Data Mining Quiz

Total points 25/99

Email address *

ranjeeth2197@msitprogram.net

Select all from the following attributes which are quantitative 0/3

□ Time in terms of AM or PM

☑ Brightness as measured by a light meter

□ Brightness as measured by people's judgement

□ Angles as measured in degrees between 0 and 360

□ Bronze, Silver, and Gold medals as awarded at the Olympics

□ Height above sea level

☑ Number of patients in a hospital

□ ISBN numbers for books. (Look up the format on the Web.)

□ Ability to pass light in terms of the following values: opaque, translucent' transparent

□ Military rank

□ Distance from the centre of campus

☑ Density of a substance in grams per cubic centimetre

□ Coat check number. When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave

Consider the interestingness measure, M = (P(B|A)−P(B))/(1−P(B)) , for an association rule A → B. The measure attains its maximum value when 0/2

○ P(B|A) = 1

○ P(B|A) = P(B)

○ P(B) = 0

● P(B/A) = 0

Classification algorithms employ usually which of the following strategy 1/1

○ Dynamic

○ Branch and Bound

○ Back Tracking

● Greedy

x = 0101010001 y = 0100011000 Jaccard Similarity =____(write exact number) 0/2

4

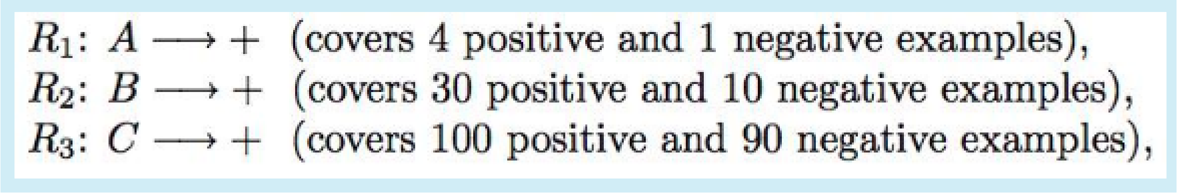
In market basket analysis, let c1, c2, and c3 be the confidence values of the rules {p} –→ {q}, {p} –→ {q, r}, and {p, r} –→ {q}, respectively. If we assume that c1, c2, and c3 have different values, which one be the lowest confidence among these three? 0/2

● C2

○ C1

○ C3

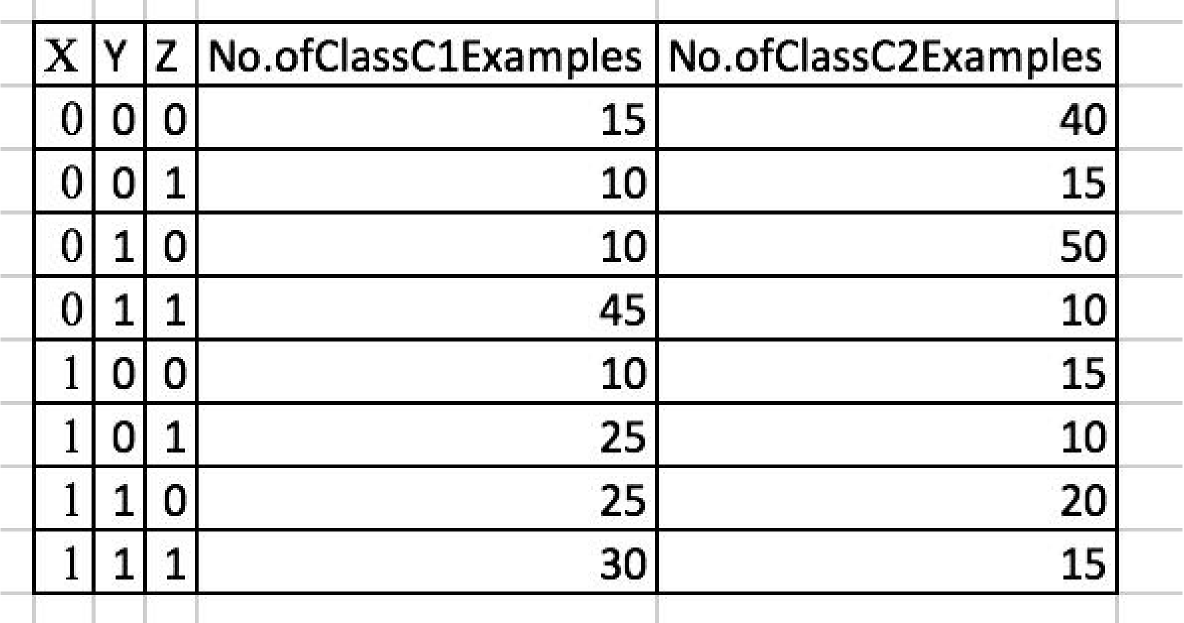
Consider a training set that contains 50 positive examples and 500 negative examples. For each of the following candidate rules and Match the following



