1. Which of the following techniques does not require bifurcation of data points into dependent and independent variables?

A) Regression

B) Classification

C) Clustering

D) All of the above

1. An international foods and beverages company wants to look at what products it should launch in India. For that, it has first tried to segment the market. It is known that people living in the same area and having similar salaries will have similar eating habits. Then which of the following can be segmented in 1 group? (All options in 1 bracket are 1 segment)

A. High Earning Individual from Bengaluru

B. Low Earning Individual from Rural Uttar Pradesh

C. Mid Earning Individual from Mumbai

D. High Earning Individual from Hyderabad

A) (A,B,C,D)

B) (B,D) - (A) - (C)

C) (A,D) - (B) - (C)

D) (A,C,D) - (B)

1. You are an analyst at a global laptop manufacturer and are given the task of deciding whether the company should enter the Indian Market. You try to estimate the market size by first breaking the market by different types of people who use a laptop such as students, working professionals and their paying capacity to get an estimate of the total market size and the characteristics of each segment. In essence, you are doing:

A) Behavioural Segmentation

B) Attitudinal Segmentation

C) Demographic Segmentation

D) None of the above

4. In PCA, project the data into the direction in which its:

A) Variance is maximum

B) Variance is minimum

C) Mean is minimum

D) None of the above

5. what will be the output if we submit the code:

vect <- c(1, 3, 5, 6)

vect2 <- vect[-3]

vect2

What will be the output of the third line (vect2) in the above script?

A) 1, 5, 6

B) 1, 3, 6

C) 5

6. A non-contributing predictor variable is added to an existing multiple linear regression model. What will be the result?

A) An increase in R-Square

B) A decrease in R-Square

C) A decrease in Mean Square Error

D) No change in R-Square

7. Linear Regression models are fitted using the:

A) Simple square method

B) Maximum square method

C) Least square method

D) Extreme square method

8. A manager wants to predict if the customer would buy a product or not.

What model must he use?

A) Linear Regression

B) Logistic Regression

C) Clustering

D) None of the above

9. A Singapore-based start-up launched an app called JustShakeIt that enables a user to send an emergency alert to emergency contacts and/or caregivers simply by shaking the phone with one hand.

The program uses a machine learning algorithm to distinguish between actual emergency shakes and everyday jostling, using data with labels to distinguish between everyday jostling and emergency shaking.

What kind of problem is this?

A) Supervised learning – Regression

B) Unsupervised learning – Clustering

C) Supervised learning – Classification

D) Can’t say with this information

10. Including redundant input variables in a regression model can:

A) Stabilize parameter estimates and increase the risk of over-fitting.

B) Destabilize parameter estimates and increase the risk of over-fitting.

C) Stabilize parameter estimates and decrease the risk of over-fitting.

D) Destabilize parameter estimates and decrease the risk of over-fitting.

11. We Use Dimensionality Reduction Techniques to deal with which Problem?

A) Heteroscedasticity

B) Autocorrelation

C) Multicollinearity

D) Non-Normal distribution

12. The parameters the Linear regression model learning are 1.\_\_\_\_\_ 2.\_\_\_\_\_

13. Logistic Regression Curve also Known as ---------

14. *Refer to the confusion matrix:*

*Predicted Outcome*

*0 1*

*Actual outcome 0 58 44*

*1 23 25*

*Calculate the accuracy and error rate (0-Negative Outcome, 1- Positive Outcome).*

*A) Accuracy=83/102 Error rate =67/102*

*B) Accuracy=58/102 Error rate =23/48*

*C) Accuracy=83/150 Error rate =67/150*

*D) Accuracy=25/150 Error rate =44/150*

15. *The total modeling data has been split into training, Validation and Test data.*

*What is the best data to use for model assessment?*

*A) Validation data*

*B) Total data*

*C) Training data*

*D) Test data*

16. If a model designer tries to strike the balance between keeping the model simple yet not making it too naive to be of any use, this phenomenon is called as \_\_\_\_\_.

17. If a model becomes way too complex than what is warranted for the task at hand and as a result suffers from bad generalization properties, this phenomenon is known as ------.

18. While Building ROC curve, what was the metric represent on Y Axis? -----

19. You have two classes in your model - spam and ham. Let's say you are building the model for a client who is a logistics company and they'll use your spam classifier for maintaining the sanity of employees' inboxes.

Now, they've demanded that your model should never misclassify a genuine email as spam (imagine your appraisal email misclassified as spam), though it is okay if some spams seep into the inbox.

Assume that 'spam' is called the 'positive class', and ham is 'negative'. Which of the following metrics should your model maximise?

A) Accuracy

B) Sensitivity/Recall (True Positive Rate)

C) Specificity (True Negative Rate)

D) Precision

20. Out of so many attributes, how does a decision tree select one for splitting? Select the best option.

A) It picks an attribute at random.

B) It calculates the improvement in homogeneity associated with each attribute and picks the one that results in the maximum increase in homogeneity.

C) It calculates the improvement in homogeneity associated with each attribute and picks the one that results in the minimum increase in homogeneity.

D) It calculates the improvement in homogeneity associated with each attribute and picks the one that results in no change in homogeneity.

21. Over-fitting will create

A) Negative impact on the model’s predictive power

B) No impact on the model’s predictive power

C) Positive impact on the model’s predictive power

22. Example of low bias algorithm is

A) Liner Regression

B) Logistic Regression

C) Decision Tree

23. When is the Gini index maximum?

A) When the homogeneity is minimum.

B) When the homogeneity is maximum.

C) The Gini index does not depend on the homogeneity.

24. When is the information gain maximum? (Select the most appropriate option.)

A) When the decrease in entropy, from the parent set to the partitions obtained after splitting, is maximum.

B) When the decrease in entropy, from the parent set to the partitions obtained after splitting, is minimum.

25. An artificially generated dataset was used to generate data, Three regression models have been created to fit the data - linear, a degree-3 polynomial and a higher degree polynomial which passes through all the training points.

The correct order of bias in the three models is:

A) Straight line > Degree-3 > Polynomial

B) Straight line > Polynomial > Degree-3

C) Polynomial > Degree-3 > Straight line

D Polynomial > Straight line > Degree-3

26. Which hyper-parameter controls the minimum no. of samples required to split an internal node?

A) min\_samples\_split

B) max\_depth

C) min\_samples\_leaf

D) max\_features

27. Assume that you have set the min\_samples\_leaf as 3 and the min\_samples\_split as 6. Consider a node with 10 data points. On splitting on an attribute, one leaf gets 2 points, and the other one gets 8 data points. This split will not be executed. Why?

A) The number of data points in the node > min\_samples\_split.

B) The number of data points in one of the leaves < min\_samples\_leaf.

C) The number of data points in one of the leaves > min\_samples\_leaf.

D) The number of data points in the node > min\_samples\_leaf.

28. Which of the following is a not a hyper-parameter for a random forest classifier?

A) Depth of a tree

B) learning rate learning rate - correct

C) The minimum number of samples required to be at a leaf node / bottom of a tree

D) The number of trees

29. Each tree is built using random dataset and random variable set, these trees are ensembled and is known as \_\_\_\_\_\_\_\_

A) Decision Tree

B) Classification Tree

C) Regression Tree

D) Random Forest

30. Say you have a 100,000 Google images as training observations and you are trying to build a neural network to classify the images in 3 classes - nature, cities and others.  You use another 50,000 observations to test it and the accuracy on the test set comes out to be 10%. Which of the following can you use to check whether the model has overfitted?

A) Sensitivity and specificity on the test set

B) Number of model parameters

C) Accuracy on the training set

D) The performance of another model like SVM on the same training set