MACHINE LEARNING

1. Between -1 and 1
2. PCA
3. Linear
4. Logistic Regression
5. 2.205x old coefficient of ‘x’
6. Increase
7. Random Forests provide a reliable feature importance estimate
8. Principal Components are calculated using supervised learning techniques
   B) Principal Components are calculated using unsupervised learning techniques
   C) Principal Components are linear combinations of Linear Variables.
9. Identifying spam or ham emails
10. max_depth, min_samples_leaf
11. Outliers are that who comes outside the bounderies of a graph. Like the points that comes outer from lower and upper bound. For IQR it helps us to find the outliers of data and helped us to removed it from our data.
12. 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.
13. Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size. Every time you add a independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines.
14. In Normalisation, the change in values is that they are at a standard scale without distorting the differences in the values. Whereas, Standardisation assumes that the dataset is in Gaussian distribution and measures the variable at different scales, making all the variables equally contribute to the analysis
15. Cross-validation is a technique that allows us to utilize our training data better for training and evaluating the model. For example, while using cross-validation, you effectively use complete data for training the model. Cross-validation also helps in finding the best hyperparameter for the model.

STATISTICS

1. The CLT is a statistical theory that states that - if you take a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from that population will be roughly equal to the population mean.
2. Sampling is a method that allows us to get information about the population based on the statistics from a subset of the population (sample), without having to investigate every individual". Simple Random Sampling , Systematic Sampling.
3. The type one error it shows the false positive rate that is actually true in the population. Type 2 error it shows false negative rate that is actually false in the population.
4. A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme. The middle of the range is also known as the mean of the distribution.
5. Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency. Correlation is a statistical measure that indicates how strongly two variables are related.
6. Univariate statistics summarize only one variable at a time. Bivariate statistics compare two variables. Multivariate statistics compare more than two variables.
7. The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.
8. H0: defendant is innocent; • H1: defendant is guilty. H0 (innocent) is rejected if H1 (guilty) is supported by evidence beyond "reasonable doubt." Failure to reject H0 (prove guilty) does not imply innocence, only that the evidence is insufficient to reject it.
9. Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data are data about numeric variables (e.g. how many; how much; or how often). Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.
10. The IQR describes the middle 50% of values when ordered from lowest to highest. To find the interquartile range (IQR), first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.
11. A bell curve is a type of graph that is used to visualize the distribution of a set of chosen values across a specified group that tend to have a central, normal values, as peak with low and high extremes tapering off relatively symmetrically on either side.
12. Data visualization method.
13. The p value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P values are used in hypothesis testing to help decide whether to reject the null hypothesis.
14. f we toss a coin, there could be only two possible outcomes: heads or tails, and if any test is taken, then there could be only two results: pass or fail. This distribution is also called a binomial probability distribution.
15. Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests. A one-way

ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.