Machine learning

1. R-squared is a better measure of the goodness of fit model in regression. It is statistical measure that determines proportion of variance in dependent variable explained by independent variable. It shows how data will fit in regression model.
2. Total sum of squares(TSS) is sum of squared differences between observed dependent variables.
   Explained sum of squares(ESS) is sum of differences between predicted value and mean of dependent variable.
   Residual sum of squares(RSS) is used to measure amount of variance in a data that is not explained by regression model.
   TSS=RSS+ESS
3. Regularization is to reduce model complexity to make new data. Its improves performance on unseen datasets. Also helps to reduce impact of noisy data.
4. Gini impurity is opposite of entropy. It is try to find lowest impurity.
5. Yes, Unregularized decision trees prone to overfitting. Overfitting can lead to poor generalization to unseen data.
6. Ensemble technique is combining multiple KNN models. (Decision tree).
7. In ensemble there are two types. Bagging and Boosting
   Bagging: It is simplest way of combine predictions belongs to same type.
   i) Bagging classifier ii) Random forest
   Boosting: It is a way of combine predictions belong to different types.
   i) Adaboost ii) G Boost iii) Extreme gradient boosting.
8. Out of bag (oob) is used for measuring prediction error of random forests.
9. K fold means split data set into K no. of subsets.
10. Huper parameter tuning directly control model structure, function and performance. It helps to over fitting.
11. Gradient decent can over fit the training data if the model is complex.
12. No, we cannot use logistic regression for classification of non-linear data because it is used to describe data. It is a method for linear classification problems.
13. Adaboost is used for binary classification problems. It boost the performance of decision tree. Gradient boost is used to solve differentiable loss function problem.
14. Bias variance trade of machine learning refer to balance between two sources of error bias and variance.
15. Linear is like a line. It follows sequence or order in single dimension.
16. RBF Radial basis function Values depends on distance between input and some fixed points.
17. Polynomial kernels used in support vector machine (SVM) handles high dimensional data sets.