**1 - Given the following set of values for the attribute Yearly Income (Euro) {2,651; 1,610; 2,994; 1,667; 2,434; 1,845; 1,570; 2,182; 2,234}. Which of the following statements is/are true?**

**Risposte:**

A- It is possible to discretize the above set of values into 4 intervals when using the unsupervised equal frequency discretization

B- (1570, 2045], (2045, 2519] and (2519, 2994] is a valid unsupervised equal width discretization into 3 intervals (rounded to the nearest integer)

C- [1570, 2045], (1667, 2234) and (2234, 2994] is a valid unsupervised equal frequency discretization into 3 intervals

D- It is not possible to discretize the above set of values into 4 intervals when using the unsupervised equal width discretization

E- (1570, 1667], (1667, 2234] and (2234, 2994] is a valid unsupervised equal frequency discretization into 3 intervals

**2- Which of the following statements about the training dataset is true?**

A- It used only to solve descriptive modeling.

B-It consists of records whose class labels (class values) are assumed to be unknown.

C- It consists of records whose class labels (class values) are known.

D- It used only to solve predictive modeling.

**3- Which of the following statements about the Itarated Hold-out method is/are true?**

A- it estimates performance measures with a bias which is smaller than the bias associated with estimates obtained from the Holdout.

B- It allows to control the number of times a given record is contained in the training and in the test set depending on the class value.

C- it does not allow to control the number of times a given record is contained in the training and in the test set.

D- It significantly improves on Holdout.

E- This could result in a strong bias with specific reference to cases where the data set does not contain dominant records (i.e. outliers).

**4- Assume you are given a dataset consisting of 7 input attributes and one class attribute. You make the decision to apply feature selection, in particular the Brute Force approach. How many possible input attributes subsets do you need to evaluate to select which is the "optimal" one?**

A- 254

B-127

C-186

D- 140

E- 169

**5- Which of the following statements about clustering is/are true?**

a. Feature extraction when applied to clustering is capable of producing features that could be of better use in uncovering the data structure.

b. Given a data set, each clustering algorithm will not produce a partition whether no particular structure in the data exists.

c. Given a data set, each clustering algorithm will produce at least two partitions (clusters) whether no particular structure in the data exists.

d. It is a device for suggesting hypotheses.

e. Given a data set, each clustering algorithm can always produce a partition whether or not there really exists a particular structure in the data.

**6- Assume you got a clustering solution C and that you use external knowledge P to evaluate such a clustering solution. You know the following; 4000 record pairs (x,y) are such that x and y belong to the same cluster of clustering C and to the same category according to external knowledge P, 100 record pairs (x,y) are such that x and y belong to the same cluster of clustering C and to different categories according to external knowledge P, 300 record pairs (x,y) are such that x and y belong to different clusters of clustering C and to the same different category according to external knowledge P, 550 record pairs (x,y) are such that x and y belong to the different clusters of clustering C and to different categories according to external knowledge P. The Jaccard Index is equal to**

Select one:

a. 0.909091 J = a/(a+b+c)

b. -0.453284

c. 0.919192

d. 1.33021

**7- Assume the following contingengy table is given; (cheese, wine)=75, (cheese, NOT wine)=25, (NOT cheese, wine)=325, (NOT cheese, NOT wine)=75. Assume you evaluate the association rule cheese -> wine, then**

Select one or more:

a. the lift of the rule can not be computed

b. the interest factor of the rule is equal to 0.1875

c. the interest factor of the rule is equal to 0.9375

d. the lift of the rule is equal to 0.4875

e. the lift of the rule is equal to 0.9375

1- **Assume you are given a data base containing an attribute Payment Category which can take the following values {excellent, good, average, bad, terrible}. Assume that the first 10 records of the database are associated wirh the following values of the Payment Category attribute; {excellent, good, average, average, average, bad, terrible, excellent, good, average}. Which are the mode of the Payment Category attribute and the associated relative frequency value?**

Select one:

a. (average; 0.4)

b. (bad; 0.1)

c. (terrible; 0.1)

d. (average; 4)

e. (bad; 0.1) and (terrible; 0.1)

2- **Given that TP=100, TN=50, FP=10 and that the number of records used to compute the confusion matrix is equal to 165 we can conclude that**

Select one or more:

a. If FN is equal to 0, then 5 records have missing value for the class attribute.

b. it must be the case that FN=5 only in the case where all the records can be used to query the inducer to compute the confusion matrix (i.e. we have no missing values for the class attribute).

c. it must be the case that FN=5 irrespectively of the records used to compute the confusion matrix.

d. FN must take value on {1, 2, 3, 4, 5}.

3-**The generalization error is**

Select one:

a. the same for any classifier that we learn on a given training data set.

b. the expected error on previously unseen records.

c. computed using the training data set.

d. such that a classifier which achieves the best accuracy on a given training data set is the one which achieves the best generalization.

4- **Which of the following statements about expected cost or cost is/are true?**

Select one or more:

a. It depends on Accuracy.

b. It depends on True Positive.

c. it depends on Error.

d. It does not depend on Error.

e. It does not depend on Accuracy.

5- **Which of the following statements about the Pearson correlation coefficient is/are true?**

Select one or more:

a. it takes values on [0,1].

b. it is equal to 0 when x and y are un-correlated (not necessarily independent), it is equal to 1 in absolute value when x and y are perfectly correlated (dependent).

c. it is typically applied to nominal attributes.

d. it is typically applied to ordinal attributes.

e. it takes values on [-1,+1].

6- **The fundamental problem of cluster validity**

Select one or more:

a. overcomes the computationally intensive nature of the validity paradigm which requires statistical testing to be applied to internal or external indices.

b. tests the following hypothesis “there is no structure on the dataset”.

c. aims to find the true number of clusters hidden in a given dataset.

d. always consists of projecting the data into a 2 or 3 dimensional Euclidean space and to use visualization techniques to get insights on the number of clusters.

7- **Assume the following contingengy table is given; (cheese, wine)=75, (cheese, NOT wine)=25, (NOT cheese, wine)=325, (NOT cheese, NOT wine)=75. Assume you evaluate the association rule cheese -> wine, then**

Select one or more:

a. the confidence of the rule is equal to 0.75

b. the confidence of the rule is equal to 0.25

c. the support of the rule is equal to 0.15

d. the confidence of the rule can not be computed

e. the support of the rule is equal to 0.75

1- **You have applied the procedure for supervised discretization of the Yearly Income attribute. In particular, you made the decision to use Entropy to find the optimal discretization for the Yearly Income attribute. Which of the following statements is/are true?**

Select one or more:

a. Any valid discretization is associated with a non negative value of Entropy.

b. It is possible to obtain a negative value for the Entropy.

c. Achieving the optimal value of Entropy means you are implementing the purest possible discretization solution.

d. The optimal discretization is the one which achieves the maximum value of Entropy.

e. The optimal discretization is the one which achieves the minimum value of Entropy.

