## **MCQ ON PREDICTIVE MODELLING-2**

Multiple-choice questions (MCQs) on various topics related to Predictive Modeling. Each question is followed by the correct answer.

Identifying Relationship in Data:

- 1. Which of the following techniques can be used to identify the correlation between two continuous variables in a dataset?
  - a) Principal Component Analysis (PCA)
  - b) Pearson correlation coefficient
  - c) K-means clustering
  - d) Decision tree

Correct Answer: b) Pearson correlation coefficient

- 2. In data analysis, what does a positive correlation coefficient indicate between two variables?
  - a) There is a strong negative relationship between the variables.
  - b) There is no relationship between the variables.
  - c) There is a strong positive relationship between the variables.
  - d) The variables are independent of each other.

Correct Answer: c) There is a strong positive relationship between the variables.

- 3. Which type of plot is commonly used to visualize the relationship between two continuous variables in predictive modeling?
  - a) Box plot
  - b) Histogram
  - c) Scatter plot

d) Bar plot

Correct Answer: c) Scatter plot

- 4. When two variables have a correlation coefficient close to -1, it indicates:
  - a) A strong positive linear relationship between the variables
  - b) A weak positive linear relationship between the variables
  - c) A weak negative linear relationship between the variables
  - d) A strong negative linear relationship between the variables

Correct Answer: d) A strong negative linear relationship between the variables

- 5. Which of the following statements about causality and correlation is true?
  - a) Correlation implies causation.
  - b) Causation implies correlation.
  - c) Correlation and causation are the same concepts.
  - d) Correlation does not imply causation.

Correct Answer: d) Correlation does not imply causation.

- 6. In data analysis, what does it mean when the correlation coefficient is close to 0?
  - a) There is a strong positive correlation between the variables.
  - b) There is no correlation between the variables.
  - c) There is a strong negative correlation between the variables.
  - d) The correlation coefficient is undefined.

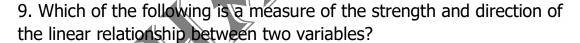
Correct Answer: b) There is no correlation between the variables.

- 7. Which of the following techniques can be used to handle missing data when identifying relationships in data?
  - a) Removing the rows with missing data
  - b) Replacing the missing data with the mean of the variable
  - c) Imputing the missing data using regression imputation
  - d) All of the above

Correct Answer: d) All of the above

- 8. When analyzing the relationship between a categorical variable and a continuous variable, which type of plot is appropriate to use?
  - a) Scatter plot
  - b) Box plot
  - c) Histogram
  - d) Bar plot

Correct Answer: b) Box plot



- a) Standard deviation
- b) P-value
- c) Correlation coefficient
- d) Mean Absolute Error (MAE)

Correct Answer: c) Correlation coefficient

- 10. In data analysis, what does it mean when the correlation coefficient is positive and close to 1?
  - a) There is a weak positive correlation between the variables.
  - b) There is a strong positive correlation between the variables.

- c) The variables are not related to each other.
- d) The correlation coefficient is undefined.

Correct Answer: b) There is a strong positive correlation between the variables.

Predictive Modelling using Clustering:

- 11. Which of the following is an unsupervised learning technique used for clustering in predictive modeling?
  - a) Decision tree
  - b) Random Forest
  - c) k-Means
  - d) Logistic Regression

Correct Answer: c) k-Means

- 12. Clustering is used in predictive modeling to:
  - a) Divide the dataset into training and testing sets
  - b) Identify groups or patterns in the data based on similarities
  - c) Handle missing data
  - d) Visualize the data distribution

Correct Answer: b) Identify groups or patterns in the data based on similarities

- 13. In k-Means clustering, the number of clusters is specified by the user beforehand. This number is known as:
  - a) Cluster size
  - b) Cluster index
  - c) Centroid value

d) Number of clusters

Correct Answer: d) Number of clusters

- 14. Which of the following clustering algorithms is based on density and connectivity?
  - a) k-Means
  - b) Agglomerative Hierarchical Clustering
  - c) Decision tree
  - d) Random Forest

