MACHINE LEARNING

Q1. A)

Q2. D)

Q3. A)

Q4. D)

Q5. B)

Q6. C)

Q7. C)

Q8. A)

Q9. D)

Q10. D)

Q11.
Number of points clubbed in a group and some points belongs to group but outside the group are called outliers. IQR is used to **measure variability** by dividing a data set into quartiles. IQR is the range between the first and the third quartiles namely Q1 and Q3: *IQR = Q3 – Q1*. The data points which fall below *Q1 – 1.5 IQR* or above *Q3 + 1.5 IQR* are outliers.

Q12.
Bagging is the method of merging same types of predictions where as boosting is the method of merging different types of predictions. Bagging decrease variance, not bias just opposite boosting decrease bias, not variance.

Q13.
With the addition of new features in the model, it is not necessary that in the model will yield better result but R2 value will increase. To solve this problem we use Adjusted R2 value which penalizes the model use of excessive features which is not correlated with the output data.
Adjusted R squared is calculated by dividing the residual mean square error by the total mean square error (which is the sample variance of the target field).

Q14.
In Normalization- Minimum and maximum no. of features are used for scaling. It is affected by the Outliers. Scale value between 0 & 1 or -1 & 1. In standardization – mean and standard deviation is used for scaling. No range bound for scaling.

Q15.

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

Cross-validation prevents the models from overfitting the training dataset. Because of splitting the dataset in number of fold for training the model, Cross validation drastically increases the training time.

STATISTICS WORKSHEET-4

Q1.

The distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution. A sufficiently large sample size can predict the characteristics of a population more accurately. Sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold.

Q2.

It is a subset containing the characteristics of a larger population. 1) Simple random sampling is ideal if every entity in the population is identical. And 2) strata or proportional random sampling -divides the overall population into smaller groups.

Q3.

The difference between a type I error and a type II error is that a type I error rejects the null hypothesis when it is true

Q4.

In Normal distribution the data is symmetrically distributed with no skew. And the mean, mode and median appear at distinct point.

Q5.

Covariance measures the joint variability of two random variables, and Correlation is best used for multiple variables that express a linear relationship with one another.

Q6.

Univariate statistics summarize only one variable at a time. Bivariate statistics compare two variables. Multivariate statistics compare more than two variables.

Q7.

Sensitivity analysis works on the simple principle: **Change the model and observe the behavior.**

The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.

Q8.

Hypothesis Testing is a type of statistical analysis in which you put your assumptions about a population parameter to the test. It is used to estimate the relationship between 2 statistical variables.

The Null Hypothesis is the assumption that the event will not occur. The Alternate Hypothesis is the logical opposite of the null hypothesis. The acceptance of the alternative hypothesis follows the rejection of the null hypothesis.

Suppose a teacher evaluates the examination paper to decide whether a student passes or fails.

H0: Student has passed

H1: Student has failed

Type I error will be the teacher failing the student [rejects H0] although the student scored the passing marks

Type II error will be the case where the teacher passes the student [do not reject H0] although the student did not score the passing marks [H1 is true].

Q9.

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.

Q10.

To calculate range, subtract the largest value with the smallest one. For calculating interquartile range, 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.

Q11.

A bell curve is a common type of distribution for a variable, also known as the normal distribution.

Q12.

There are four ways to identify outliers: Sorting method. Data visualization method. Statistical tests (z scores) and Interquartile range method.

Q13.

The level of statistical significance is usually represented as a P-value between 0 and 1. The smaller the p-value, the more likely it is that you would reject the null hypothesis.

Q14.

The binomial distribution formula calculates the probability of getting x successes in the n trials of the independent binomial experiment.

Q15.

ANOVA-Analysis of Variance, test with more than one independent variables or factor.