# ***Machine learning***

**1.R-squared or Residual Sum of Squares (RSS) which one of these two is a better measure of goodness of fit model in regression and why?**

**Ans:** R-***squared is*** a goodness-of-fit measure for linear regression models.

This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively.

**2. What are TSS (Total Sum of Squares), ESS (Explained Sum of Squares) and RSS (Residual Sum of Squares) in regression. Also mention the equation relating these three metrics with each other.**

Ans: **Total Sum of Squares (TSS):**

The Total Sum of squares tells you how much variation there is in dependent variable.

**Total SS = Σ (Yi – mean of Y)$^2$.**

**Explained Sum of Squares (Ess):** The explained sum of squares (ESS) is the sum of the squares of the deviations of the predicted values from the mean value of a response variable, in a standard reg model.

ESS = $\sum (\hat{Y}i - \bar{Y})^2$,

where $\hat{Y}i$ is the predicted value of the response variable for observation i.

RSS = $\sum (Yi - \hat{Y}i)2$,

which is the sum of squared differences between the actual and predicted values of the response variable.

**3. What is the need of regularization in machine learning?**

**Ans:** Regularization is one of the most important concepts of machine learning.

- It is a technique to prevent the model from overfitting by adding extra information to it.
- Sometimes the machine learning model performs well with the training data but does not perform well with the test data.

**4.What is Gini–impurity index?**

**Ans:** Gini impurity is a measurement used to built Decision trees to determined how the features of a data set should splits the nodes to form a tree.

**5.Are unregularized decision trees prone o over fitting? If yes? why:**

**Ans:** *Decision trees are prone to overfitting if they are allow to grow to deep and complex.*

*Regularization in machine learning is the process of regularizing the parameters that constrain, regularizes, or shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model. Avoiding risk of overfitting*

**6.What is an ensemble technique in machine learning?**

**Ans:** Join multiple decision and take final decision.

Bagging and Boosting are ensemble techniques.

**7. What is the difference between Bagging and Boosting techniques?**

**Ans:** Bagging attempts to tackle the over-fitting issue.

- Bagging is decreasing the variance without increasing bias.
- If the classifier is unstable (high variance), then we need to apply bagging.
  Boosting tries to reduce bias.
- If the classifier is steady and straightforward (high bias), then we need to apply boosting.

**8. What is out-of-bag error in random forests?**

**Ans:** Out of bag error is a method of measuring the prediction error in Random Forest, Decision trees and other Machine learning models.

**9. What is K-fold cross-validation?**

**Ans:** K-fold cross-validation is a technique for evaluating models.

In k-fold, the dataset is divided into k subsets or folds. The model is trained and evaluated k times, using a different

folds as the validation set each time.

**10. What is hyper parameter tuning in machine learning and why it is done?**

**Ans:** Hyperparameter tuning is an essential part of controlling the behaviour of a machine learning model.

If we don't correctly tune our hyperparameters, our estimated model parameters produce suboptimal results, as they don't minimize the loss function. This means our model makes more errors.

**11. What issues can occur if we have a large learning rate in Gradient Descent?**

**Ans:** When the learning rate is too large, gradient descent can suffer from divergence. This means that weights increase exponentially, resulting in exploding gradients which can cause problems such as instabilities and overly high loss values.

**12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?**

**Ans:** Non-linear problems can't be solved with logistic regression.

Logistic regression is a type of linear model that predicts the probability of a binary outcome, such as yes or no, true or false, or 0 or 1.

**13. Differentiate between Ada boost and Gradient Boosting.**

**Ans:** Adaptive Boosting or AdaBoost, it minimises the exponential loss function that can make the algorithm sensitive to the outliers.

With Gradient Boosting, any differentiable loss function can be utilised. Gradient Boosting algorithm is more robust to outliers than AdaBoost.

**14. What is bias-variance trade off in machine learning?**

**Ans:** Bias-variance trade off describes the relationship between Models complexity, the accuracy of its predictions.

**15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.**

**Ans:** Linear kernel SVM assumes that the input data is linearly separable, whereas polynomial kernel SVM can handle non-linearly separable data by transforming it into a higher-dimensional space.

The Gaussian kernel, also known as the radial basis function (RBF) kernel, is a popular kernel function used in machine learning, particularly in SVMs (Support Vector Machines). It is a nonlinear kernel function that maps the

input data into a higher-dimensional feature space using a Gaussian function.

**The Gaussian kernel can be defined as:**

1. K(x, y) = exp(-gamma * ||x - y||^2)

=