ID: **NAME:**

Birla Institute of Technology & Science (BITS), Pilani
2nd SEMESTER 2021-22,
Predictive Analytics MPBA G513
Comprehensive Examination (Closed Book) – Part A

Max. Time: 30 Minutes Date: 21-05-2022 Max. Marks: 20

Instructions:

I. Please write your answer in the response space as provided in the second page of the question paper. The answers should be neatly written and free of any over writing/cutting. Otherwise it will not be evaluated.
II. Please submit the Part-A paper upon completion to the invigilator to get the question paper of Part-B. Suggested maximum time to complete Part-A is 20 minutes.
III. No negative marking is there for incorrect answers

---

1. Identify the type of learning in which labeled training data is used
   a. Semi Supervised Learning
   b. Supervised Learning
   c. Reinforcement Learning
   d. Unsupervised Learning
2. In Principal component analysis
   a. Number of components is always equal to the number of dimensions
   b. Number of components will always be less than the number of dimensions
   c. Number of components can be more than the number of dimensions
   d. All of the above are correct
3. Dimensionality reduction reduces
   a. Entropy
   b. Overfitting
   c. Performance
   d. Collinearity
4. Which of the following machine learning algorithm is based upon the idea of bagging?
   a. Decision Tree
   b. Random Forest
   c. Support Vector Machine
   d. GBM
5. Choose a disadvantage of decision trees among the following
   a. They are robust to outliers
   b. They are prone to over fitting
   c. They are non-parametric models
   d. All of the above
6. Among the following options identify the one which is false regarding regression
   a. It is used for prediction
   b. It is used for interpretation
   c. It related inputs and outputs
   d. It can be used for any kind of target
7. Choose the most widely used metrics and tools to assess the classification models
   a. Area under the ROC curve
   b. Confusion Matrix
   c. Cohens Kappa
   d. All the above
8. What does K stand for in K means algorithm?
   a. Number of clusters
   b. Number of observations
   c. Number of features
   d. Number of dimensions
9. Which of the following is not a supervised learning
   a. PCA
   b. Naïve Bayes
   c. Linear Regression
   d. Decision Tree
10. Which of the following CANNOT be achieved by using machine learning?
    a. classify respondents into groups based on their response pattern
    b. accurately predict the outcome using supervised learning algorithms
    c. proving causal relationships between variables
    d. forecast the outcome variable into the future
11. Which of the following is TRUE about unsupervised machine learning?
    a. a semi-autonomous machine learning where researchers control some parts of the modelling process
    b. learning algorithms with no control over quality of their predictions
    c. a fully autonomous machine learning with no human interference
    d. unsupervised learning comprises algorithms with no pre-existing outcomes
12. What is the Elbow method?
    a. a method used to determine the optimal number of clusters in unsupervised learning, for example K-mean clustering
    b. an approach to estimating 'black-box' predictions in supervised learning
    c. a way of assessing the fit of a machine learning algorithm

d. a method of forecasting in machine learning

13. Which of the following is TRUE about over-fitting in machine learning?
    a. a model with too many outcome categories
    b. a model with too many predictors
    c. a situation where a model predicts training data very well but fares much worse in testing data
    d. predictions that very highly accurate

14. Which of the following is NOT true about feature importance scores?
    a. show the size and direction of the effect of a predictor on an outcome
    b. pertains mostly to supervised machine learning
    c. there are different ways of measuring feature importance, for example by using mean decrease accuracy scores
    d. quantify the unique contribution of a given feature (predictor)

15. What is the purpose of cross-validation?
    a. to validate the performance of a machine learning algorithm by resampling the original dataset
    b. to re-formulate a research hypothesis
    c. to find the optimal number of clusters in unsupervised learning
    d. to apply different thresholds of statistical significance

16. Random forest is
    a. a method of linear regression analysis based on multiple resampling
    b. a learning algorithm where a researcher fits a limited number of decision trees in a specified order
    c. an ensemble machine learning method comprised of multitude of decision trees
    d. a method of visualizing the effect of a given predictor on an outcome