2- **Which of the following statements about the classification model is true?**

Select one:

a. The learning algorithm learns it by using the data of the training data set.

b. The learning algorithm learns it by using the data of the test data set.

c. The learning algorithm learns it by using both the data of the training and test data set.

3- **A good classification model**

Select one:

a. must not fit the training set well.

b. must only be concerned with fitting the training set well.

c. must not only fit the training set well, but it must also accurately classify records it has never seen before (test set).

d. must overfit the training set.

e. must underfit the training set.

4- **The wrapper approach to feature selection**

Select one or more:

a. can be applied only when a specific classification model is selected.

b. a classifier is used to find the optimal subset of the available attributes.

c. can be applied also in the case where no classification model is selected.

d. selects attributes before learning the classifier

5- **In the clustering algorithm design or selection step of the Components of Cluster Analysis you are asked to …**

Select one:

a. select ideal features to be used to distinguish patterns belonging to different clusters, immune to noise, and easy to obtain and interpret.

b. provide users with meaningful insights from the original data.

c. determine the proximity measure to be used and construct a criterion function.

d. provide users with a degree of confidence for the clustering results.

6- **Which of the following statements about internal validation measures is/are true?**

Select one or more:

a. Higher values of separation and better.

b. Given the clustering solution C={C1, C2, ..., Ck}, the validity of C is computed as a weighted sum of the validity of individual clusters Ci.

c. Lower values of cohesion and better.

d. The validity function can be cohesion, separation or any combination of them.

7- **Given the following transactions t1={eggs, beer, coke, bread, butter, honey}, t2={beer, onion, water, bread, honey}, t3={eggs, coke, butter, honey, bisquits}, t4={apple, eggs, beer, bread, butter}, you have that ...**

Select one or more:

a. the support for the association rule {onion} -> {water} is equal to 0.75

b. the support for the association rule {eggs, coke} -> {honey} is equal to 0.75

c. the support for the association rule {eggs} -> {beer} is equal to 0.5

d. the support for the association rule {eggs, beer} -> {bread} is equal to 0.5

1- **Assume you are given a data base containing an attribute Food Quality which can take the following values {excellent, very good, good, average, bad, terrible}. Assume you want to apply binarization to the Food Quality attribute, How many new attributes do you need to create to achieve this goal, in case you can accept to induce correlation among them?**

Select one:

a. 2

b. 6

c. 4

d. 3

2- **Which of the following statements is/are true?**

Select one or more:

a. A predictive classification model is a black box that automatically assigns a class label when presented with the attribute set of an unknown record.

b. A descriptive classification model serves as explanatory tool to distinguish between objects of different classes

c. A descriptive classification model can be developed only when you have to predict a binary class attribute

d. A predictive classification model serves as explanatory tool to distinguish between objects of different classes

3- **Assume you learn a classification model to solve a binary classification problem. Assume then that your classifier misses to corrrectly classify 2 instances on the training set and 10 on the test set. (Assume that training and test set both consitst of 100 records, different records)**

Select one:

a. I do not have sufficient information to tell wheter it underfits, overfits or whether it generalizes well.

b. It generalizes well to the test set.

c. The classifier underfits the training data set.

d. The classifier overfits the training data set.

4- **When comparing two classifiers achieving on the same test set different values for Precision and Recall, which of the following statements is true?**

Select one or more:

a. The optimal classifier is the one which maximizes the Recall for the positive class.

b. The optimal classifier is the one which maximizes the Precision for the positive class.

c. To select the "optimal classifier" we must always compute some combination between Precision and Recall.

d. It may happen that to select the "optimal classifier" we must resort to the computation of a function of Precision and Recall.

5- **Complete clustering**

Select one:

a. can only be overlapping i.e. each observation must be assigned to more than one group (cluster).

b. assigns every data object to more than a cluster.

c. assigns every data object to a cluster.

d. assigns every data object to exaclty two clusters.

6- **Which of the following statements is/are true?**

Select one or more:

a. "Cohesion" of a cluster is typically computed as the sum of the weights of the links in the proximity graph that connects points within the given cluster.

b. The "Separation" between two given clusters (C1,C2) is computed to be the sum of weights of the links from points in one cluster (C1) to points in the second cluster (C2).

c. The greater "Separation" is the better the clustering solution.

d. The smaller "Cohesion" is the better the clustering solution.

7- **Given the following transactions t1={eggs, beer, coke, bread, butter, honey}, t2={beer, onion, water, bread, honey}, t3={eggs, coke, butter, honey, bisquits}, t4={apple, eggs, beer, bread, butter}, you have that ...**

Select one or more:

a. the confidence for the association rule {onion} -> {water} is equal to 0.75

b. the confidence for the association rule {eggs, beer} -> {bread} is equal to 1

c. the confidence for the association rule {eggs} -> {beer} is equal to 2/3

d. the confidence for the association rule {eggs, coke} -> {honey} is equal to 0.25

1-**Assume you are given a data base containing the following attribute, Credit Card owner. Assume that Credit Card owner can take values {yes, no}. Assume you have the following 10 records for the Credit Card owner attribute {yes, yes, no, yes, no, no, no, no, no, no}. Assume you want to form a sample contisting of 4 records. Then, you apply random sampling with stratification (stratified sampling) using the equal number option. Which of the following samples is a valid one?**

Select one or more:

a. (no, yes, yes, no)

b. (yes, no, no, no)

c. (yes, yes, yes, no)

d. (yes, yes, no, no)

2- **Which of the following statements about the Inducer is/are true?**

Select one or more:

a. the inducer is obtained by applying the learning algorithm to the test data

b. the inducer is obtained by learning a classification model when using the training data

c. the inducer is obtained by applying the learning algorithm to the training data

d. the inducer is queried on the testing data

e. the inducer is obtained by learning a classification model using the testing data

3-**Assume you have learnt two classifiers (Classifier A and Classifier B) to solve the same binary classification problem, you applied K-folds cross validation and get the following accuracy values 0.95 for Classifier A and 0.93 for Classsifier B.**

Select one:

a. the accuracy achieved on new data by Classifier A is always greater than or equal to the accuracy achieved by Classifier B.

b. Classifier B is not the optimal classifier and thus it must not be used in place of Classifier A when we are asked to predict new data.

c. the accuracy achieved on new data by Classifier A is always greater than the accuracy achieved by Classifier B.

d. with the available information I can not made the decision about which is the optimal classifier, i.e. the one to be used when predicting new data.

e. Classifier A is the optimal classifier and thus it must be used in place of Classifier B when we are asked to predict new data.

