WORKSHEET – 6

MACHINE LEARNING

1.D] None of the above

2.B] Decision trees are highly prone to overfitting

3.A] SVM

4.A] Accuracy

5.B] Model B

6.D] Lasso

7.A] Adaboost

8.A] Pruning

9.B] A tree in the ensemble focuses more on the data points on which the previous tree was not performing well

10. 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. Typically, the adjusted R-squared is positive, not negative. It is always lower than the R-squared. The adjusted R-squared compensates for the addition of variables and only increases if the new predictors enhances the model above what would be obtained by probability. Conversely, it will decrease when a predictor improves the model less than what is predicted by chance.

11. LASSO REGRESSION is a modification of linear regression, where the model is penalized for the sum of absolute values of the weights. Thus, the absolute values of weight will be reduced, and many will tend to be zeros. **RIDGE REGRESSION** takes a step further and penalizes the model for the sum of squared value of the weights. Thus, the weights not only tend to have smaller absolute values, but also really tend to penalize the extremes of the weights, resulting in a group of weights that are more evenly distributed.

12. A variance inflation factor is a tool to help identify the degree of multicollinearity. Multiple regression is used when a person wants to test the effect of multiple variables on a particular outcome. The dependent variables is the outcome that is being acted upon by the independent variables the inputs into the model. Multicollinearity exists when there is a linear relationship, or correlation, between one or more of the independent variables or inputs.

13. Similarly, in many machine learning algorithms, to bring all features in the same standing, we need to do scaling so that one significant number doesn't impact the model just because of their large magnitude. Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one. The most common techniques of feature scaling are Normalization and Standardization. Normalization is used when we want to bound our values between two numbers, typically, between

[0,1] or [-1,1]. While StandardizSation transforms the data to have zero mean and a variance of 1, they make our data unitless.

14. There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model they are **Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE).**

STATISTICS

1.d] All of the mentioned

2.a] Discrete

3.a] pdf

4.c] Mean

5.a] Variance

6.b] Standard deviation

7.c] 0 and 1

8.b] Bootstrap

9.b] Summarized data

10. Histogram and box plots are graphical representation for the frequency of numeric data values. They aim to describe the data and explore the central tendency and variability before using advanced statistical analysis techniques. Both histogram and boxplots allow to visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points. Both histograms and boxplots are used to explore and present the

data in an easy and understandable manner. **Histogram** are preferred to determine the underlying probability distribution of a data. **Boxplot** on the other hand are more useful when comparing between several data sets. They are less detailed than histograms and take up less space.

11. Prioritize objectives, examine which metric consistently predicts their achievement and identify which activities influence predictors in that order. And continuously re-evaluate this process to keep up with the times.

12. Sample size is an important component of statistical significance in that larger sample are less prone and flukes. Only randomly chosen, representative samples should be used in significance testing. The level at which one can accept whether an event is statistically significant is known as significance level. Researcher use a measurement known as the p-value to determine statistical significance, if the p-value falls below the significance level, then the result is statistically significant. The p-value is a function of the means and standard deviations of the data samples. The p-value indicates the probability under which the given statistical result occurred, assuming chance alone is responsible for the result. If this probability is small, then the researcher can conclude that some other factor could be responsible for the observed data.

14. Income is the classic example of when to use the median instead of mean because its distribution tends to be skewed.

15. The likelihood function represents the probability of random variable realizations conditional on particular values

of the statistical parameter. Thus, when evaluated on a given sample, the likelihood function indicates which parameter values are more likely than other, in the sense that they would have made the observed data more probable.