17. Which of the following statement is False in the case of the KNN Algorithm?
    a. For a very large value of K, points from other classes may be included in the neighborhood
    b. For the very small value of K, the algorithm is very sensitive to noise.
    c. KNN is used only for classification problem statements
    d. KNN is a lazy learner

18. Which of the following statement is TRUE?
    a. Outliers should be identified and removed always from a dataset.
    b. Outliers can never be present in the testing dataset.
    c. Outliers is a data point that is significantly close to other data points.
    d. The nature of our business problem determines how outliers are used

19. Which one of the following statements is TRUE for a Decision Tree?
    a. Decision tree is only suitable for the classification problem statement
    b. In a decision tree, the entropy of a node decreases as we go down a decision tree
    c. In a decision tree, entropy determines purity
    d. Decision tree can only be used for only numeric valued and continuous attributes

20. How do you choose the right node while constructing a decision tree?
    a. An attribute having high entropy
    b. An attribute having high entropy and information gain
    c. An attribute having the lowest information gain
    d. An attribute having the highest information gain

Response space

Q.1	Q.2	Q.3	Q.4	Q.5	Q.6	Q.7	Q.8	Q.9	Q.10
Q.11	Q.12	Q.13	Q.14	Q.15	Q.16	Q.17	Q.18	Q.19	Q.20

Birla Institute of Technology & Science (BITS), Pilani
2nd SEMESTER 2021-22,
Predictive Analytics MPBA G513
Comprehensive Examination (Closed Book) – Part B

Max. Time: 150 Minutes | Date: 21-05-2022 | Max. Marks: 40

Question 1: Answer any 12 of the following questions very briefly (Unnecessary points and wrong points will result in deduction of marks). Each question carries 2 marks (12x2 = 24 Marks)

(i) Name any three important metrics that you use to test the best regression model on the testing data set along with a broad mechanism to calculate the same
(ii) Name any three nonlinear transformations that you plan to do on the data in case the data is not normally distributed and one key aspect of each of the transformations.
(iii) Name the three most important parameters of the Gradient boosting classification algorithm and the key purpose of each of the parameters.
(iv) Mention three key aspects of KNN algorithm which differentiate it from the rest of the algorithms. (Writing lengthier paragraphs on the algorithm will not fetch any marks, rather the three bullet points need to be answered)
(v) Name and briefly explain any three important approaches to select features explicitly in a regression problem. Please do not name the mechanisms that are common to classification and regression problems.
(vi) Name any three metrics that can be used with computing the cross validation score for a classification problem.
(vii) Name the three most important parameters of the Random Forest classification algorithm and the key purpose of each of the parameters.
(viii) Mention three key aspects of Density based clustering algorithm which differentiate it from the rest of the algorithms. (Writing lengthier paragraphs on the algorithm will not fetch any marks, rather the three bullet points need to be answered)
(ix) Write a sentence on the role of (a) Heat map on correlation matrix (b) Kernel density estimation plots and (c) Count plot in a prediction exercise.
(x) Name any three metrics that can be used with computing the cross validation score for a regression problem.
(xi) Name the three most important parameters of the Support vector machines classification algorithm and the key purpose of each of the parameters.
(xii) Name any three types of scaling you would like to do on a data before using it in a prediction exercise.
(xiii) Name at least three different ways in which you can identify outliers in a data. For each method, write one or two sentences on the methodology used for computation of outliers.
(xiv) What is the relationship between the explained variance ratio and the dimensionality of the dataset in a principal component analysis exercise?
(xv) Write three important points about the ROC curve and the role of area under the ROC curve in the context of classification problems

Question 2 (16 Marks)

You are asked to design a model that can be used to predict fee that can be charged by a business school for a newly created MBA program. You need to identify 10 important variables which can be used to build the model. There should be at least 3 categorical variables and at least 3 quantitative variables. Create a model data set of 10 rows with appropriate values filled in.