- Which of the following is NOT a property of a *metric* distance function? **BOUNDEDNESS**

- Given the two binary vectors below, which is their similarity according to the *Simple Matching Coefficient*? **0.5**

  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

- Given the two binary vectors below, which is their similarity according to *the Jaccard Coefficient*? **0.375**

  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

- What is the *single linkage*? **A method to compute the distance between two sets of items, it can be used in hierarchical clustering**

- Given the definitions below:
  - TP = True Positives
  - TN = True Negatives
  - FP = False Positives
  - FN = False Negatives

  Which of the formulas below computes the *recall (*or *hit rate* or *sensitivity)* of a binary classifier? **TP/(TP+FN)**

- Given the definitions below:
  - TP = True Positives
  - TN = True Negatives
  - FP = False Positives
  - FN = False Negatives

  Which of the formulas below computes the *accuracy* of a binary classifier?
  (TP+TN) /(TP+FP+TN+FN)

- Given the definitions below:
  - TP = True Positives
  - TN = True Negatives
  - FP = False Positives
  - FN = False Negatives

  Which of the formulas below computes the *precision (*or *positive predictive value)* of a binary classifier? **TP/(TP+FP)**

- Given the definitions below:
    - TP = True Positives
    - TN = True Negatives
    - FP = False Positives
    - FN = False Negatives

  Which of the formulas below computes the *specificity* of a binary classifier? **TN/(TN+FP)**

- Why do we *prune* a decision tree? **To eliminate parts of the tree where the decisions could be influenced by random effects**

- A Decision Tree is… **A tree-structured plan of tests on single attributes to forecast the target**
- When training a neural network, what is the *learning rate*? **A multiplying factor of the correction to be applied to the connection weights**

- Which of the following is a strength of the clustering algorithm DBSCAN?
    - **Ability to find cluster with concavities**
    - **Ability to separate outliers from regular data**

- Which of the following is *not* a strength point of DBSCAN with respect to k-means? **The efficiency even in large datasets**

- Which of the following characteristic of data can reduce the effectiveness of K-Means? **Presence of outliers**

- After fitting DBSCAN with the default parameter values the result are: 0 clusters, 100% of noise points. Which will be your next trial?
    - **Reduce the minimum number of object in the neighborhood**
    - **Increase the radius of the neighborhood**

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

- Consider the transactional dataset below

  IDItems
  1 A,B,C
  2 A,B,D
  3 B,D,E
  4 C,D
  5 A,C,D,E

  Which is the *confidence* of the rule A,C => B? **50%**

- Consider the transactional dataset below

  IDItems
  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? **20%**

- Consider the transactional dataset below

  IDItems
  1 A,B,C
  2 A,B,D
  3 B,D,E
  4 C,D
  5 A,C,D,E
  Which is the *confidence* of the rule B => E? **33%**

- When is polynomial regression appropriate? **When the relationship between the predicting variable and the target cannot be approximated as linear**

- Which is the purpose of discretization/discretisation? **Reduce the number of distinct values in an attribute, in order to put in evidence possible patterns and regularities**

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

- Which is the main reason for the *standardization* of numeric attributes? **Map all the numeric attributes to a new range such that the mean is zero and the variance is one**

- What is *Gini Index*? **An impurity measure of a dataset alternative to the *Information Gain* and to the *Misclassification Index***

- Which of the following measure can be used as an alternative to the *Information Gain*? **Gini Index**

- In a decision tree, the number of objects in a node… **is smaller than the number of objects in its ancestor**

- Which of the following is a base hypothesis for a bayesian classifier? **The attributes must be statistically independent inside each class**

- With reference to the total *sum of squared errors* and *separation* of a clustering scheme, which of the statements below is true? **They are strictly correlated, if, changing the clustering scheme, one increases, then the other decreases**

- Which of the statements below is true (one or more)?
  - **Sometimes k-means stops to a configuration which does not give the minimum distortion for the chosen value of the number of clusters**
  - **K-means is quite efficient even for large datasets**
  - **K-means is very sensitive to the initial assignment of the centers**

- Which of the statements below is true (one or more)?
  - **Sometimes DBSCAN stops to a configuration which does not include any cluster**
  - **DBSCAN can give good performance when clusters have concavities**
  - **Increasing the radius of the neighborhood can decrease the number of noise points**
- What is the meaning of the statement: *"the support is anti-monotone"*? **The support of an itemset never exceeds the support if its subsets**

- What is the coefficient of determination R^2? **Provide an index of goodness for a linear regression model**

- What does K-means try to minimize/minimise? **The *distortion*, that is the sum of the squared distances of each point with respect to its centroid**

- Which of the activities below is part of "Business Understanding" in the CRISP methodology? **Which are the resources available (manpower, hardware, software,…)**

- Which of the following statements is *true* (one or more)?
  - **Outliers can be due to noise**
  - **The noise can generate outliers**

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

- What is the *cross validation*? **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**

