Review Questions?

What are the differences between regression and classification?

Explain the steps in making decision tree. What is the problem of decision tree?

How is decision tree different than random forest? How bagging is different from random forest? Can we say theoretically random forest always performs better than decision tree at test data? (Answer is Yes)? Why or why not?

What is Bias Variance tradeoff? What is Bias and Variance in Bias Variance tradeoff?

What is underfitting or overfitting?

How can you avoid overfitting your model?

How can you treat missing value if variable is categorical or numerical variable?

How Boosting is different from Bagging? What is the algorithm for Gradient Boosting Algorthitms like XGBoost?

Can we say theoretically boosting algorithms like XGBoost always performs better than decision tree at test data (Answer is No)? Why or why not?

What is regularization? Why is it needed? **Describe different regularization methods, such as L1 and L2 regularization** What parameters is needed for regularization?

How is

In case of CoVID test, which test is more important? Precison or Recall? Why?

What is ROC Curve? What is the meaning of AUC?

Write 3 python library.

 Explain cross-validation.

Evaluation metrics for classification.

Evaluation metrics for regressions

**What is the difference between a box plot and a histogram**

**What are type 1 and type 2 error?**

**How is Pearson correlation and spearman correlation different?**

**What is the difference between causality and correlation?**

**What is the difference between supervised learning and unsupervised learning?**

**Differentiate between univariate, bivariate and multivariate analysis.**

**Can you cite some examples where a false positive is important than a false negative?**

**Can you cite some examples where a false negative important than a false positive?**

### **What is Ensemble Learning?**

### **Relation between Central limit theorem and Bagging**

### Graphical user interface, text Description automatically generated

### Text, letter Description automatically generated

### Text, letter Description automatically generated

### Graphical user interface, text, Word Description automatically generated

1. a = 1
2. while True:
3. if a % 7 == 0:
4. break
5. print(a)
6. a += 1

What is correct output of this program?

1. How is the root mean squared is different than mean absolute error value?

**2. \_\_\_\_\_\_ is the basic data structure of pandas that can be thought of SQL table or a spreadsheet data representation.**

1. Series
2. List
3. Dataframe
4. ndarray

**3) Which of the following is/are true about bagging trees?**

1. In bagging trees, individual trees are independent of each other
2. Bagging is the method for improving the performance by aggregating the results of weak learners

A) 1  
B) 2  
C) 1 and 2  
D) None of these

**4) Which of the following is/are true about boosting trees?**

1. In boosting trees, individual weak learners are independent of each other
2. It is the method for improving the performance by aggregating the results of weak learners

A) 1  
B) 2  
C) 1 and 2  
D) None of these

**5) Which of the following is/are true about Random Forest and Gradient Boosting ensemble methods?**

1. Both methods can be used for classification task
2. Random Forest is use for classification whereas Gradient Boosting is use for regression task
3. Random Forest is use for regression whereas Gradient Boosting is use for Classification task
4. Both methods can be used for regression task

A) 1  
B) 2  
C) 3  
D) 4  
E) 1 and 4

**6) In Random forest you can generate hundreds of trees (say T1, T2 …..Tn) and then aggregate the results of these tree. Which of the following is true about individual(Tk) tree in Random Forest?**

1. Individual tree is built on a subset of the features
2. Individual tree is built on all the features
3. Individual tree is built on a subset of observations
4. Individual tree is built on full set of observations

A) 1 and 3  
B) 1 and 4  
C) 2 and 3  
D) 2 and 4

**7) Which of the following is true about “max\_depth” hyperparameter in Gradient Boosting?**

1. Lower is better parameter in case of same cross validation accuracy
2. Higher is better parameter in case of same cross validation accuracy
3. Increase the value of max\_depth may overfit the data
4. Increase the value of max\_depth may underfit the data

A) 1 and 3  
B) 1 and 4  
C) 2 and 3  
D) 2 and 4

**8) Which of the following algorithm would you take into the consideration in your final model building on the basis of performance?**

Suppose you have given the following graph which shows the ROC curve for two different classification algorithms such as Random Forest(Red) and Logistic Regression(Blue)

[Chart, line chart

Description automatically generated](https://cdn.analyticsvidhya.com/wp-content/uploads/2017/07/12181501/7_image_dt.jpg)

A) Random Forest  
B) Logistic Regression  
C) Both of the above  
D) None of these

**9) Which of the following is true about training and testing error in such case?**

Suppose you want to apply AdaBoost algorithm on Data D which has T observations. You set half the data for training and half for testing initially. Now you want to increase the number of data points for training T1, T2 … Tn where T1 < T2…. Tn-1 < Tn.

