First phase evaluation project:-

*Baseball Case Study:-*

***Article* - Baseball statistics** refers to a variety of metrics used to evaluate player and team performance in the game of [baseball](https://en.wikipedia.org/wiki/Baseball).

\***It’s a sports** case study where data for 2014 to develop an algorithm that predicts the number of wins for a given team in the 2015 season based on several different indicators of success. It will help the team management to replace the team member.

**\* Statistical analysis** reveals slight right skewness in feature distributions. Consistent counts signify a complete dataset. Potential outliers, especially in 'Errors' and 'Run Scored,' warrant attention. Heatmap highlights the need for feature scaling.

\* From **Data Analysis** we can easily find out outlier in ('Runs Scored', 'Earned Run Average', 'Shut Outs', 'Saves' and 'Errors') that columns.

\* From Skewness Checking we can get ('Runs Scored', 'Complete Game', 'Shut Outs' and 'Errors' ) columns have slightly skewed and it will need to be treated accordingly. Rest columns are **Normally Distributed.**

**\* df head** gives us an Optimal range for skewness is -0.5 to 0.5.

\* **Correlation of dataset** gives us the strength of a linear relationship between two random variables.

\* **Standard Scaling** gives us a technique wherein it makes the data scale free by converting the statistical distribution of the data set into the mean- 0 standard deviation – 1.

\* **PCA(Principal Component Analysis)** is used to reduce the dimensionality of large data sets, by transforming a large data set of the variables into a smaller one that still contain most of the information in the large set.

\* **Logistic Regression** predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes ,No, 0 or 1, true, false, etc, but instead of giving the exact values as 0 and 1 , it gives the probabilistic values which lie between 0 and 1.

\* **Hyperparameters** directly control the Model Structure, function, and performance.

\* **Saving Model** allows us to efficiently fit the linear models to non-linear data without explicitly transforming them to feature spaces where they are linear.

***Example:-***

**[import** joblib

joblib**.**dump(model,'Baseball\_model.pkl')]

***Thank You………..***