Correct Answer: b) Agglomerative Hierarchical Clustering

- 15. The silhouette score is a metric used to evaluate the quality of clusters in clustering algorithms. It measures:
  - a) The number of clusters in the data
  - b) The density of the clusters
  - c) The separation between clusters and the cohesion within clusters
  - d) The correlation between variables in the clusters

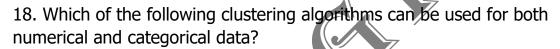
Correct Answer: c) The separation between clusters and the cohesion within clusters

- 16. Which of the following statements about clustering algorithms is true?
- a) Clustering algorithms can only be used for supervised learning tasks.
  - b) Clustering algorithms require labeled data for training.
- c) Clustering algorithms are used to classify data into predefined classes.
  - d) Clustering algorithms are used to group data based on similarities.

Correct Answer: d) Clustering algorithms are used to group data based on similarities.

- 17. In hierarchical clustering, the process of combining individual data points into larger clusters is known as:
  - a) Aggregation
  - b) Linkage
  - c) Centroid computation
  - d) Initialization

Correct Answer: b) Linkage



- a) k-Means
- b) DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
  - c) Hierarchical clustering
  - d) Principal Component Analysis (PCA)

Correct Answer: b) DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

- 19. Which method is used to determine the optimal number of clusters in k-Means clustering?
  - a) The Elbow method
  - b) The Silhouette method
  - c) The Root Mean Squared Error (RMSE)
  - d) The R-squared value

Correct Answer: a) The Elbow method

- 20. Which of the following statements about clustering is true?
- a) The number of clusters is always determined by the algorithm automatically.
- b) Clustering can be used for both unsupervised and supervised learning tasks.
  - c) Clustering is used to predict a continuous target variable.
  - d) Clustering does not require any input parameters.

Correct Answer: b) Clustering can be used for both unsupervised and supervised learning tasks.

Predicting the Future using Classification:

- 21. In predictive modeling, classification is used for:
  - a) Grouping similar data points together
  - b) Identifying the relationship between two variables
  - c) Predicting continuous target variables
  - d) Assigning data points to predefined categories or classes Correct Answer d) Assigning data points to predefined

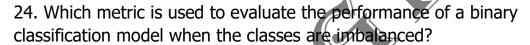
categories or classes

- 22. Which of the following is a binary classification algorithm?
  - a) Decision tree
  - b) k-Means
  - c) Random Forest
  - d) Logistic Regression

## Correct Answer: d) Logistic Regression

- 23. The receiver operating characteristic (ROC) curve is used to evaluate the performance of a classification model by plotting the trade-off between:
  - a) True Positive Rate (TPR) and False Negative Rate (FNR)
  - b) Sensitivity and Specificity
  - c) Precision and Recall
  - d) Accuracy and Error Rate

Correct Answer: b) Sensitivity and Specificity



- a) Mean Absolute Error (MAE)
- b) Mean Squared Error (MSE)
- c) F1 score
- d) R-squared value

Correct Answer: c) F1 score

- 25. The Confusion Matrix is a table used to evaluate the performance of a classification model by showing the counts of:
  - a) True Positive, True Negative, False Positive, and False Negative
  - b) Positive, Negative, True, and False
  - c) Correct and Incorrect predictions
  - d) Sensitivity and Specificity values

Correct Answer: a) True Positive, True Negative, False Positive, and False Negative

- 26. Which of the following is a method used to handle imbalanced classes in a classification model?
  - a) Feature scaling
  - b) Oversampling the minority class
  - c) Principal Component Analysis (PCA)
  - d) K-means clustering

Correct Answer: b) Oversampling the minority class

- 27. In classification, the term "precision" refers to the number of:
- a) Correct positive predictions divided by the total number of positive predictions
- b) Correct positive predictions divided by the total number of actual positive cases
  - c) Correct predictions divided by the total number of predictions
- d) True Positive predictions divided by the total number of positive predictions

Correct Answer: a) Correct positive predictions divided by the total number of positive predictions

- 28. Which of the following statements about the K-nearest neighbors (KNN) algorithm is true?
  - a) KNN is a supervised learning algorithm.
  - b) KNN can only be used for binary classification tasks.
  - c) KNN assigns each data point to the nearest cluster centroid.
- d) KNN is a non-parametric algorithm used for classification and regression.