R_1: $A \longrightarrow +$ (covers 4 positive and 1 negative examples),
R_2: $B \longrightarrow +$ (covers 30 positive and 10 negative examples),
R_3: $C \longrightarrow +$ (covers 100 positive and 90 negative examples),

	R1	R2	R3	Score
Best rule according to Rule accuracy	●	○	○	1/1
Best rule according to FOIL's information gain	○	●	○	0/1
Worst rule according to Rule accuracy	○	●	○	0/1
Worst rule according to FOIL's information gain	○	○	●	0/1

Compute a two-level decision tree using the greedy approach. Use the classification error rate as the criterion for splitting. Match the following error rates at root node of the induced tree



X	Y	Z	No.ofClassC1Examples	No.ofClassC2Examples
0	0	0	15	40
0	0	1	10	15
0	1	0	10	50
0	1	1	45	10
1	0	0	10	15
1	0	1	25	10
1	1	0	25	20
1	1	1	30	15

	0.405797101	0.31884058	0.449275362	Score
X	○	○	●	0/1
Y	●	○	○	0/1
Z	○	●	○	1/1

Given a similarity measure with values in the interval [0,1], select all valid ways from the following to transform this similarity value into a dissimilarity value in the interval [0,∞] 0/2

□ d = (1−s)/s

☑ d = − log s

□ d= log s

□ d= (s-1)/s

Select all the true statements in the following 0/4

☑ The dimensionality of PCA or SVD can be viewed as a projection of the data onto a reduced set of dimensions

□ In aggregation, groups of dimensions are combined

☑ The aggregation can be viewed as a change of scale

☑ The dimensionality reduction provided by PCA and SVD do not have any interpretation with respect to scaling of variables

□ Meaningful aggregation may not be possible but PCA and SVD are always possible

☑ Meaningful aggregation is always possible but PCA and SVD may not be possible

If x = (1, 1, 1, 1) and y = (2, 2, 2, 2), Euclidean(x, y) =____ 2/2

2

Select only True statements for the following activities to be a data mining task. 0/4

☑ Dividing the customers of a company according to their gender

☑ Dividing the customers of a company according to their profitability

☐ Computing the total sales of a company

☑ Predicting the future stock price of a company using historical records

☑ Sorting a student database based on student identification numbers

☑ Predicting the outcomes of tossing a (fair) pair of dice

☑ Monitoring the heart rate of a patient for abnormalities

☑ Monitoring seismic waves for earthquake activities

☐ Extracting the frequencies of a sound wave

☐ Option 10

Select all statements from the following which are false? 3/3

☐ Text files can be easily inspected by typing the file or viewing it with a text editor

☑ Binary files are more portable than text files, both across systems and programs

☐ Text files can be more easily modified, for example, using a text editor

If x = (0, 1, 0, 1) and y = (1, 0, 1, 0), corr(x, y)= ____ 0/2

-3

Consider the following set of frequent 2-itemsets: {1, 2}, {1, 3},{2, 3},{3, 4}. Assume that there are only four items in the data set. Select all the candidate 3-itemsets that survive the candidate pruning step of the Apriori algorithm 0/6