4- **Assume using a classifier you learnt to solve a binary classification problem (-1, +1) you got TP=4, FP=32, TN=48 and FN=4. Which of the following statements is true?**

Select one or more:

a. Precision is equal to 0.3333 and Recall is equal to 0.1111.

b. Precision is equal to 0.1111 and Recall is equal to 0.5000.

c. Precision is equal to 0.2543 and Recall is equal to 0.1111.

d. Precision is equal to 0.3333 and Recall is equal to 0.2543.

**5- Which of the following statements about the distance usage is/are true?**

Select one or more:

a. When analyzing sparse data, and asymmetric binary data, similarity measures ignoring 00 matches are typically used.

b. It is useful to consult literature to find one or more similarity measures which have been already used to the application domain we are analyzing.

c. When analyzing dense and continuous data the metric distance measures as the Euclidean distance are typically used.

d. When analyzing dense and continuous data the metric distance measures as the Jaccard distance are typically used.

e. When analyzing sparse data, and asymmetric binary data, the metric distance measures are typically used.

6- **Which of the following statements about external validation measures is/are true?**

Select one or more:

a. The smaller the Gamma Statistics is the better the clustering solution is.

b. The Rand, Jaccard and Fowlkes and Mallows as well as the Gamma Statistics indices takes value over the same interval.

c. The greater the Rand, Jaccard and Fowlkes and Mallows and Gamma Statistics indices are the better the clustering solution is.

d. The Rand, Jaccard and Fowlkes and Mallows indices takes value over the same interval.

7-**Assume the following contingengy table is given; (cheese, wine)=75, (cheese, NOT wine)=25, (NOT cheese, wine)=325, (NOT cheese, NOT wine)=75. Assume you evaluate the association rule cheese -> wine, then**

Select one or more:

a. the φ-coefficient of the rule is equal to 0.375

b. the IS-measure of the rule is equal to 0.375

c. the φ-coefficient of the rule is equal to -0.0625

d. the IS-measure of the rule is equal to -0.3698

1- **Given the following set of values for the attribute Yearly Income (Euro) {2,651; 1,610; 2,994; 1,667; 2,434; 1,845; 1,570; 2,182; 2,234} which are the percentiles of order 1/3 and 2/3?**

Select one:

a. (1,845, 2,994)

b. (2,994; 1,845)

c. (1,667; 2,234)

d. (1,667, 2,994)

2-**Given that TP=200, TN=140, FP=10 and FN=8, the accuracy is**

Select one:

a. 0.9497

b. 0.9794

c. 0.9000

d. 0.8875

e. 0.9611

3- **Assume you have learnt two classifiers (Classifier A and Classifier B) to solve the same binary classification problem, you applied K-folds cross validation and get the following error values 0.02 for Classifier A and 0.05 for Classsifier B. You also computed the Confidence Interval with confidence alpha equal to 0.05 for the true difference in the error (Error Classifier A - Error Classifier B) to obtain (-0.0050982 -0.009018). Which of the following statement is/are true?**

Select one or more:

a. You are sure that the classifier A achieves an error level which is significantly different from the error achieved by Classifier B when the confidence is set to 0.01. (hint, mind the standard normal quantile value)

b. Classifier B achieves an error level which is significantly smaller than the error achieved by Classifier A when the confidence is set to 0.05.

c. Classifier A achieves an error level which is significantly different from the error achieved by Classifier B when the confidence is set to 0.05.

d. Classifier A achieves an error level which is significantly different from the error achieved by Classifier B when the confidence is set to 0.10. (hint, mind the standard normal quantile value)

4- **Which of the following statements about true negative rate is/are true?**

Select one or more:

a. it is the fraction of positive records predicted correctly by a classifier.

b. it is the fraction of negative records predicted correctly by a classifier.

c. It is equal to the ratio TN/(TN+FP) where TN stands for True Negative while FP stands for False Positive

d. It is equal to the ratio TP/(TP+FN) where TP stands for True Positive while FN stands for False Negative

e. it is the number of negative records predicted correctly by a classifier.

5- **Partitional clustering**

Select one:

a. permits clusters to have sub-clusters, in this case the clustering is a set of clusters that are organized as a tree.

b. can only be exclusive, i.e. each observation must be assigned exactly to one group (cluster).

c. can only be overlapping i.e. each observation must be assigned to more than one group (cluster).

d. is a division of the set of data objects (records) into non-overlapping subsets (clusters) such that each data object is in exactly one subset (cluster).

6-**Assume you got a clustering solution C and that you use external knowledge P to evaluate such a clustering solution. You know the following; 4,000 record pairs (x,y) are such that x and y belong to the same cluster of clustering C and to the same category according to external knowledge P, 100 record pairs (x,y) are such that x and y belong to the same cluster of clustering C and to different categories according to external knowledge P, 300 record pairs (x,y) are such that x and y belong to different clusters of clustering C and to the same different category according to external knowledge P, 550 record pairs (x,y) are such that x and y belong to the different clusters of clustering C and to different categories according to external knowledge P. which is the number of records of the dataset associated with the clusterin solution C?**

Select one:

a. 200

b. 50

c. 150

d. 100

7- **Which of the following statements is/are correct?**

Select one or more:

a. A rule achieving a good Support also achieves a good Confidence

b. Given the rule X->Y we have that the higher the confidence the more likely is for Y to be present in transactions that contain X

c. Confidence measures the reliability of the inference made by a rule

d. A rule achieving a good Confidence also achieves a good Support

e. Support is often used to eliminate uninteresting rules

1- **Assume you are given a data base containing an attribute Payment Category which can take the following values {excellent, good, average, bad, terrible}. Assume that the first 10 records of the database are associated wirh the following values of the Payment Category attribute; {excellent, good, average, average, average, bad, terrible, excellent, good, average}. In case you apply random sampling without replacement to form a sample consisting of 5 records, which of the following is/are a valid sample/s?**

Select one or more:

a. (average, average, average, average, average)

b. (good, average, excellent, average, terrible)

c. (good, good, bad, average, good)

d. (good, good, bad, average, terrible)

e. (average, average, excellent, average, average)

2- **Which of the following statements about a confusion matrix and its components is/are true?**

Select one or more:

a. FP+FN represents the number of wrong predictions.

b. TP+TN represents the number of correct predictions.

c. FP+FN+TP+TN represents the number of records of the data set only under the assumption of no missing data.

d. FP+FN+TP represents the number of records of the data set.

e. FP+FN+TP+TN represents the number of records of the data set used to compute the confusion matrix.

