MACHINE LEARNING

A1) The residual sum of squares (RSS) is the absolute amount of explained variation, whereas R-squared is the absolute amount of variation as a proportion of total variation.

RSS can be mathematically represented as:

$RSS = \sum^{n}_{i=1} (y^{i} - y_{i})^{2}$

R- squared can be mathematically represented as:

$R2=1-(\text{sum squared regression (SSR)/total sum of squares (SST)})$

$R2=1-(\sum(yi-{}^\wedge yi)2/\sum(yi-{}^{-}y)2)$

R squared is a better than RSS as it explains the variation with respect to variation of entire data.

A2) TSS (Total Sum of Squares): It is the sum of square of the deviation of the observed data from the mean of the entire data.

$$TSS = \Sigma(Yi - \text{mean of } Y)^{2}$$

$$TSS=ESS+RSS$$

Where,

Yi is the the observed data

ESS (Explained Sum of Squares): It is the sum of the squares of the deviations of the predicted values from the mean value of entire data.

$$ESS = \Sigma(Yp - \text{mean of } Y)^{2}$$

Where,

Yp is the predicted value

RSS (Residual Sum of Squares): It is the sum of the squares of the deviations of the predicted values from the actual values.

$$RSS= \Sigma(\text{actual value-predicted value})^{2}$$

A3) Regularization is a important in machine learning because it is used to reduce the errors by fitting the model appropriately on the given training set and avoid overfitting.

A4) Gini Impurity is a measurement used to build Decision Trees to determine how the features of a dataset should split nodes to form the tree. The feature split with least impurity is considered as the root node.

A5) Decision trees are prone to overfitting, especially when a tree is particularly deep. This is due to the amount of specificity we look at leading to smaller sample of events that meet the previous assumptions.

A6) 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. For example: Random forest and ada boost algorithms.

A7)

Bagging	**Boosting**
Bagging is a method of merging the same type of predictions.	Boosting is a method of merging different types of predictions.
Bagging decreases variance, not bias, and solves over-fitting issues in a model.	Boosting decreases bias, not variance.
Each model receives an equal weight.	Models are weighed based on their performance.
Models are built independently in Bagging.	New models are affected by a previously built model's performance in Boosting.
training data subsets are drawn randomly with a replacement for the training dataset.	every new subset comprises the elements that were misclassified by previous models.
Bagging is usually applied where the classifier is unstable and has a high variance.	Boosting is usually applied where the classifier is stable and simple and has high bias.

A8) The out-of-bag (OOB) error is the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample. This allows the Random Forest Classifier to be fit and validated whilst being trained.

A9) K-fold Cross-Validation is when the dataset is split into a K number of folds and is used to evaluate the model's ability when given new data. K refers to the number of groups the data sample is split into.

A10) Hyperparameter tuning consists of finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set.

Hyperparameter tuning is important because 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.

A11) If the learning rate is very large we will skip the optimal solution.

A12) Non-linear classification problems can't be solved with logistic regression because it has a linear decision surface.

A13) AdaBoost is the first designed boosting algorithm with a particular loss function. On the other hand, Gradient Boosting is a generic algorithm that assists in searching the approximate solutions to the additive modelling problem.

A14) The bias–variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters.

A15) Linear Kernel- It is the most basic type of kernel, usually one dimensional in nature. It proves to be the best function when there are lots of features. The linear kernel is mostly preferred for text-classification problems as most of these kinds of classification problems can be linearly separated.

RBF Kernel- It is one of the most preferred and used kernel functions in svm. It is usually chosen for non-linear data. It helps to make proper separation when there is no prior knowledge of data.

Polynomial Kernel- It is a more generalized representation of the linear kernel. It is not as preferred as other kernel functions as it is less efficient and accurate.