○ {1, 3, 4}

● {1, 2, 4}

○ {1, 2, 3}

○ {2, 3, 4}

State true or false the following statement. “The goal of both tasks regression and classification is to learn a model that minimizes the error between the predicted and true values of a target variable” 0/1

○ True

● False

Suppose we have market basket data consisting of 100 transactions and 20 items. If the support for item a is 25%, the support for item b is 90% and the support for itemset {a, b} is 20%. Let the support and confidence thresholds be 10% and 60%, respectively. The confidence percentage of the association rule {a}→{b} is ___ 0/2

17

Select all measures from the following which are used to evaluate quality of a classification rule 0/1

☑ Accuracy

☑ Length of the rule (number of conditions)

☑ Coverage

Which of the following are the steps of the pre-processing of data? a. fusing data from multiple sources b. Visualizing the data c. cleaning data to remove noise and duplicate observations d. selecting records and features that are relevant to the data mining task at hand. 1/1

○ a b c d

○ a b c

● a c d

○ b c d

Which of the following are called datamining tasks?a. To scour large databases in order to find novel and useful patterns that might otherwise remain unknown. b. To predict the outcome of a future observation, such as predicting whether a newly arrived. customer will spend more than $100 at a department store c. Looking up individual records using a database management system or finding particular Web pages via a query to an Internet search engine

