Predicting Baseball Team Wins Using Machine Learning

**1. Problem Definition**

In professional baseball, predicting a team's performance based on various statistics is a crucial aspect of strategic planning. Wins (W) are one of the most critical metrics to determine how well a team is performing during a season. Several factors, such as runs, home runs, strikeouts, and errors, contribute to the number of games a team wins. In this project, the objective is to develop a machine learning model to predict the number of wins a team can achieve based on these statistics. Using data from 30 baseball teams, we will analyze the relationships between these variables and build a predictive model.

**2. Data Analysis**

The dataset consists of 30 entries, each representing a baseball team with 17 features, including metrics like runs (R), hits (H), home runs (HR), and earned run average (ERA). Our target variable is the number of wins (W) for each team. Below is a brief overview of the key columns in the dataset:

- Wins (W): Total number of wins by the team.

- Runs (R): Total number of runs scored by the team.

- Home Runs (HR): Total number of home runs hit.

- Strikeouts (SO): Total number of strikeouts by the team.

- Errors (E): Number of errors committed by the team.

We aim to explore how these features affect the number of wins and identify the most significant predictors.

**3. EDA Concluding Remarks**

Exploratory Data Analysis (EDA) reveals several key insights:

- Teams that score more runs (R) tend to win more games, as expected.

- Teams with fewer errors (E) generally perform better, showing a negative correlation with wins.

- Home runs (HR) and strikeouts (SO) are moderately correlated with wins, but not as strongly as runs.

- Earned run average (ERA), which measures the effectiveness of a team's pitching, also shows a significant negative correlation with wins, as lower ERA values indicate better pitching performance.

These insights provide a foundation for feature selection and model building.

**4. Pre-processing Pipeline**

Before building the machine learning model, several preprocessing steps are required:

- Missing Data Handling- No missing data was observed in the dataset.

- Feature Scaling- To ensure that all features are on the same scale, we will apply standardization to normalize the values.

- Train-Test Split- The data will be split into training and testing sets in an 80-20 ratio to evaluate model performance on unseen data.

**5. Building Machine Learning Models**

We will apply several machine learning models to predict the number of wins (W). The models include

- Linear Regression- A basic model to capture linear relationships between wins and other features.

- Decision Tree Regressor- A non-linear model that creates decision nodes based on feature values.

- Random Forest Regressor- An ensemble learning method that builds multiple decision trees and averages their results.

- XGBoost Regressor- A gradient-boosting algorithm that iteratively improves model performance by minimizing errors. We will evaluate the models based on metrics such as Mean Absolute Error (MAE) and Rsquared (R²) to determine their accuracy.

**6. Concluding Remarks**

After evaluating multiple models, we found that ensemble methods like Random Forest and XGBoost performed better than simple linear regression. These models captured complex interactions between the features and wins, leading to improved predictions. However, models still showed some variance, indicating that factors beyond the available data could affect a team's performance. Overall, this project demonstrates the importance of data driven decision making in sports and provides a framework for predicting team performance using machine learning.