MACHINE LEARNING

1. C) between -1 and 1
2. B) PCA
3. A) linear
4. A) Logistic Regression
5. A) 2.205 × old coefficient of 'X'
6. C) decreases
7. A) Random Forests reduce overfitting

8. B) Principal Components are calculated using unsupervised learning techniques

C) Principal Components are linear combinations of Linear Variables.

9. A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index

D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

10. D) min_samples_leaf

11. Any data point less than lower bond or more than the upper bond is considered as Outliers. The difference between Q3 and Q1 is known as Inter Quartile Range. IQR = Q3-Q1.

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 is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected. It is calculated using the formula **$1 – [((1 – R^2) * (n – 1)) / (n – k – 1)]$** where R square is corelation coefficient square, n is number of data, k is number of independent data.

14. In normalization minimum and maximum number of features are used for scaling whereas in standardisation mean and standard deviation are used for scaling. Normalisation is used when features are of different scales and standardisation in used when we want to ensure zero mean and unit standard deviation. Normalisation is effected by outliers and standardisation is not much effected by outliers.

15. Cross validation is a statistical method used to estimate the performance of the machine learning model. It is used to protect against overfitting in a predictive model.

The advantage of cross validation is, it reduces the overfitting. In the method we split the dataset into multiple folds and train the algorithm on different folds. This prevents from overfitting.

The disadvantage of cross validation is, it increases the training time. As we train the model by splitting the dataset it increases the training time.

STATISTICS WORKSHEET-4

1. Central limit theorem is a statistical theory which states that when the large sample size has a finite variance, the samples will be normally distributed and mean of samples will be approximately equal to the mean of the whole population. The central limit theorem is important because it allows us to safely assume that the sampling distribution of mean will be normal in most cases.

2. Sampling is the process of selecting a number of cases from all the cases in a particular group.

Types of sampling – random sampling, systematic sampling, convenience sampling, cluster sampling, stratified sampling.

3. Type 1 error is known as false positive that is when we reject the correct nul hypothesis, whereas type 2 error is known as false negative that is when we fail to reject the correct nul hypothesis.

4. The term normal distribution means the data in the distribution is normally distributed i.e., the graph of normal is bell curved.

5. Correlation refers to the statistical relationship between the two entities. It measures the extend to which two variables are linearly related. Covariance is the measurement of relationships between two random variables and to what extent they change together.

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 of a test is the proportion of people who test positive among all those who actually have the disease. It is calculated by true positive divided by true positive + false negative.

8. Hypothesis testing is a statistical testing where we tests an assumption regarding a population parameter. H0 is null hypothesis and H1 is alternative hypothesis. H0 in two tail test is equal to 1 and H1 is not equal to 1.

9. Quantitative data are measures of values or counts and are expressed in numbers. Qualitative data are measures of types and are represented by names, symbol or number codes.

10. The range is calculated by subtracting lowest value from highest value. The difference between Q1 and Q3 is interquartile range.

11. If the graph is bell curved in a distribution, we can understand that the distribution is normally distributed.

12. Outliers can be found by interquartile range method.

13. The p value of hypothesis testing is 0.05.

14. The binomial formula is $P(x) = {}^{n}C_{x} \cdot p^{x} (1 - p)^{n-x}$ where n = total number of events, x = total number of successful events, p = probability of success on a single trial.

15. Analysis of variance is a statistical method used to test differences between two or more means. ANOVA is helpful for testing three or more variables. It is similar to multiple two sample t test. It results in fewer type 1 errors and is appropriate for a range of issues. It groups differences by comparing the means of each group and includes spreading out the variance into diverse sources.