Correct Answer: d) KNN is a non-parametric algorithm used for classification and regression.

29. Which of the following classification algorithms is based on the idea of "voting" from multiple decision trees?
a) Decision tree
b) k-Means
c) Random Forest
d) Logistic Regression
Correct Answer: c) Random Forest
30. In binary classification, which threshold value is commonly used to convert predicted probabilities into class labels?
a) 0.5
b) 0
c) 1
d) The threshold value is determined by the algorithm.
Correct Answer: a) 0.5
31. The area under the ROC curve (AUC-ROC) is a metric used to measure the performance of a classification model. The AUC value ranges from:
a) 0 to 1 b) 1 to 1
c) -∞ to ∞
d) 0,to ∞
Correct Answer: a) 0 to 1
32. Which of the following statements about the Naive Bayes classifier is true?
a) Naive Bayes is based on the idea of k-Nearest Neighbors (KNN).

- b) Naive Bayes assumes that features are dependent on each other.
- c) Naive Bayes is a non-parametric classifier.
- d) Naive Bayes is based on Bayes' theorem and assumes that features are conditionally independent.

Correct Answer: d) Naive Bayes is based on Bayes' theorem and assumes that features are conditionally independent.

- 33. Which of the following is NOT a classification algorithm?
  - a) Decision tree
  - b) Linear regression
  - c) Logistic Regression
  - d) Support Vector Machine (SVM)

Correct Answer: b) Linear regression

- 34. Which technique can be used to handle imbalanced classes by generating synthetic samples for the minority class?
  - a) Undersampling
  - b) Oversampling
  - c) Feature scaling
  - d) Feature selection

Correct Answer: b) Oversampling

- 35. In classification, the term "recall" refers to the number of:
- a) Correct positive predictions divided by the total number of positive predictions
- b) Correct positive predictions divided by the total number of actual positive cases
  - c) Correct predictions divided by the total number of predictions

d) True Positive predictions divided by the total number of actual positive cases

Correct Answer: b) Correct positive predictions divided by the total number of actual positive cases

- 36. Which of the following is a method used to handle imbalanced classes in a classification model?
  - a) Feature scaling
  - b) Oversampling the minority class
  - c) Principal Component Analysis (PCA)
  - d) K-means clustering

Correct Answer: b) Oversampling the minority class

- 37. In classification, the term "precision" refers to the number of:
- a) Correct positive predictions divided by the total number of positive predictions
- b) Correct positive predictions divided by the total number of actual positive cases
  - c) Correct predictions divided by the total number of predictions
- d) True Positive predictions divided by the total number of positive predictions

correct Answer: a) Correct positive predictions divided by the total number of positive predictions

- 38. Which of the following statements about the K-nearest neighbors (KNN) algorithm is true?
  - a) KNN is a supervised learning algorithm.
  - b) KNN can only be used for binary classification tasks.
  - c) KNN assigns each data point to the nearest cluster centroid.

d) KNN is a non-parametric algorithm used for classification and regression.

Correct Answer: d) KNN is a non-parametric algorithm used for classification and regression.

- 39. Which of the following classification algorithms is based on the idea of "voting" from multiple decision trees?
  - a) Decision tree
  - b) k-Means
  - c) Random Forest
  - d) Logistic Regression

Correct Answer: c) Random Forest

- 40. In binary classification, which threshold value is commonly used to convert predicted probabilities into class labels?
  - a) 0.5
  - b) 0
  - c) 1
  - d) The threshold value is determined by the algorithm.

Correct Answer: a) 0.5

- 41. The area under the ROC curve (AUC-ROC) is a metric used to measure the performance of a classification model. The AUC value ranges from:
  - a) 0 to 1
  - b) -1 to 1
  - c)  $-\infty$  to  $\infty$
  - d) 0 to ∞

Correct Answer: a) 0 to 1

42. Which of the following statements about the Naive

Bayes classifier is true?

- a) Naive Bayes is based on the idea of k-Nearest Neighbors (KNN).
- b) Naive Bayes assumes that features are dependent on each other.
- c) Naive Bayes is a non-parametric classifier.
- d) Naive Bayes is based on Bayes' theorem and assumes that features are conditionally independent.