3/3

○ All three

● Only a and b

○ Only a

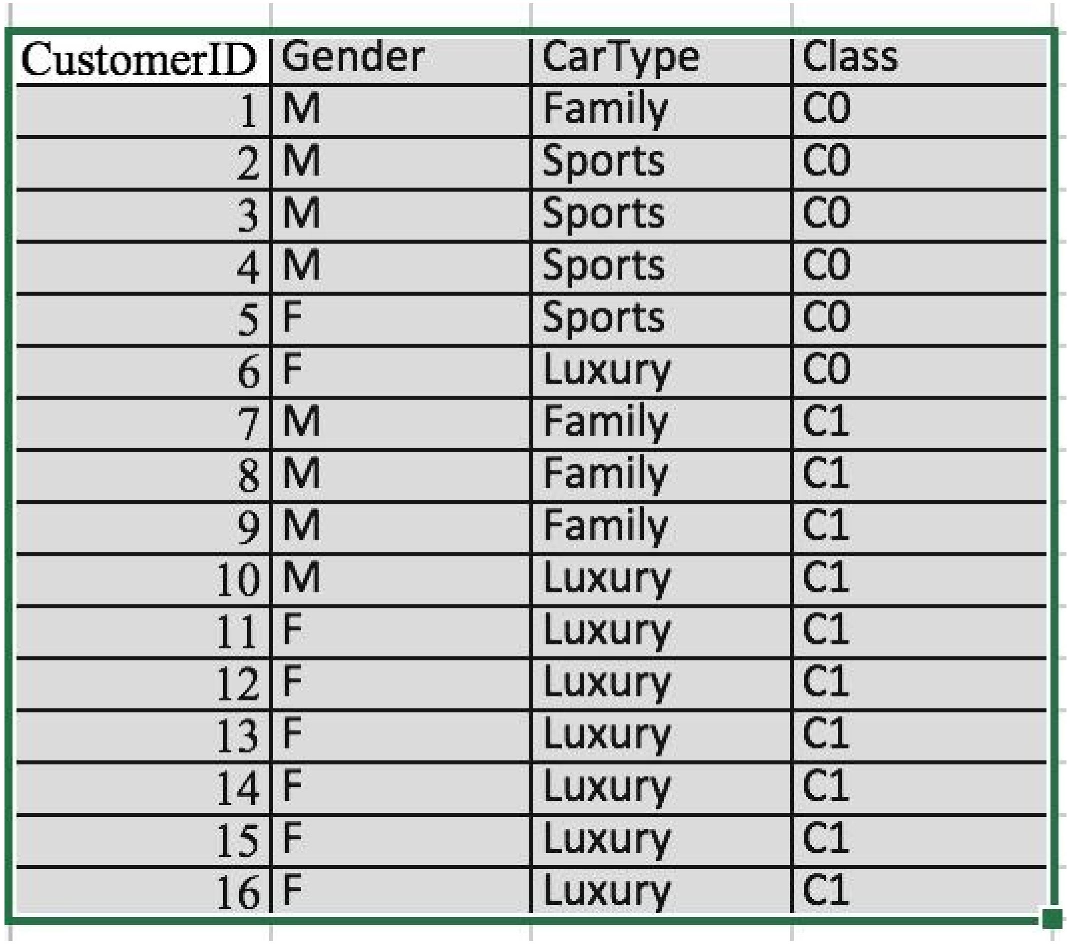
○ Only b and c

Select all from the following attributes which are qualitative nominal 0/3

- ☐ Time in terms of AM or PM
- ☐ Brightness as measured by a light meter
- ☐ Brightness as measured by people's judgement
- ☐ Angles as measured in degrees between 0 and 360
- ☑ Bronze, Silver, and Gold medals as awarded at the Olympics
- ☐ Height above sea level
- ☐ Number of patients in a hospital
- ☐ ISBN numbers for books. (Look up the format on the Web.)
- ☑ Ability to pass light in terms of the following values: opaque, translucent' transparent
- ☑ Military rank
- ☐ Distance from the centre of campus
- ☐ Density of a substance in grams per cubic centimetre
- ☑ Coat check number. When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave

What is the best split according to the information gain

0/2



CustomerID	Gender	CarType	Class
1	M	Family	C0
2	M	Sports	C0
3	M	Sports	C0
4	M	Sports	C0
5	F	Sports	C0
6	F	Luxury	C0
7	M	Family	C1
8	M	Family	C1
9	M	Family	C1
10	M	Luxury	C1
11	F	Luxury	C1
12	F	Luxury	C1
13	F	Luxury	C1
14	F	Luxury	C1
15	F	Luxury	C1
16	F	Luxury	C1

○ Customer ID

○ Gender

● Car Type

○ Class

Select all from the following attributes which are binary 0/3

- □ Time in terms of AM or PM
- ☑ Brightness as measured by a light meter
- ☑ Brightness as measured by people's judgement
- □ Angles as measured in degrees between 0 and 360
- □ Bronze, Silver, and Gold medals as awarded at the Olympics
- □ Height above sea level
- □ Number of patients in a hospital
- □ ISBN numbers for books. (Look up the format on the Web.)
- □ Ability to pass light in terms of the following values: opaque, translucent' transparent
- □ Military rank
- □ Distance from the centre of campus
- □ Density of a substance in grams per cubic centimetre
- □ Coat check number. When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave

The triangle inequality in Euclidean space is 0/1

○ $d(y, z) \geq d(x, y)+d(x, z)$

○ $d(x, z) \geq d(x, y)+d(y, z)$

● $d(x, y) \leq d(x, z)+d(y, z)$

○ $d(x, y) \leq d(x, z)-d(y, z)$

Which of the following data sets, data privacy is an important issue 2/2

○ Census data collected from 1900-1950

● IP addresses and visit times of Web users who visit your Website

○ Images from Earth-orbiting satellites

○ Names and addresses of people from the telephone book

○ Names and email addresses collected from the Web

Select all advantages of using colour to visually represent information 0/2

- □ Grayscale figures are not understandable
- ☑ Even if proper colour is not used, it is better than grey visualisation
- ☑ Colour makes it much easier to visually distinguish visual elements from one another
- ☑ Figures with colour are more interesting to look at

Which of the following sciences have complete control on data quality 1/1

- ○ Observation
- ● Carefully designed experiments

Consider a document-term matrix, where tfij is the frequency of the i th word (term) in the jth document and m is the number of documents. Consider the variable transformation that is defined by tfij ` = tfij ∗ log m dfi , where dfi is the number of documents in which the i th term appears and is known as the document frequency of the term. This transformation is known as the inverse document frequency transformation.Select all true statements from the following. 0/2

☑ The effect of this transformation is that the terms that occur in every document have 0 weight, while those that occur in one document have maximum weight, i.e., log m