3- **Assume you have to solve a binary classification problem. You have a training dataset consisting of 1,000 record. You apply 5-folds cross validation to estimate the accuracy of your classifier. Which of the following is a valid 5-folds partitioning?**

Select one:

a. (201; 199; 200; 200; 200)

b. (199; 199; 199; 199; 204)

c. (200; 200; 200; 200; 200)

d. (199; 199; 199; 201; 202)

4-**Assume you are using a classifier associated with a cumulative gains curve such that the following holds; %Subset size = 10% gives %Positive records = 40% and %Subset size = 30% gives %Positive records = 75%. Which of the following statements is/are true?**

Select one or more:

a. The Lift value associated with %Subset size = 25% must be smaller than or equal to 4.

b. The Lift value associated with %Subset size = 10% is 0.4 while the lift value associated with %Subset size = 30% = 0.25.

c. The Lift value associated with %Subset size = 10% is equal to 4 while the lift value associated with %Subset size = 30% is equal to 2.5.

d. The Lift value associated with %Subset size = 20% must be greater than 5.

e. The Lift value associated with %Subset size = 20% must be smaller than 2.

5- **Which of the following statements about partial clustering is/are true?**

Select one or more:

a. Partial clustering is constrained to assign every data object to at least one cluster.

b. Partial clustering can be used only on continuous attributes.

c. Partial clustering is not constrained to assign every data object to a cluster.

d. The main motivation for partial clustering is that some data objects may not belong to well-defined groups. (noise and/or outliers)

e. Partial clustering is constrained to assign every data object to at least two clusters.

6- **Assume you got a clustering solution C and that you use external knowledge P to evaluate such a clustering solution. You know the following; 4000 record pairs (x,y) are such that x and y belong to the same cluster of clustering C and to the same category according to external knowledge P, 100 record pairs (x,y) are such that x and y belong to the same cluster of clustering C and to different categories according to external knowledge P, 300 record pairs (x,y) are such that x and y belong to different clusters of clustering C and to the same different category according to external knowledge P, 550 record pairs (x,y) are such that x and y belong to the different clusters of clustering C and to different categories according to external knowledge P. The Rand Index is equal to**

Select one:

a. 1.33021

b. 0.909091

c. 0.919192

d. -0.453284

7-**Association rule mining is usually splitted into two phases, which of the following is one of such a phases?**

Select one or more:

a. Rule Generation, which aims to extract all the high-confidence rules from the Frequent Itemsets.

b. Frequent Itemeset Generation, which aims to extract all the high-confidence rules from the Frequent Itemsets.

c. Frequent Itemeset Generation, which aims to find all the Itemsets that satisfy the minsup threshold.

d. Rule Generation, which aims to find all the Itemsets that satisfy the minsup threshold.

1- **Assume you are given a data base containing an attribute Payment Category which can take the following values {excellent, good, average, bad, terrible}. Assume that the first 10 records of the database are associated wirh the following values of the Payment Category attribute; {excellent, good, average, average, average, bad, terrible, excellent, good, average}. In case you apply random sampling with replacement to form a sample consisting of 5 records, which of the following is/are a valid sample/s?**

Select one or more:

a. (good, good, bad, average, good)

b. (good, average, excellent, average, terrible)

c. (average, average, average, average, average)

d. (average, average, excellent, average, average)

e. (good, good, bad, average, terrible)

2- **Given a binary classification problem, which of the following statements about the Accuracy and Error is/are true?**

Select one or more:

a. Accuracy and Error are not related to each other

b. We aim to learn a classification model which achieves high Accuracy and low Error

c. Error is the Accuracy complement to 1, i.e. Error = 1 - Accuracy

d. Accuracy ranges in the [0, 1] interval

e. We aim to learn a classification model which achieves high Accuracy and high Error

3- **Which of the following statements about LOOCV is true?**

Select one:

a. it allows to apply stratified sampling.

b. it is obtained by assuming that each record is a partition of the dataset, thus the value of K equals the number of records of the dataset D.

c. it is faster to apply than a 3-folds cross validation.

d. It is applied when the dataset consists of a huge numebr of records.

4- **The cumulative gains plot**

Select one or more:

a. reports on the Y axis the % positve records and on the X axis the percentage of negative records.

b. reports on the Y axis the % Subset size and on the X axis the percentage of negative records.

c. reports on the X axis the % Subset size and on the Y axis the percentage of negative records.

d. reports on the X axis the % Subset size and on the Y axis the percentage of positive records.

6- **Which of the following statements abour relative indices is/are true?**

Select one or more:

a. The value of K corresponding to the minimum value of the Calinski and Harabasz is taken to be the optimal number of clusters.

b. The value of K corresponding to the minimum of the Minimum Description Lenght is taken to be the optimal number of clusters.

c. The value of K corresponding to the maximum of the Akaike Information Criterion is taken to be the optimal number of clusters.

d. The value of K corresponding to the maximum value of the Dunn index is taken to be the optimal number of clusters.

e. The value of K corresponding to the minimum value of the Davies-Bouldin index is taken to be the optimal number of clusters.

7-**Given the following market basket {eggs, eggs, beer, coke, bread, butter, chips, coke}, it's transaction width is equal to**

Select one:

a. 8

b. 5

c. 6

d. 7

**1- Assume you are given a data base containing an attribute Food Quality which can take the following values {excellent, very good, good, average, bad, terrible}. Assume you want to apply binarization to the Food Quality attribute, How many new attributes do you need to create to achieve this goal, in case you can NOT accept the induced correlation among them?**

Select one:

a. 4

b. 6

c. 5

d. 10

2- **Which of the following statements about a classification model is/are true?**

Select one or more:

a. It outputs the output attribute.

b. It gets in input output attributes.

c. It's input attributes can be nominal attributes.

d. It's input attributes must be continuous. // sbagliata

e. It outputs the explanatory attribute.

f. It gets in input explanatory attributes.

3- **Iterated Hold-out**

Select one:

a. requires less computation time than Holdout.

b. significantly improves on Holdout, it estimates performance measures with a bias which is smaller than the bias associated with estimates obtained from the Holdout.

c. allows to control the number of times a given record is contained in the training and in the test set

d. allows to control the computation time for any given training set and test set.

