Iniziato	mercoledì, 11 febbraio 2026, 14:40
Stato	Completato
Terminato	mercoledì, 11 febbraio 2026, 16:37
Tempo impiegato	1 ora 56 min.
Valutazione	**28** su un massimo di 30 (**93**%)
Feedback	Passed

Domanda 1

Completo

Punteggio ottenuto 19,00 su 20,00

- The task is described in this document.
- The data file is here.

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

lab4_079_riccardo.gardenghi.ipynb

Commento:

T5: minor error -1 - You dind't use cross validation in
ex 5

Domanda 2

Risposta corretta

Punteggio ottenuto 0,50 su 0,50

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 correct.

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 0,50 su 0,50

Which of the following *is not* an objective of feature selection

Scegli un'alternativa:

- a. Select the features with higher range, which have more influence on the computations ✔
- b. Avoid the *curse of dimensionality*
- c. Reduce time and memory complexity of the mining algorithms
- d. Reduce the effect of noise

Risposta corretta.

La risposta corretta è: Select the features with higher range, which have more influence on the computations

Domanda 4

Risposta corretta

Punteggio ottenuto 0,50 su 0,50

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 0,50 su 0,50

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 0,50 su 0,50

Why do we *prune* a decision tree?

Scegli un'alternativa:

- a. To eliminate parts of the tree where the decisions could be influenced by random effects ✔
- b. To eliminate parts of the tree where the decision could generate *underfitting*
- c. To eliminate attributes which could be influenced by random effects
- d. To eliminate rows of the dataset which could be influenced by random effects

Your answer is correct.

La risposta corretta è: To eliminate parts of the tree where the decisions could be influenced by random effects

Domanda 7

Risposta corretta

Punteggio ottenuto 0,50 su 0,50

In a Decision Tree for classification, what is a *leaf node*?

- a. A node which assigns a class value to the objects passing the tests on the path from the root to the node itself ✔
- b. A node where all the objects belong to the same class
- c. A node which allows classification without errors
- d. A node which assigns a class value only by majority of the examples

Your answer is correct.

La risposta corretta è:
A node which assigns a class value to the objects passing the tests on the path from the root to the node itself

Domanda 8

Risposta corretta

Punteggio ottenuto 0,50 su 0,50

In a Neural Network, what is the *backpropagation?*

- a. The technique used to adjust the connection weights according to the difference between the desired output and the output generated by the network ✔
- b. The technique used to adjust the node weights according to the difference between the desired output and the output generated by the network
- c. The technique used to adjust the output according to the difference between the desired weights and the actual weights
- d. The technique used to adjust the weights limiting the probability of overfitting

Your answer is correct.

La risposta corretta è:
The technique used to adjust the connection weights according to the difference between the desired output and the output generated by the network

Domanda 9

Risposta corretta

Punteggio ottenuto 0,50 su 0,50

Which of the following is a strength of the clustering algorithm DBSCAN?

Scegli una o più alternative:

- a. Ability to find cluster with concavities ✔
- b. Ability to separate outliers from regular data ✔
- c. Very fast computation
- d. Requires to set the number of clusters as a parameter

Your answer is correct.

Le risposte corrette sono: Ability to find cluster with concavities, Ability to separate outliers from regular data

Domanda 10

Risposta corretta

Punteggio ottenuto 0,50 su 0,50

Which of the statements below is true? (Only one)

Scegli un'alternativa:

- a. Sometimes k-means stops to a configuration which does not give the minimum distortion for the chosen value of the number of clusters. ✔
- b. K-means always stops to a configuration which gives the minimum distortion for the chosen value of the number of clusters.
- c. K-means finds the number of clusters which gives the minimum distortion
- d. K-means works well also with datasets having a very large number of attributes

Your answer is correct.

La risposta corretta è: Sometimes k-means stops to a configuration which does not give the minimum distortion for the chosen value of the number of clusters.

Domanda 11

Risposta corretta

Punteggio ottenuto 0,50 su 0,50

What can impact the results of agglomerative clustering?

Scegli un'alternativa:

- a. The size of the dataset.
- b. The computational complexity of the algorithm.
- c. The number of clusters formed.
- d. The choice of a distance metric and linkage method. ✔

La risposta corretta è: The choice of a distance metric and linkage method.

Domanda 12

Risposta errata

Punteggio ottenuto 0,00 su 0,50

Which of the following statements regarding the discovery of association rules is true? (One or more)

Scegli una o più alternative:

- a. The confidence of a rule can be computed starting from the supports of itemsets ✔
- b. The support of an itemset is anti-monotonic with respect to the composition of the itemset
- c. The confidence of an itemset is anti-monotonic with respect to the composition of the itemset ✖ This is wrong, because the "confidence of an itemset" does not make any sense, we consider only the "confidence of a rule"
- d. The support of a rule can be computed given the confidence of the rule

Your answer is incorrect.

