**Work Sheet- 4**

**Machine Learning Assignment: 4**

1.The value of correlation coefficient will always be: -1 and +1 Ans: C

2. Which of the following cannot be used for dimensionality reduction? Ans:C

3. Which of the following is not a kernel in Support Vector Machines? Ans:C

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries? Ans:B

5. In a Linear Regression problem, ‘X’ is independent variable and ‘Y’ is dependent variable, where ‘X’ represents weight in pounds. If you convert the unit of ‘X’ to kilograms, then new coefficient of ‘X’ will be? Ans:C

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model? Ans:B

7.Which of the following is not an advantage of using random forest instead of decision trees? Ans:A

8. Ans: B,C

9. Ans: A,B,C,D

10. A,B,D

11. Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations.

We can use the IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3. Any values that fall outside of this fence are considered outliers. To build this fence we take 1.5 times the IQR and then subtract this value from Q1 and add this value to Q3. This gives us the minimum and maximum fence posts that we compare each observation to. Any observations that are more than 1.5 IQR below Q1 or more than 1.5 IQR above Q3 are considered outliers.

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 adjusts the statistic based on the number of independent variables in the model. R 2 shows how well terms (data points) fit a curve or line.

R2adj=1−[(1−R2)(n−1)n−k−1]

14. Normalization Standardization

1. Minimum and maximum value of features are used for scaling Mean and standard deviation is used for scaling.

2. It is used when features are of different scales. It is used when we want to ensure zero mean and unit standard deviation.

3. Scales values between [0, 1] or [-1, 1]. It is not bounded to a certain range.

4. It is really affected by outliers. It is much less affected by outliers.

15. Cross-validation (CV) is a technique used to assess a machine learning model and test its performance (or accuracy). It involves reserving a specific sample of a dataset on which the model isn't trained. Later on, the model is tested on this sample to evaluate it.

Advantage: Cross-validation is used to protect a model from overfitting, especially if the amount of data available is limited. It's also known as rotation estimation or out-of-sample testing and is mainly used in settings where the model's target is prediction.

Disadvantage: Increases Testing and Training Time: CV significantly increases the training time required for an ML model

**Statistics Worksheet:4**

1. The central limit theorem states that the sampling distribution of the sample mean approaches a normal distribution as the sample size gets larger no matter what the shape of the population distribution. The central limit theorem is important because it is used in hypothesis testing and also to calculate confidence intervals.
2. sampling technique is the process of studying the population by gathering information and analyzing that data. It is the basis of the data where the sample space is enormous. There are 2 types of sampling, Probability sampling and non-probability sampling.
3. Type I error is an error that takes place when the outcome is a rejection of null hypothesis which is, in fact, true. Type II error occurs when the sample results in the acceptance of null hypothesis, which is actually false.
4. Normal Distribution, also called the Gaussian Distribution, is the most significant continuous probability distribution. Sometimes it is also called a bell curve. A large number of random variables are either nearly or exactly represented by the normal distribution, in every physical science and economics. Furthermore, it can be used to approximate other probability distributions.
5. Correlation and covariance are two statistical concepts used to determine the relationship between two random variables. Correlation defines how a change in one variable will impact the other, while covariance defines how two items vary together.
6. Univariate analysis looks at one variable, Bivariate analysis looks at two variables and their relationship. Multivariate analysis looks at more than two variables and their relationship.
7. Sensitivity: The term sensitivity was introduced by Yerushalmy in the 1940s as a statistical index of diagnostic accuracy. It is also called the true positive rate, the recall, or probability of detection. It has been defined as the ability of a test to identify correctly all those who have the disease, which is “true-positive”. Sensitivity = No of true positives/ (No of true positives + No of false negatives).
8. Hypothesis testing in statistics refers to analyzing an assumption about a population parameter. It is used to make an educated guess about an assumption using statistics. With the use of sample data, hypothesis testing makes an assumption about how true the assumption is for the entire population from where the sample is being taken. The null hypothesis (H0) is a statement of no difference, no association, or no treatment effect. The alternative hypothesis, Ha is a statement of difference, association, or treatment effect. H0 is assumed to be true until proven otherwise. However, Ha is the hypothesis the researcher hopes to bolster.
9. Quantitative data is countable or measurable, relating to numbers; qualitative data is descriptive, relating to words. Quantitative data lends itself to statistical analysis; qualitative data is grouped and categorized according to themes.
10. Interquartile range = Upper Quartile – Lower Quartile = Q­3 – Q­1
11. A bell curve is a common type of distribution for a variable, also known as the normal distribution. The term "bell curve" originates from the fact that the graph used to depict a normal distribution consists of a symmetrical bell-shaped curve.

The highest point on the curve, or the top of the bell, represents the most probable event in a series of data (its mean, mode, and median in this case), while all other possible occurrences are symmetrically distributed around the mean, creating a downward-sloping curve on each side of the peak. The width of the bell curve is described by its standard deviation.

1. Hypothesis Testing, Boxplot.
2. The P-value is known as the probability value. It is defined as the probability of getting a result that is either the same or more extreme than the actual observations. The P-value is known as the level of marginal significance within the hypothesis testing that represents the probability of occurrence of the given event. The P-value is used as an alternative to the rejection point to provide the least significance at which the null hypothesis would be rejected. If the P-value is small, then there is stronger evidence in favour of the alternative hypothesis.
3. The binomial probability formula for any random variable x is given by

P (x : n, p) = nCx px qn-x

1. Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

Applications: Retail: Store are often interested in understanding whether different types of promotions, store layouts, advertisement tactics, etc. lead to different sales. ...

Medical: Researchers are often interested in whether or not different medications affect patients differently, which is why they often use one-way or two-way ANOVA’s in these situations.