- Which of the following preprocessing activities is useful to build a Naive Bayes classifier if the independence hypothesis is violated? **Feature selection**

- Which is the main reason for the *MinMax scaling* (also known as "*rescaling*") of attributes? **Map all the numeric attributes to the same range, in order to prevent the attributes with higher range from having prevalent influence**

- Which is the main reason for the *normalization* (also known as "*rescaling*") of numeric attributes? **Map all the numeric attributes to the same range, in order to prevent the attributes (without altering the distribution) with higher range from having prevalent influence**

- Which of the following *is not* an objective of feature selection? **Select the features with higher range, which have more influence on the computations**

- For each type of data choose the best suited distance function:
  - Vector space with real values: **Euclidean Distance**
  - Boolean data: **Jaccard coefficient**
  - Vectors of terms representing documents: **Cosine distance**
  - High dimensional spaces: **Manhattan distance**

- When developing a classifier, which of the following is a symptom of overfitting? **The error rate in the test set is much greater than the error rate in the training set**

- In a decision tree, an attribute which is used only in nodes near the leaves… **gives little insight with respect to the target**

- Which of the statements below about *Hierarchical Agglomerative Clustering* is true? **Requires the definition of *distance between set of objects***

- Match the rule evaluation formulas with their names:
  - $\frac{\sup(A => C)}{\sup(A)}$ **Confidence**
  - $\frac{\sup(A => C)}{\sup(C)}$ **Lift**
  - $\sup(A \cup C) - \sup(A)\sup(C)$ **Leverage**
  - $\frac{1 - \sup(C)}{1 - \sup(A => C)}$ **Conviction**

- In data preprocessing, which of the following ***is not*** an objective of the *aggregation* of attributes? **Obtain a more detailed description of data**

- In data preprocessing, which of the following ***is*** an objective of the *aggregation* of attributes?
  - **Obtain a less detailed scale**
  - **Reduce the variability of data**
  - **Reduce the number of attributes or distinct values**

- Which of the statements below best describes the strategy of Apriori in finding the frequent itemsets? **Evaluation of the support of the itemsets in an order such that uninteresting parts of the search space are pruned as soon as possible**

- In order to reduce the dimensionality of a dataset, which is the advantage of Multi Dimensional Scaling (MDS), with respect to Principal Component Analysis (PCA)? **MDS can be used also with categorical data, provided that the matrix of the distance is available, while PCA is limited to vector spaces**

- Which is different from the others? **Decision tree** [only supervised method - between K-means, Expectation Maximization, Apriori]

- Which is different from the others? **Decision tree** [not a clustering method - between K-means, Expectation Maximization, Dbscan]

- Which is different from the others? **Dbscan** [not a classification method - between SVM, Neural Network, Decision Tree]

- Which is different from the others? **Silhouette Index** [not a index for the evaluation pf purity – between Misclassification Error, Gini Index, Entropy]

- How does *pruning* work when generating frequent itemsets? **If an itemset is not frequent, then none of its superset can be frequent, therefore the frequencies of the supersets are not evaluated.**

- What measure is maximized by the Expectation Maximization algorithm for clustering? **The *likelihood* of a class label, given the values of the attributes of the example**
- The *information gain* is used to **… select the attribute which maximizes, for a given training set, the ability to predict the class value**

- In *data preparation* which is the effect of *normalization*? **Map all the numeric attributes to the same range, without altering the distribution, in order to avoid that attributes with large ranges have more influence**

- Which of the following clustering methods is *not* based on distances between objects? **Expectation Maximization**

- In a dataset with D attributes, how many subsets of attributes should be considered for feature selection according to an exhaustive search? **$O(2^D)$**

- Which is the effect of the *course of dimensionality*? **When the number of dimensions increases the Euclidean distance becomes less effective to discriminate between points in the space**

- Which is the main purpose of *smoothing* in Bayesian classification? **Classifying an object containing attribute values which are missing from some classes in the training set**

- Which of the following characteristic of data can reduce the effectiveness of DBSCAN? **Presence of clusters with different densities**

- Which of the following types of data allows the use of the Euclidean distance? **Point in a vector space**

- Which is the effect of the curse of *dimensionality*? **When the number of dimensions increases the Euclidean distance becomes less effective to discriminate between points in the space**

- What are the hyperparameters of a Neural Network (possibly non exhaustive)? **Hidden layer structure, learning rate, activation function, number of epochs**

- How can we measure the quality of a trained regression model? **With a formula elaborating the difference between the forecast values and the true ones**

- What is the difference between classification and regression? **Classification has a categorical target, while regression has a numeric target**

- In *feature selection*, what is the Principal Component Analysis? **A mathematical technique used to transform a set of numeric attributes into a smaller set of numeric attributes which capture most of the variability in data**

- In a Neural Network, what is the *backpropagation*? **The technique used to adjust the connection weights according to the difference between the desire output and the output generated the network**