☑ This transformation (normalization) reflects the observation that terms that occur in every document will have more power to distinguish one document from another, while those that are relatively rare do

Select all true statements with respect to in market basket analysis from the following 0/4

☑ Lowering the support threshold often results in more itemsets being declared as frequent

☑ The maximum size of frequent itemsets tends to increase with higher support thresholds

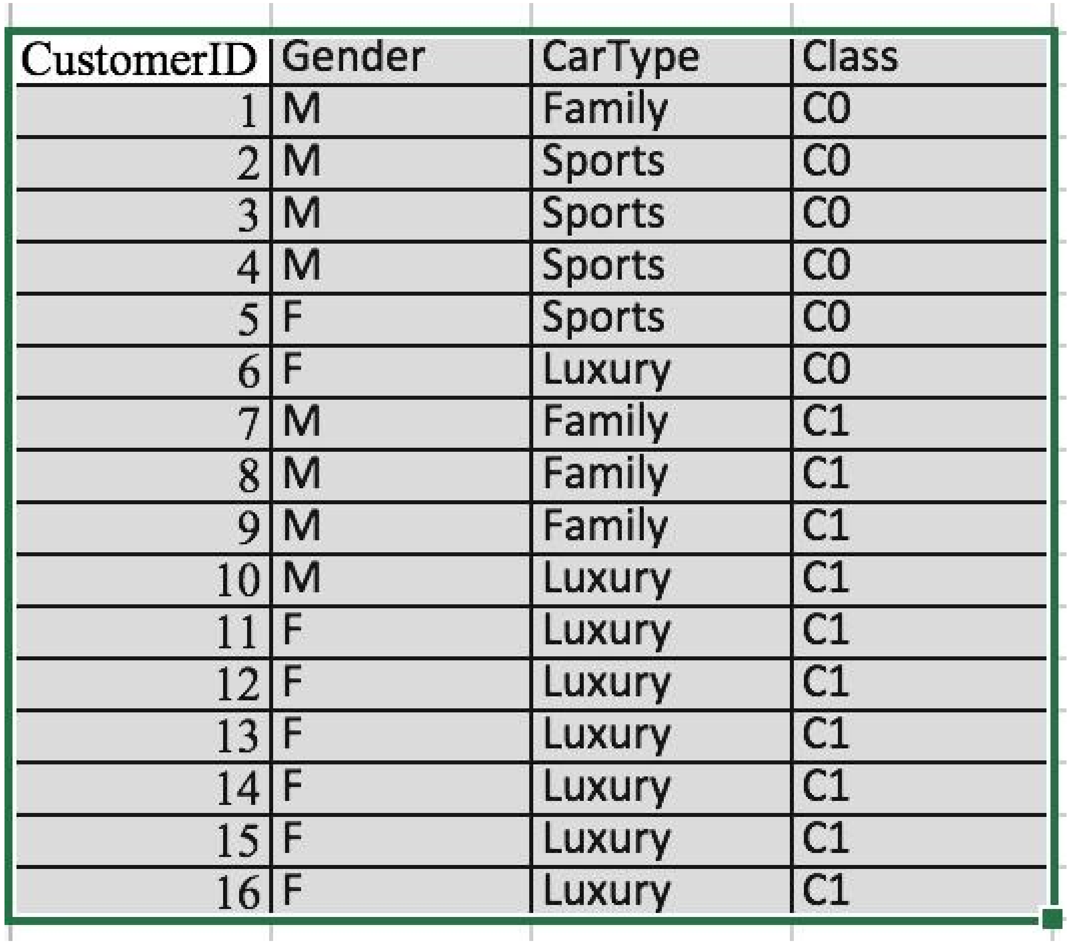
☑ As the number of items increases, more space will be needed to store the support counts of items