4- **Assume you learnt a classifier to predict whether a customer is loyal or not. Assume you used a data set consisting of 10,000 records where 9,900 are associated with loyal customers. Furthermore, assume you applied 10-folds cross validation and estimated the accuracy of your classifier to be 0.99. What can be said about your classifier?**

Select one or more:

a. I can not make any statement about its usefulness with such a limited information about the problem, data set and estimated accuracy value.

b. It is an excellent classifier but nothing can be told about its effectiveness to predict which new customers are loyal and which are not.

c. It is an excellent classifier and will be effective to predict which new customers are loyal and which are not.

d. It is a poor classifier and will be ineffective to correctly predict which new customers are loyal and which are not.

5- **Which of the following statements about clustering is/are true?**

Select one or more:

a. The goal of clustering is that of grouping objects (clustering) in such a way that the objects within a group (cluster) are similar (related) to one another and different from (unrelated) the objects in the other groups (clusters).

b. The greater the similarity (or homogeneity) within a group (cluster) and the greater the difference between groups (clusters), the better or more distinct the clustering.

c. Clustering is not useful to understand classes, or conceptually meaningful groups (clusters) of objects that share common characteristics.

d. Clustering is useful only for summarizing objects.

6- **The Silhouette index for a given record is defined as follows s=(b-a)/max(a,b), which of the following statements is/are true?**

Select one or more:

a. The Silhouette index takes value in [-1,1]

b. the term "a" represents the average distance of the considered record to all other recordss in its cluster.

c. The Silhouette index is always greater than 0.5.

d. The Silhouette index can not be negative.

e. the term "b" represents the minimum of the average distances of the considered record to all the records in each given cluster different from the cluster to which the considered record belongs to.

7- **An itemset**

Select one or more:

a. is said to be an empty set whether it contains no items

b. is said to be a 4-itemset whether it consists of at least 4 items

c. is said to be a 4-itemset whether it consists of at the most 4 items

d. is a collection of items (repetitions of the same item are allowed)

e. is a collection of items (no repetitions of the same item are allowed)

1- **Which of the following statements about standardization and normalization is/are true?**

Select one or more:

a. They can be both applied to continuous attributes.

b. Standardization can be applied only to non negative continuous attributes.

c. They can be both applied to categorical attributes.

d. Normalization can be applied only to non negative continuous attributes.

e. Standardization transform the processed attribute in such a way that the resulting attribute has zero mean and unit standard deviation.

f. Standardization transform the processed attribute in such a way that the resulting attribute has zero standard deviation and unit mean.

**2- Which of the following statements about the test dataset is/are true?**

Select one or more:

a. the inducer is learnt on testing dataset

b. the test dataset is always 1/3 of train dataset

c. the test dataset is used to have an estimate of the performance of the classifier learnt on the training dataset

d. we assume its instances are unknown

3- **Assume you have a dataset consisting of 1,000 records. You want to solve a binary classification problem and thus apply 5-folds cross validation. You got the following confusion matrices from the first up to the fifth folds; (TP=90, FP=10, TN=85, FN=15), (TP=83, FP=12, TN=95, FN=10), (TP=98, FP=22, TN=70, FN=10), (TP=91, FP=4, TN=101, FN=4), and (TP=100, FP=2, TN=92, FN=6). Which is the Accuracy of the trained classifier?**

Select one:

a. we do not have enough information to compute the accuracy

b. 0.925

c. 0.894

d. 0.915

e. 0.905

4- **Feature selection**

Select one or more:

a. may apply all possible subsets of the available attributes as the input attributes to the Classification Model to find irrelevant and redundant attributes.

b. ensures we find the optimal set of attributes.

c. is applied to remove redundant and irrelevant attributes.

d. always applies all possible subsets of the available attributes as the input attributes to the Classification Model to find irrelevant and redundant attributes.

e. can be applied only to quantitative attributes.

5- **Which of the following statements about Cosine Similarity is/are true?**

Select one or more:

a. cos(x,y) is equal to 0 when magnitudes of x and y are the same while it is equal to 1 when x and y share no terms (maximum difference).

b. cos(x,y) is equal to 0 when x and y are the same (except for magnitude) while it is equal to 1 when x and y share no terms (maximum difference).

c. it is able to handle non-binary attributes.

d. It can not be used for comparing sparse records.

e. it ignores 00 matches like the Jaccard Coefficient.

6- **Which of the following statements about external or supervised evaluation measures for cluster analysis is/are true?**

Select one:

a. They determine how closely related the objects in a cluster are using only the dataset.

b. They determine how distinct or well-separated a cluster is from other clusters using only the dataset.

c. They measure the extent to which the clustering structure discovered by a clustering algorithm matches some external structure.

d. They measure the goodness of a clustering structure without respect to external information.

7-**Given the following transactions t1={hard drive, external drive, mouse, monitor, windows OS}, t2={monitor, mouse, hard drive, windows OS}, t3={tablet, SD card, windows OS, mouse}, t4={hard drive, pendrive, tablet, apple OS}, you have that ...**

Select one or more:

a. the support count for the itemset (mouse, hard drive, windows OS) is equal to 3

b. the support count for the itemset (SD card, tablet) is equal to 3

c. the support count for the itemset (windows OS, mouse) is equal to 3

d. the support count for the itemset (monitor, hard drive, windows OS, mouse) is equal to 2

1- **Assume you have a dataset consisting of 100,000 records and 1,000 continuous attributes. You know that the probability of missing value for each attribute is equal to 0.005. Furthermore, assume that the missing mechanism is at random, i.e. that each attribute value can be missing or not and this does not influence in any way the missingness of other attributes. Which of the following statements is/are true?**

Select one or more:

a. The exact number of records containing at least one missing value is equal to 100 (rounded down to the nearest integer)

b. The expected number of records containing at least one missing value is equal to 100 (rounded down to the nearest integer)

c. The probability that a record contains no missing values is equal to 0.000995 (six digits representation)

d. The exact number of records containing at least one missing value is equal to 665 (rounded down to the nearest integer)

e. The probability that a record contains no missing values is equal to 0.006654 (six digits representation)

f. The expected number of records without any missing value is equal to 665 (rounded down to the nearest integer)

2- **Given that TP=200, TN=50, FP=50 and that the number of records used to compute the confusion matrix is equal to 323 we can conclude that**

Select one or more:

a. it must be the case that FN=23 only in the case where all the records can be used to query the inducer to compute the confusion matrix (i.e. we have no missing values for the class attribute).

b. If FN is equal to 10, then 13 records have missing value for the class attribute.

c. FN must take value in [0, 10]

d. it must be the case that FN=23 irrespectively of the records used to compute the confusion matrix.