Le risposte corrette sono: The confidence of a rule can be computed starting from the supports of itemsets, The support of an itemset is anti-monotonic with respect to the composition of the itemset

Domanda 13

Risposta corretta

Punteggio ottenuto 0,50 su 0,50

How does *pruning* work when generating frequent itemsets?

Scegli un'alternativa:

- a. If an itemset is not frequent, then none of its supersets can be frequent, therefore the frequencies of the supersets are not evaluated ✔
- b. If an itemset is frequent, then none of its supersets can be frequent, therefore the frequencies of the supersets are not evaluated
- c. If an itemset is not frequent, then none of its subsets can be frequent, therefore the frequencies of the subsets are not evaluated
- d. If an itemset is frequent, then none of its subsets can be frequent, therefore the frequencies of the subsets are not evaluated

Risposta corretta.

La risposta corretta è: If an itemset is not frequent, then none of its supersets can be frequent, therefore the frequencies of the supersets are not evaluated

Domanda 14

Risposta corretta

Punteggio ottenuto 0,50 su 0,50

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 0,50 su 0,50

Which is different from the others?

Scegli un'alternativa:

- a. Silhouette Index ✔ This is not a index for the evaluation of purity
- b. Gini Index
- c. Misclassification Error
- d. Entropy

Risposta corretta.

La risposta corretta è: Silhouette Index

Domanda 16

Risposta corretta

Punteggio ottenuto 0,50 su 0,50

Which of the activities below is part of "Business Understanding" in the CRISP methodology?

- a. Which machine learning functions are necessary for my problem?
- b. Which data are available?
- c. Which data must be collected with a specific campaign?
- d. Which are the resources available (manpower, hardware, software, ...) ✔

Your answer is correct.

La risposta corretta è:
Which are the resources available (manpower, hardware, software, ...)

Domanda 17

Risposta corretta

Punteggio ottenuto 0,50 su 0,50

Which of the following sentences describes an advantage of a Data Warehouse with respect to a standard DBMS

Scegli una o più alternative:

- a. Allows analysis along the time dimension ✔
- b. Allows efficient execution of key-based queries
- c. Allows efficient execution of multi-dimensional queries ✔
- d. Has tools for helping to solve inconsistencies ✔
- e. Manages efficiently data updates

Risposta corretta.

Le risposte corrette sono: Allows analysis along the time dimension, Allows efficient execution of multi-dimensional queries, Has tools for helping to solve inconsistencies

Domanda 18

Risposta corretta

Punteggio ottenuto 0,50 su 0,50

Talking about ETL, which of the following activities is related to the **Cleansing** step?

Scegli una o più alternative:

- a. **Snapshot** of the operational data
- b. Association of a **timestamp** to the operational data
- c. Elimination of **duplicates** ✔
- d. Usage of dictionaries to solve **inconsistencies** ✔

Risposta corretta.

Le risposte corrette sono: Elimination of **duplicates**, Usage of dictionaries to solve **inconsistencies**

Domanda 19

Risposta corretta

Punteggio ottenuto 0,50 su 0,50

Which of the definition below describes the OLAP operation **Drill-Down**?

Scegli un'alternativa:

- a. Creates a link between concepts in interrelated cubes, to compare them
- b. Causes an increase in data aggregation and removes a detail level in a hierarchy
- c. Reduces data aggregation and adds a detail level to a hierarchy ✔
- d. Reduces the number of cube dimensions after setting one of the dimensions to a specific value
- e. Changes the layout, in order to analyse a group of data from a different viewpoint

Risposta corretta.

La risposta corretta è: Reduces data aggregation and adds a detail level to a hierarchy

Domanda 20

Parzialmente corretta

Punteggio ottenuto 0,33 su 0,50

Talking about the general idea of database, what is the purpose of the "Schema on read" strategy?

Scegli una o più alternative:

- a. Possibility to extract data in various shapes
- b. Optimisation for various types of queries
- c. Flexibility for any kind of query ✔
- d. Avoid preprocessing of data before writing ✔

Risposta parzialmente esatta.

Hai selezionato correttamente 2.
Le risposte corrette sono: Possibility to extract data in various shapes, Flexibility for any kind of query, Avoid preprocessing of data before writing

Domanda 21

Risposta errata

Punteggio ottenuto 0,00 su 0,50

What is *Data Ingestion*?

a. A process that copies data from sources to a repository, taking care of possible differences in speed between the generation and the storing process

b. A process that copies data from sources to a repository, ensuring high data quality

c. A process that copies data from sources to a repository, making the transformation required by the users ✖

d. A process that copies data from sources to a Data Warehouse guaranteeing the correctness of data with respect to the schema

Your answer is incorrect.

La risposta corretta è:
A process that copies data from sources to a repository, taking care of possible differences in speed between the generation and the storing process