Correct Answer: d) Naive Bayes is based on Bayes' theorem and assumes that features are conditionally independent.

- 43. Which of the following is NOT a classification algorithm?
  - a) Decision tree
  - b) Linear regression
  - c) Logistic Regression
  - d) Support Vector Machine (SVM)

Correct Answer: b) Linear regression

- 44. Which technique can be used to handle imbalanced classes by generating synthetic samples for the minority class?
  - a) Undersampling
  - b) Oversampling
  - c) Feature scaling
  - d) Feature selection

Correct Answer: b) Oversampling

- 45. In classification, the term "recall" refers to the number of:
- a) Correct positive predictions divided by the total number of positive predictions
- b) Correct positive predictions divided by the total number of actual positive cases
  - c) Correct predictions divided by the total number of predictions
- d) True Positive predictions divided by the total number of actual positive cases

Correct Answer: b) Correct positive predictions divided by the total number of actual positive cases

- 46. Which of the following statements about classification algorithms is true?
- a) Classification algorithms are only used for unsupervised learning tasks.
- b) Classification algorithms can be used for both binary and multiclass classification problems.
  - c) Classification algorithms are used to identify patterns in the data.
  - d) Classification algorithms are based on clustering techniques.

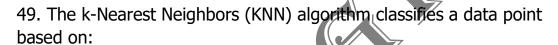
Correct Answer: b) Classification algorithms can be used for both binary and multiclass classification problems.

- 47. The decision boundary in a classification model represents:
- a) The line or surface that separates the data points into different classes.
  - b) The accuracy of the model in predicting the target variable.
  - c) The threshold value used to convert probabilities into class labels.
  - d) The centroid of each cluster in the data.

Correct Answer: a) The line or surface that separates the data points into different classes.

- 48. Which of the following is NOT a common distance metric used in clustering and classification?
  - a) Euclidean distance
  - b) Manhattan distance
  - c) Mahalanobis distance
  - d) Pearson correlation

Correct Answer: d) Pearson correlation



- a) The centroid of the cluster to which it belongs
- b) The mean of the target variable in its neighborhood
- c) The majority class of its k nearest neighbors
- d) The distance to the farthest data point in the dataset

  Correct Answer: c) The majority class of its k nearest neighbors
- 50. In classification, what does the term "precision-recall trade-off" refer to?
  - a) The trade-off between precision and accuracy in the model
  - b) The trade-off between sensitivity and specificity in the model
- c) The trade-off between the number of features and model complexity
- d) The trade-off between the number of clusters and the model's performance

Correct Answer: b) The trade-off between sensitivity and specificity in the model

- 51. In classification, the F1 score is the harmonic mean of:
  - a) Precision and recall
  - b) Sensitivity and specificity
  - c) Accuracy and error rate
  - d) True Positive Rate (TPR) and False Positive Rate (FPR)

Correct Answer: a) Precision and recall

- 52. Which of the following techniques can be used to improve the performance of a classification model with high-dimensional data?
  - a) Feature scaling
  - b) Feature selection
  - c) Oversampling the minority class
  - d) Principal Component Analysis (PCA)

Correct Answer: d) Principal Component Analysis (PCA)

- 53. Which of the following statements about Support Vector Machine (SVM) is true?
  - a) SVM is a clustering algorithm.
  - b) SVM is used only for binary classification tasks.
  - c) SVM can only handle linearly separable data.
- d) SVM aims to find the hyperplane that maximizes the margin between classes.

Correct Answer: d) SVM aims to find the hyperplane that maximizes the margin between classes.

54. Which of the following is a method used to handle class imbalance by reducing the number of instances in the majority class?

- a) Feature scaling
- b) Undersampling
- c) Oversampling
- d) Feature selection

Correct Answer: b) Undersampling

- 55. The k-Nearest Neighbors (KNN) algorithm is an example of
  - a) Parametric classification algorithm
  - b) Non-parametric classification algorithm
  - c) Clustering algorithm
  - d) Linear classification algorithm

Correct Answer: b) Non-parametric classification algorithm

- 56. In classification, which metric can be used to measure the balance between precision and recall?
  - a) F1 score
  - b) Accuracy
  - c) R-squared value
  - d) Mean