3- **A classifier is said to be scalable …**

Select one:

a. if it decreases its accuracy when more training data are available.

b. if it is capable to learn from huge amount of data.

c. if it is capable to increase its accuracy when more training data are available.

d. if it achieves a stable accuracy, no matter how many training data is available.

4- **Assume using a classifier you learnt to solve a binary classification problem (-1, +1) you got TP=4, FP=32, TN=48 and FN=4. Which of the following statements is true?**

Select one or more:

a. It could happen that Recall can not be computed, i.e. it is undefined.

b. Recall can be computed only when TP+TN is not zero.

c. Precision can always be computed, i.e. it is always well defined.

d. It could happen that Precision can not be computed, i.e. it is undefined.

5- **Assume the similarity between two objects can take value between 1 (minimum similarity) and 6 (maximum similarity). You are asked to transform the given similarity measure to obtain a transformed similarity which takes values on the interval [0,1]. Which is the value of the transformed similarity for a pair of objects which have similarity equal to 2?**

Select one:

a. 1/5

b. 1/6

c. 4/10

d. 3/10

e. 1/4

6- **Which of the following statements about clustering algorithms, cluster analysis and cluster validation is/are true?**

Select one or more:

a. The issue of finding the “correct number of clusters” is not an issue for cluster validation.

b. In cluster analysis, cluster evaluation is a well-developed and clear step.

c. One of the main issues in cluster validation is distinguishing whether non-random structure actually exists in the data.

d. A clustering algorithm is designed to find clusters and thus it will find clusters in a data set, even if that data set has no natural cluster structure.

7- **Given the following transactions t1={hard drive, external drive, mouse, monitor, windows OS}, t2={monitor, mouse, hard drive, windows OS}, t3={tablet, SD card, windows OS, mouse}, t4={hard drive, pendrive, tablet, apple OS}, you have that ...**

Select one or more:

a. the support for the association rule {tablet} -> {SD card} is equal to 0.75

b. the support for the association rule {monitor, mouse} -> {hard drive} is equal to 0.5

c. the support for the association rule {monitor} -> {mouse} is equal to 0.5

d. the support for the association rule {windows OS} -> {mouse} is equal to 0.75

1- **The main advantages of aggregation are … Choose one or more of the following statements**

Select one or more:

a. Maintain the scale and scope focus, it avoids us to make the decision whether to use high-level or low-level views of data.

b. Smaller data sets; we need less memory and processing time, thus more time consuming and possibly more effective algorithms can be used.

c. Increased attention to details; we increase our focus to specific aspect and details of the dataset.

d. Reduced variance; attributes computed on aggregated records are more stable than those associated with the native records.

2- **A predictive classification model**

Select one:

a. Predicts the class of an unknown record. It can be treated as a black box that automatically assigns a class label when presented with the attribute set of an unknown record.

b. Can be developed only when you have to predict a binary class attribute

c. Can serve as explanatory tool to distinguish between objects of different classes

3- **In cross validation we have that**

Select one or more:

a. the classification algorithm undergoes K training-testing phases, while the K estimates of the classifier’s accuracy used to compute the arithmetic mean.

b. the classification algorithm undergoes K-1 training-testing phases, while the K-1 estimates of the classifier’s accuracy used to compute the arithmetic mean.

c. the dataset is partitioned into K disjoint subsets each of which always consists of exaclty the same number of records.

d. the dataset is partitioned into K disjoint subsets, exhaustive (a partition of the dataset) and with almost a constant number of records.

e. each record of the dataset is included into the training sets the same number of times and exactly one time in the test set.

4- **ANOVA, t-test and mutual information**

Select one or more:

a. ignore that attributes can be dependent.

b. are independent on the classifier

c. are univariate and parametric filters.

d. together with Relief are univariate filters.

e. depend on the classifier.

5- **The simple matching coefficient is such that**

Select one or more:

a. counts both presences and absences not in the same way, i.e. it is an asymettric measure.

b. counts presences and absences in the same way as the Jaccard coefficient.

c. takes values on [0,1].

d. it is equal to 0 when the value x of the considered attribute for the first record is different from the value y of the considered attribute for the second record while it is equal to 1 when the values x and y are the same.

6- **Relative index/ices for probabilistic mixture model-based clustering is/are**

Select one or more:

a. Akaike Information Criterion.

b. Davies-Bouldin index.

c. Dunn index.

d. Bayesian Information Criterion.

7- **An objective interestingness measure is**

Select one or more:

a. domain-independent and requires minimal input from the users (threshold only)

b. computed based on the frequency counts tabulated in a Contingency Table

c. applied only to itemsets exceeding a given support count

d. a model-driven approach for evaluating the quality of association patterns

1- **Given the following set of values for the attribute Yearly Income (Euro) {2,651; 1,610; 2,994; 1,667; 2,434; 1,845; 1,570; 2,182; 2,234} which are the values of the mean and trimmed mean?**

Select one:

a. (2,133; 2,090)

b. (2,133; 2,182)

c. (2,434; 2,090)

d. (2,111; 2,234)

2- **A descriptive classification model**

Select one:

a. Predicts the class of an unknown record. It can be treated as a black box that automatically assigns a class label when presented with the attribute set of an unknown record.

b. Can be developed only when you have to predict a binary class attribute

c. Can serve as explanatory tool to distinguish between objects of different classes

3- **Which of the following statements about accuracy is/are true?**

Select one or more:

a. A good indicator of the accuracy achieved by a classifier is represented by the percentage of test records which are correctly classified. // giusta anche questa

b. Accuracy measures the capability of the classification model to give reliable predictions on new records (i.e. records not available when the model was developed).

c. Accuracy is the complement to one of Error, i.e. Accuracy = 1 - Error

d. Accuracy allows to “select” the instance of the classification model which likely provide the “best prediction performance” on new records.

**4- Assume you learnt two classifiers and compare them on a given test data set. In particular you got the following confusion matrix for classsifier A {TN=900, FP=50, FN=30, TP=100} and the following confusion matrix for classifier B {TN=900, FP=30, FN=30, TP=120}. Which is the following statements is/are true?**

Select one or more:

a. The optimal classifier is always Classifier A, no matter whether we evaluate on Accuracy or on Expected Cost (or simply cost).

b. The optimal classifier when evaluating on Expected cost (or simply cost) using the following cost matrix {0, 1, 80, 2} is classifier B.

c. The optimal classifier when evaluating on Accuracy is classifier B.

d. The optimal classifier when evaluating on Accuracy is classifier A.

e. The optimal classifier when evaluating on Expected cost (or simply cost) using the following cost matrix {0, 1, 80, 2} is classifier A.