A) The difference between training error and test error increases as number of observations increases  
B) The difference between training error and test error decreases as number of observations increases  
C) The difference between training error and test error will not change  
D) None of These

Both can handle real valued features.

**10) Which of the following algorithm are not an example of ensemble learning algorithm?**

A) Random Forest  
B) Adaboost  
C) Gradient Boosting  
D) Decision Trees

**11) Suppose you are using a bagging based algorithm say a RandomForest in model building. Which of the following can be true?**

1. Number of tree should be as large as possible
2. You will have interpretability after using RandomForest

A) 1  
B) 2  
C) 1 and 2  
D) None of these

**Context 12-15**

Consider the following figure for answering the next few questions. In the figure, X1 and X2 are the two features and the data point is represented by dots (-1 is negative class and +1 is a positive class). And you first split the data based on feature X1(say splitting point is x11) which is shown in the figure using vertical line. Every value less than x11 will be predicted as positive class and greater than x will be predicted as negative class.

[Chart, scatter chart

Description automatically generated](https://cdn.analyticsvidhya.com/wp-content/uploads/2017/07/14112922/12_image_con.jpg)

**12) How many data points are misclassified in above image?**

A) 1  
B) 2  
C) 3  
D) 4

**13) Which of the following splitting point on feature x1 will classify the data correctly?**

A) Greater than x1  
B) Less than x1  
C) Equal to x1  
D) None of above

**14) If you consider only feature X2 for splitting. Can you now perfectly separate the positive class from negative class for any one split on X2?**

A) Yes  
B) No

**15) Now consider only one splitting on both (one on X1 and one on X2) feature. You can split both features at any point. Would you be able to classify all data points correctly?**

A) TRUE  
B) FALSE

**19) Which of the following is true about the Gradient Boosting trees?**

1. In each stage, introduce a new regression tree to compensate the shortcomings of existing model
2. We can use gradient decent method for minimize the loss function

A) 1  
B) 2  
C) 1 and 2  
D) None of these

**20) True-False: The bagging is suitable for high variance low bias models?**

A) TRUE  
B) FALSE

22)

Suppose, you are building a Gradient Boosting model on data, which has millions of observations and 1000’s of features. Before building the model you want to consider the difference parameter setting for time measurement.

**Consider the hyperparameter “number of trees” and arrange the options in terms of time taken by each hyperparameter for building the Gradient Boosting model?**

Note: remaining hyperparameters are same

1. Number of trees = 100
2. Number of trees = 500
3. Number of trees = 1000

A) 1~2~3  
B) 1<2<3

C) 1>2>3  
D) None of these

**25) [True or False] Cross validation can be used to select the number of iterations in boosting; this procedure may help reduce overfitting.**

A) TRUE  
B) FALSE

**26) When you use the boosting algorithm you always consider the weak learners. Which of the following is the main reason for having weak learners?**

1. To prevent overfitting
2. To prevent under fitting

A) 1  
B) 2  
C) 1 and 2  
D) None of these

**27) To apply bagging to regression trees which of the following is/are true in such case?**

1. We build the N regression with N bootstrap sample
2. We take the average the of N regression tree
3. Each tree has a high variance with low bias

A) 1 and 2  
B) 2 and 3  
C) 1 and 3  
D) 1,2 and 3

**28) How to select best hyperparameters in tree based models?**

A) Measure performance over training data  
B) Measure performance over cross validation data  
C) Both of these  
D) None of the

Suppose that you would like to get graph for that. Which one is more appropriate?

A picture containing chart

Description automatically generated

* 1. sales['Unit\_Cost'].plot(kind='hist')
  2. sales['Unit\_Cost'].plot(kind='bar')
  3. sales['Unit\_Cost'].plot(kind='scatter')
  4. sales['Unit\_Cost'].plot(kind='density')