**Q1.** **You are given a train data set having 1000 columns and 1 million rows. The data set is based on a classification problem. Your manager has asked you to reduce the dimension of this data so that model computation time can be reduced. Your machine has memory constraints. What would you do? (You are free to make practical assumptions.)**

**Q2.** **Is rotation necessary in PCA? If yes, Why? What will happen if you don’t rotate the components?**

**Q3.** **You are given a data set. The data set has missing values which spread along 1 standard deviation from the median. What percentage of data would remain unaffected? Why?**

**Q4. You are given a data set on cancer detection. You’ve build a classification model and achieved an accuracy of 96%. Why shouldn’t you be happy with your model performance? What can you do about it?**

**Q5**. **Why is naive Bayes so ‘naive’ ?**

**Q6. Explain prior probability, likelihood and marginal likelihood in context of naiveBayes algorithm?**

**Q7.** **You are working on a time series data set. You manager has asked you to build a high accuracy model. You start with the decision tree algorithm, since you know it works fairly well on all kinds of data. Later, you tried a time series regression model and got higher accuracy than decision tree model. Can this happen? Why?**

**Q8.** **You are assigned a new project which involves helping a food delivery company save more money. The problem is, company’s delivery team aren’t able to deliver food on time. As a result, their customers get unhappy. And, to keep them happy, they end up delivering food for free. Which machine learning algorithm can save them?**

**Q9.** **You came to know that your model is suffering from low bias and high variance. Which algorithm should you use to tackle it? Why?**

**Q10.** **You are given a data set. The data set contains many variables, some of which are highly correlated and you know about it. Your manager has asked you to run PCA. Would you remove correlated variables first? Why?**

**Q11. After spending several hours, you are now anxious to build a high accuracy model. As a result, you build 5 GBM models, thinking a boosting algorithm would do the magic. Unfortunately, neither of models could perform better than benchmark score. Finally, you decided to combine those models. Though, ensembled models are known to return high accuracy, but you are unfortunate. Where did you miss?**

**Q12.** **How is kNN different from kmeans clustering?**

**Q13.** **How is True Positive Rate and Recall related? Write the equation.**

**Q14. You have built a multiple regression model. Your model R² isn’t as good as you wanted. For improvement, your remove the intercept term, your model R² becomes 0.8 from 0.3. Is it possible? How?**

**Q15.** **After analyzing the model, your manager has informed that your regression model is suffering from multicollinearity. How would you check if he’s true? Without losing any information, can you still build a better model?**

**Q16. When is Ridge regression favorable over Lasso regression?**

**Q17.** **Rise in global average temperature led to decrease in number of pirates around the world. Does that mean that decrease in number of pirates caused the climate change?**

**Q18.** **While working on a data set, how do you select important variables? Explain your methods.**

**Q19.** **What is the difference between covariance and correlation?**

**Q20. Is it possible capture the correlation between continuous and categorical variable? If yes, how?**

**Q21. Both being tree based algorithm, how is random forest different from Gradient boosting algorithm (GBM)?**

**Q22. Running a binary classification tree algorithm is the easy part. Do you know how does a tree splitting takes place i.e. how does the tree decide which variable to split at the root node and succeeding nodes?**

**Q23.** **You’ve built a random forest model with 10000 trees. You got delighted after getting training error as 0.00. But, the validation error is 34.23. What is going on? Haven’t you trained your model perfectly?**

**Q24.** **You’ve got a data set to work having p (no. of variable) > n (no. of observation). Why is OLS as bad option to work with? Which techniques would be best to use? Why?**

**Q25. What is convex hull ? (Hint: Think SVM)**

**Q26.** **We know that one hot encoding increasing the dimensionality of a data set. But, label encoding doesn’t. How ?**

**Q27. What cross validation technique would you use on time series data set? Is it k-fold or LOOCV?**

**Q28.** **You are given a data set consisting of variables having more than 30% missing values? Let’s say, out of 50 variables, 8 variables have missing values higher than 30%. How will you deal with them?**

**29.** **‘People who bought this, also bought…’ recommendations seen on amazon is a result of which algorithm?**

**Q30.** **What do you understand by Type I vs Type II error ?**

**Q31**. **You are working on a classification problem. For validation purposes, you’ve randomly sampled the training data set into train and validation. You are confident that your model will work incredibly well on unseen data since your validation accuracy is high. However, you get shocked after getting poor test accuracy. What went wrong?**

**Q32.** **You have been asked to evaluate a regression model based on R², adjusted R² and tolerance. What will be your criteria?**

**Q33.** **In k-means or kNN, we use euclidean distance to calculate the distance between nearest neighbors. Why not manhattan distance ?**

**Q34.** **Explain machine learning to me like a 5 year old.**

**Q35.** **I know that a linear regression model is generally evaluated using Adjusted R² or F value. How would you evaluate a logistic regression model?**

**Q36.** **Considering the long list of machine learning algorithm, given a data set, how do you decide which one to use?**

**Q37.** **Do you suggest that treating a categorical variable as continuous variable would result in a better predictive model?**

**Q38.** **When does regularization becomes necessary in Machine Learning?**

**Q39.** **What do you understand by Bias Variance trade off?**

**Q40.** **OLS is to linear regression. Maximum likelihood is to logistic regression. Explain the statement.**