5- **Which of the following statements about the Mahalanobis Distance is/are true?**

Select one or more:

a. it is equal to 0 when x and y are the same.

b. it takes values on [0,+Infinite].

c. it takes values on [0,+1].

d. it is typically used when attributes are correlated and have different different variances, and their distribution is approximately bivariate Gaussian.

e. it takes values on [-1,+1].

6- **When two clusters are relatively cohesive but not well separated which is an appropriate thing to do?**

Select one:

a. split the two clusters into sub-clusters.

b. remove some records of the two clusters.

c. merge them into a single cluster.

d. define a new clustering solution by excluding the two clusters.

7- **Which of the following statements is/are correct**

Select one or more:

a. odds ratio and Kappa measures may not be suitable for analyzing asymmetric binary attributes

b. cosine and jaccard measures are preferred for non sparse data

c. interest, cosine and jaccard measures are preferred for asymmetric attributes

d. Jaccard measure can not be used other than binary attributes

1- **Assume you are given a data base containing the customer data, in particular the following attributes; Name, Age, Purchase date, Yearly Income, Credit Card owner, and Payment Category. Assume that Credit Card owner can take values {yes, no}, Payment Category can take values {excellent, good, average, bad, terrible}. Which of the following statements is/are true?**

Select one or more:

a. It is possible to compute mean and variance for "Payment Category"

b. It is possible to compute the correlation between "Age" and "Yearly Income"

c. It is possible to compute the contingency correlation between "Age" and "Yearly Income"

d. It is possible to compute the mode of the attribute "Credit Card owner"

e. It is possible to compute percentile for "Payment Category"

f. It is possible to compute the mean of the attribute "Credit Card owner"

g. It is possible to compute rank correlation between "Payment Category" and "Credit Card owner"

2- **Given that TP=200, TN=140 and that the number of records used to compute the confusion matrix is equal to 380 we can conclude that**

Select one or more:

a. FN+FP=40 only if no missing records for the class attribute are included into the 380 records.

b. FN = 10 and FP = 30.

c. FN = 40.

d. If out of the 380 records we have 10 missings values for the class attribute, then FN+FP=30.

**3- When applying LOOCV, the accuracy on each test folder (single record) is**

Select one:

a. always equal to the probability of the most probable class.

b. can be used only when solving binary classification problems.

c. it can take any value in [0, 1]

d. either equal to 1 or to 0

4- **Which of the following statements about Precision is/are true?**

Select one or more:

a. It determines the fraction of records that actually turns out to be positive in the group the classification model has declared as a positive class.

b. For a given number of True Positive, the higher the Precision is, the lower the number of False Positive errors committed by the Classification Model.

c. For a given number of True Positive, the higher the Precision is, the higher the number of False Positive errors committed by the Classification Model.

d. For a given number of True Negative the lower the Precision is, the lower the number of False Positive errors committed by the Classification Model.

e. For a given number of True Positive, the lower the Precision is, the lower the number of False Positive errors committed by the Classification Model.

5- **Distances are dissimilarities with certain properties and they are to be used when the considered attributes are numeric. Which of the following statements about distances is/are true?**

Select one or more:

a. Minkowski, Euclidean and Supremum distances are special cases of the Manhattan distance.

b. Manhattan, Euclidean and Supremum distances satisfy non negativity, simmetry but not necessarily the triangle inequality.

c. Similarity does not satisfy the triangle inequality property but symmetry and non negativity are typically satisfied.

d. Manhattan, Euclidean and Supremum distances satisfy non negativity, simmetry and triangle inequality.

e. Manhattan, Euclidean and Supremum distances are special cases of the Minkowski distance.

6-**Which of the following statements about internal or unsupervised evaluation measures for cluster analysis is/are true?**

Select one or more:

a. They are based on cohesion measures or on separation measures.

b. They compare different clusterings or clusters. It can be supervised or unsupervised.

c. They can be of relative type and thus it compares different clusterings or clusters according to supervised or unsupervised methods.

d. They measure the goodness of a clustering structure without respect to external information.structure.

7- **Assume you have the following content of a market basket {apple, beer, chips, beer, bread}. In case we take into account the number of occurences of an item, which of the following representations is/are valid for the given market basket?**

Select one:

a. apple=1, banana=0, orange=0, water=0, beer=1, coke=0, chips=1, bread=1, cheese=1

b. apple=1, banana=0, orange=0, water=0, beer=1, coke=0, chips=1, bread=1, cheese=0

c. apple=1, banana=0, orange=0, water=0, beer=1, coke=1, chips=1, bread=1, cheese=0

d. apple=1, banana=0, orange=0, water=0, beer=2, coke=0, chips=1, bread=1, cheese=0

e. apple=1, banana=0, orange=1, water=0, beer=1, coke=0, chips=1, bread=1, cheese=0

6-**The Validity Paradigm is such that ...**

Select one or more:

a. uses only external indices.

b. a common hypothesis is the Random Position Hypothesis, i.e. all the locations of the available records in some specific region of a n-dimensional space are equally likely.

c. tests the following hypothesis “there is no structure on the dataset”.

d. a common hypothesis is Random label Hypothesis, i.e. all permutations of "m" records are equally likely.

5-**Which of the following statements about Proximity/Similarity is/are true?**

Select one or more:

a. Proximities are defined or transformed to take values on [0,1], in such a way to use a scale in which proximity indicates the fraction of similarity (dissimilarity) between two records.

b. The similarity between two objects (records) is a numerical measure of the degree to which the two objects (records) are alike.

c. Similarity is usually positive and is often between 0 (no similarity) and 1 (complete similarity).

d. Similarity is usually non-negative and is often between 0 (no similarity) and 1 (complete similarity).

e. The similarity is lower for pairs of objects (records) that are more alike.

5- **Which of the following statements about notions of a cluster is/are true?**

Select one or more:

a. According to graph-based clusters, nodes are objects and links represent connections among objects. A cluster is a connected component (group of objects that are connected to one another but have no connection to objects outside the group).

b. According to well separated clusters, a cluster is a dense region of objects that is surrounded by a region of low density.

c. According to density-based cluster, a cluster is a dense region of objects that is surrounded by a region of low density.

d. According to density-based clusters, given a cluster, each object in the cluster is closer (more similar) to every other object in the same cluster than to any object not in the cluster.

e. According to well separated clusters, given a cluster, each object in the cluster is closer (more similar) to every other object in the same cluster than to any object not in the cluster.

