Machine Learning

3. Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.

4. The Gini impurity measure is one of the methods used in decision tree algorithms to decide the optimal split from a root node, and subsequent splits. It is the most popular and the easiest way to split a decision tree and it works only with categorical targets as it only does binary splits.

5. In the case of decision tree's they can learn a training set to a point of high granularity that makes them easily overfit. Allowing a decision tree to split to a granular degree, is the behavior of this model that makes it prone to learning every point extremely well — to the point of perfect classification

6. Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. This has been the case in a number of machine learning competitions, where the winning solutions used ensemble methods

7. Bagging and boosting are two main types of ensembles learning methods. The main difference between these learning methods is the way in which they are trained. In bagging, weak learners are trained in parallel, but in boosting, they learn sequentially.

8. The out-of-bag error is the average error for each predicted outcome calculated using predictions from the trees that do not contain that data point in their respective bootstrap sample. This way, the Random Forest model is constantly being validated while being trained.

9. k-Fold Cross-Validation Cross validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into.

10. Hyperparameter tuning in machine learning is a technique where we tune or change the default parameters of the existing model or algorithm to achieve higher accuracies and better performance. Sometimes when we use the default parameters of the algorithms, it does not suit the existing data as the data can vary according to the problem statement.

12. Yes, it might work, but logistic regression is more suitable for classification task and we want to prove that logistic regression yields better results than linear regression. Let's see how logistic regression classifies our dataset. Logistic regression model, a sigmoid curve that fit the training dataset

13. The main differences, therefore, are that Gradient Boosting is a generic algorithm to find approximate solutions to the additive modeling problem, while AdaBoost can be seen as a special case with a particular loss function. Hence, Gradient Boosting is much more flexible.

14. The bias-variance tradeoff is a fundamental concept in machine learning and statistics that relates to the balance between the complexity of a model and its ability to generalize to new, unseen data. A model with high bias is too simplistic and underfits the data, while a model with high variance is too complex and overfits the data.

15. Support Vector Machines (SVMs) are a type of supervised learning algorithm that can be used for classification or regression tasks. The main idea behind SVMs is to find a hyperplane that maximally separates the different classes in the training data