□ Apriori algorithm run time does not depend on number of transactions

☑ Apriori algorithm run time depends on Average Transaction Width

The Gini index for the CustomerID attribute is ___

2/2



CustomerID	Gender	CarType	Class
1	M	Family	C0
2	M	Sports	C0
3	M	Sports	C0
4	M	Sports	C0
5	F	Sports	C0
6	F	Luxury	C0
7	M	Family	C1
8	M	Family	C1
9	M	Family	C1
10	M	Luxury	C1
11	F	Luxury	C1
12	F	Luxury	C1
13	F	Luxury	C1
14	F	Luxury	C1
15	F	Luxury	C1
16	F	Luxury	C1

0

Select from the following valid definitions for the proximity among a group of objects. 0/2

☑ Based on pairwise proximity, i.e., minimum pairwise similarity or maximum pairwise dissimilarity

□ For points in Euclidean space compute a centroid and then compute the maximum distance of any point to the centroid

☑ For points in Euclidean space compute a centroid and then compute the minimum distance of any point to the centroid

□ For points in Euclidean space compute a centroid and then compute the sum or average of the distances of the points to the centroid

Which of the following are the part of descriptive tasks? 3/3

☑ Correlations

☑ Trends

☑ Clusters

☑ Trajectories

☑ Anomalies

□ Predict the value of a particular attribute based on the values of other attributes

Select all from the following which are the applications of association analysis 0/4

☑ Understanding the relationships between different elements of Earth's climate system

☑ Group sets of related customers

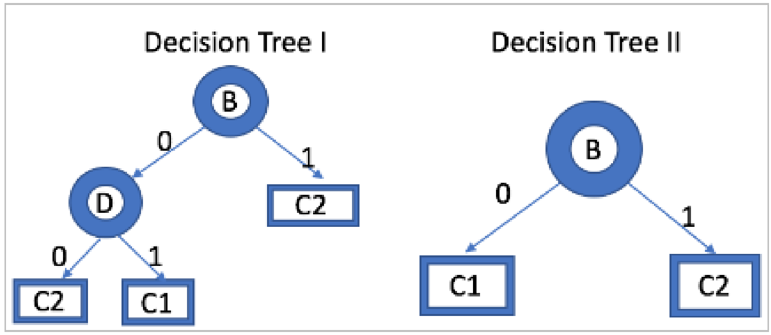
☑ Find areas of the ocean that have a significant impact on the Earth's climate, and compress data

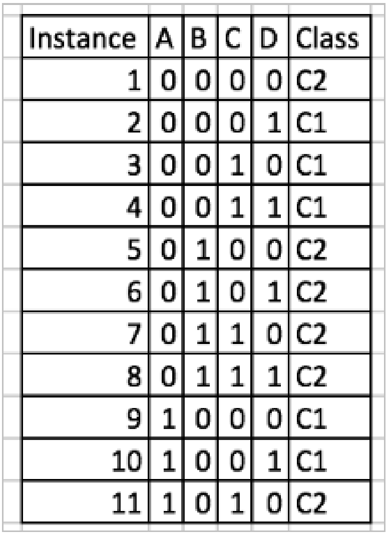
☑ Finding groups of genes that have related functionality

□ Identifying Web pages that are accessed together

□ Trends

Consider the decision tree shown below for given dataset. If we apply classification error rate as criteria which of the following statement is true 0/4





Instance	A	B	C	D	Class
1	0	0	0	0	C2
2	0	0	0	1	C1
3	0	0	1	0	C1
4	0	0	1	1	C1
5	0	1	0	0	C2
6	0	1	0	1	C2
7	0	1	1	0	C2
8	0	1	1	1	C2
9	1	0	0	0	C1
10	1	0	0	1	C1
11	1	0	1	0	C2

○ Only Decision Tree I is valid

○ Only Decision Tree II is valid

○ Both decisions trees are valid

● Both are not Valid

Which of the following measures you think would be more appropriate for comparing the genetic makeup of two organisms. Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise. Note: Two human beings share >99.9% of the same genes.Select all measures appropriate

2/2

□ Jaccard

☑ Hamming

Which of the following statements are true 0/5

☑ The coverage of a classification rule depends on the number of records that satisfy the rule antecedent.

□ The coverage of a classification rule depends on the total number of records.

□ The coverage of a classification rule depends on the number of records that satisfy both the antecedent and consequent.

