances of each point with respect to its centroid

Domanda **11**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

Which of the following characteristic of data can reduce the effectiveness of DBSCAN?

Scegli un'alternativa:

- a. Presence of clusters with different densities ✔
- b. All the variables are the same range of values
- c. Clusters have concavities
- d. Presence of outliers

Your answer is correct.

La risposta corretta è: Presence of clusters with different densities

Domanda **12**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

Match the rule evaluation formulas with their names

$$\frac{sup(A \Rightarrow C)}{sup(A)}$$ Confidence ✔

$$\frac{conf(A \Rightarrow C)}{sup(C)}$$ Lift ✔

$$sup(A \cup C) - sup(A)sup(C)$$ Leverage ✔

$$\frac{1 - sup(C)}{1 - conf(A \Rightarrow C)}$$ Conviction ✔

Your answer is correct.

La risposta corretta è: $\frac{sup(A \Rightarrow C)}{sup(A)}$ → Confidence,

$\frac{conf(A \Rightarrow C)}{sup(C)}$ → Lift,

$$sup(A \cup C) - sup(A)sup(C)$$ → Leverage,

$$\frac{1 - sup(C)}{1 - conf(A \Rightarrow C)}$$ → Conviction

Domanda **13**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

Consider the transactional dataset below

ID	Items
1	A,B,C
2	A,B,D
3	B,D,E
4	C,D
5	A,C,D,E

Which is the *support* of the rule A,C ⇒ B?

Scegli un'alternativa:

- a. 20% ✔ 1 / 5
- b. 100%
- c. 40%
- d. 50%

Risposta corretta.

La risposta corretta è: 20%

Domanda **14**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

How can we measure the quality of a trained regression model?

- a. With a formula elaborating the difference between the forecast values and the true ones ✔
- b. With a confusion matrix
- c. With precision, recall and accuracy
- d. Counting the number of values correctly forecast

Your answer is correct.

La risposta corretta è:
With a formula elaborating the difference between the forecast values and the true ones

Domanda **15**

Risposta corretta

Punteggio ottenuto 0,67 su 0,67

What does K-means try to minimise?

Scegli un'alternativa:

- a. The *distortion*, that is the sum of the squared distances of each point with respect to its centroid ✔
- b. The *separation*, that is the sum of the squared distances of each cluster centroid with respect tho the global centroid of the dataset
- c. The *distortion*, that is the sum of the squared distances of each point with respect to the points of the other clusters
- d. The *separation*, that is the sum of the squared distances of each point with respect to its centroid

Risposta corretta.

La risposta corretta è: The *distortion*, that is the sum of the squared distances of each point with respect to its centroid

Domanda **16**

Risposta errata

Punteggio ottenuto 0,00 su 0,67

In which part of the CRISP methodology we perform the **test design** activity?

a. Modeling

b. Evaluation ✖

c. Business Understanding

d. Data Understanding

Your answer is incorrect.

La risposta corretta è:
Modeling