2- **Given a binary classification problem, which of the following statements about the Confusion Matrix is/are true?**

Select one or more:

a. False negative means that the Inducer predicted a negative class (-1) while the actual class is positive (+1)

b. False positive means that the Inducer predicted a negative class (-1) while the actual class is positive (+1)

c. False negative means that the Inducer predicted a positive class (+1) while the actual class is negative (-1)

d. The main diagonal elements are the counts of true positive and true negative records

e. False positive means that the Inducer predicted a positive class (+1) while the actual class is negative (-1)

4- **When costs coefficients are symmetric**

Select one or more:

a. the cost is proportional to accuracy.

b. cost and accuracy are not related to eahc other.

c. a classifier which maximizes accuracy, minimizes the expected cost (or cost).

d. the optimal classifier selected using the expected cost (or cost) is different from the optimal classifier selected using the accuracy.

4-**Which of the following statements about ROC is/are true?**

Select one or more:

a. It may happen that a classifier is the optimal one for a given range of %FP and another one is the optimal one for another range of %FP.

b. The ROC curve shows the behaviour of the performance of a classifier without regard to class distribution or error costs.

c. The ROC curve is not related to the Cumulative Gains.

d. The ROC curve is always a smooth line and never shows stairs like behaviours.

e. We are sure to find a classifier whose ROC curve dominates the ROC curves of all other classifiers (for each value of %FP the value of its' %TP is greter than or equal to that of any other classifier)

7- **Given the following transactions t1={eggs, beer, coke, bread, butter, honey}, t2={beer, onion, water, bread, honey}, t3={eggs, coke, butter, honey, bisquits}, t4={apple, eggs, beer, bread, butter}, you have that ...**

Select one or more:

a. the support count for the itemset (butter, eggs) is equal to 3

b. the support count for the itemset (beer, bread, honey) is equal to 2

c. the support count for the itemset (butter, honey) is equal to 2

d. the support count for the itemset (beer, bread, honey) is equal to 3

3- **Assume you have learnt two classifiers (Classifier A and Classifier B) to solve the same binary classification problem. You computed the Confidence Interval for the true difference in the error (Error Classifier A - Error Classifier B). Which of the following statements is/are true?**

Select one or more:

a. If the upper limit of the confidence interval is smaller than 0, then Classifier B is better than Classifier A with at the associated level of confidence.

b. If the upper limit of the confidence interval is smaller than 0, then Classifier A is better than Classifier B with at the associated level of confidence.

c. If the Confidence Interval spans the value 0, we conclude that the observed difference in erros is not Statistically Significant and thus we can not tell which is the optimal classifier.

d. If the lower limit of the confidence interval is greater than 0, then Classifier A is better than Classifier B with at the associated level of confidence.

e. If the lower limit of the confidence interval is greater than 0, then Classifier B is better than Classifier A with at the associated level of confidence.

6- **Which of the following statements about the SIlhouette index is/are true?**

Select one or more:

a. An overall measure of goodness of a clustering can be obtained by computing the Average Silhouette index of all records.

b. We can compute the Average Silhouette index of a cluster by simply taking the average of the Silhouette index of records belonging to the considered cluster.

c. The Silhouette index is defined for hierarchical clustering while for partitional clustering a different evaluation measure is used.

d. A positive value for the Silhouette index means that the average distance to records in its cluster is greater than the minimum average distance to records in another cluster.

6- **You are must select the optimal number of clusters K using the Davies-Bouldin index. You are given the following set of pairs (2, 3), (3, 2), (4, 1), (5, 3), (6, 6), (7, 6), (8, 5) where the first element of the pair represents the number of clusters while the second represents the associated value of the Davies-Bouldin index. Which is the optimal number of clusters K?**

Select one:

a. 6

b. 4

c. 6 and 7

d. 7

2- **Assume you are given a data base containing the customer data, in particular the following attributes; Cutomer Name, Age, Purchase date, Yearly Income, Credit Card owner, Payment Category and Profitable. Assume you want to predict the value of the Profitable attribute which can take the following values {yes, no}. Which of the following statements is/are true?**

Select one or more:

a. "Credit Card owner" can not be used as an input attribute

b. "Yearly Income" can not be used as an input attribute

c. The attribute "Customer Name" is a valid input attribute to predict the "Profitable" attribute (mind that you learn a model on the training set to predict on new records, i.e. different customers)

d. The attribute "Profitable" is usually called the class attribute or output attribute

e. The attribute "Age" is usually called the explanatory attribute or input attribute

1- **Assume you are given a data base containing an attribute Payment Category which can take the following values {excellent, good, average, bad, terrible}. Assume that the first 10 records of the database are associated with the following values of the Payment Category attribute; {excellent, good, average, average, average, bad, ?, excellent, good, average}, where "?" stands for missing value. In case you apply mode replacement which is the value used to replace "?"?**

Select one:

a. bad

b. average

c. equally "bad" or "terrible"

d. we can not apply mode replacement to the Payment Category attribute

e. terrible

3-**Which of the following statements about the Hold-out method is/are true?**

Select one or more:

a. consists of limiting the data which are used to learn the classifier, i.e. some records are saved (hold back) to estimate the reliability level of the classifier for the task to be solved.

b. gives always a reliable accuracy estimate.

c. can over(under)estimate the true value of the classification accuracy.

d. gives an accuracy estimate which does not depend on the choice of the test set.

e. gives an accuracy estimate which depends on the choice of the test set.

7- **Given the following transactions t1={hard drive, external drive, mouse, monitor, windows OS}, t2={monitor, mouse, hard drive, windows OS}, t3={tablet, SD card, windows OS, mouse}, t4={hard drive, pendrive, tablet, apple OS}, you have that ...**

Select one or more:

a. the confidence for the association rule {monitor, mouse} -> {hard drive} is equal to 1

b. the confidence for the association rule {tablet} -> {SD card} is equal to 0.75

c. the confidence for the association rule {windows OS} -> {mouse} is equal to 0.75

d. the confidence for the association rule {monitor} -> {mouse} is equal to 1

1- **Assume you are given a data base containing the customer data, in particular the following attributes; Name, Age, Purchase date, Yearly Income, Credit Card owner, and Payment Category. Assume that Credit Card owner can take values {yes, no}, Payment Category can take values {excellent, good, average, bad, terrible}. Which of the following statements is/are true?**

Select one or more:

a. "Yearly Income" and "Age" are interval attributes

b. "Payment Category" is a categorical attribute

c. The only binary attribute is "Credit Card owner"

d. "Yearly Income" is a ratio attribute

e. "Yearly Income" and "Age" are quantitative attributes, while "Purchase date" is a categorical attribute

f. "Credit Card owner", "Name" and "Payment Category" are all nominal attributes

g. "Payment Category" is a nominal attribute