☑ The coverage of a classification rule depends the number of records that satisfy consequent alone

☑ When the prior probabilities are different, the decision boundary shifts toward the class with higher prior probability

☑ Bayesian network is a directed acyclic graph (dag) encoding the dependence relationships among a set of variables

□ Cost matrix is both scale- invariant and translation-invariant

☑ Minimizing the total cost is equivalent to maximizing accuracy in all the cases

While the .632 bootstrap approach is useful for obtaining a reliable estimate of model accuracy, it has a known limitation. Consider a two-class problem, where there are 70% of positive and 30% negative examples in the data. Suppose the class labels for the examples are generated randomly with 70% positives and 30% negatives. The classifier used is an unpruned decision tree (i.e., a perfect memorizer). Match the accuracy of the classifier using each of the following methods

	0.3	0.616	0.4346	0.7	Score
The holdout method (two thirds for training)	○	○	●	○	0/1
Ten-fold cross-validation	●	○	○	○	1/1

Select all true statements from the following 0/2

☑ Hamming distance is a similarity measure

□ The Jaccard distance is similar to the SMC(Simple Matching Coefficient)

□ Hamming measure is similar to the cosine measure

☑ SMC = Hamming distance / number of bits

The following attributes are measured for members of a herd of Asian elephants: weight, height, tusk length, trunk length, and ear area. Based on these measurements, what sort of similarity measures would you use to compare or group these elephants? (select all appropriate measures) 2/2

□ Cosine Measure

□ Correlation Measure

□ Euclidean distance

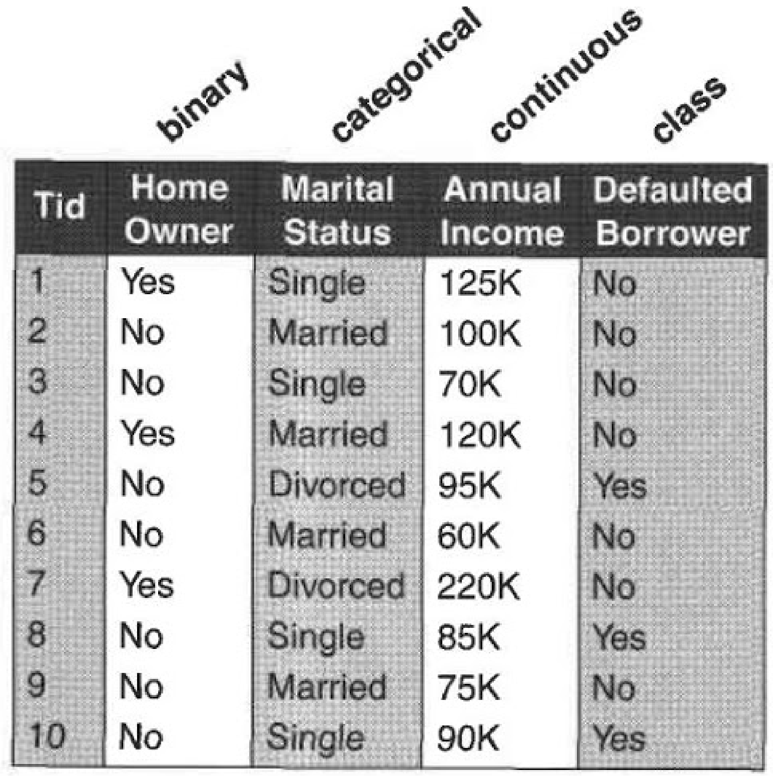
☑ Euclidean distance, applied after standardizing the attributes to have a mean of 0 and a standard deviation of 1

Which of the following is the right measure for customer satisfaction of a product? 0/1

● Ratio of number of complaints for the product and total number of sales for the product

○ Number of customer complaints for each product

The number of leaf nodes we get if we construct decision tree by using Hunt's algorithm is 0/2


binary
categorical
continuous
class

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training set for predicting borrowers who will default on loan payments.

○ 3

○ 4

○ 5

● 6

This form was created inside of Msitprogram.net.

Google Forms