

Name: Hamza Javaid.

Roll No: SU92-BSAIM-F23-002.

Section: 3A

AI Lab Project Documentation

Submitted to: Sir Rasikh Ali

**Project name:**

Cricket Analytics of sixes in the PSL.

**Introduction**

This project is a data analysis and machine learning techniques which is used to predict whether a cricket inning from the Pakistan Super League (PSL) is dominated by sixes. Using performance data such as runs, balls, strike rate, and boundary statistics, we classify innings as "mostly sixes" or not.

**Purpose.**

The purpose of this project is to create a machine learning pipeline to analyse player performance and predict innings where sixes contribute to more than half the runs scored.

**1. Data Loading and Exploration**

The dataset Most sixes in an innings in PSL csv file contains records of player performances. We begin by loading the data using pandas and checking for missing values:

**2. Data Preprocessing**

Steps include:

1. Cleaning the Runs column by removing non-numeric characters (e.g., \*) and converting it to integers.
2. Creating a target column mostly sixes, defined as True if sixes account for more than 50% of the runs scored.

**3. Feature Selection**

We select the following features for training the model:

* Balls: Number of balls faced by the player.
* 4s: Number of fours hit.
* 6s: Number of sixes hit.
* SR: Strike rate.

The target variable is mostly sixes.

**Model Building**

**1. Splitting Data**

The dataset is split into training and testing sets, with **70%** used for training and **30%** for testing.

**2. Training the Model**

A Random Forest Classifier is trained on the data to predict the target variable.

**3. Evaluating the Model**

We use accuracy and a classification report to evaluate model performance:

**Results after analysing the data in the File which is imported**

**Performance Metrics**

| **Metric** | **Value** |
| --- | --- |
| **Accuracy** | 86% |
| **Precision** | 86% |
| **Recall** | 86% |
| **F1-Score** | 86% |

**Conclusion**

This project runs successfully and utilizes the machine learning to analyse and predict cricket dominated by sixes based on player performance statistics from the Pakistan Super League (**PSL**):

1. **Data Preparation**: The preprocessing steps ensured the dataset was clean and ready for analysis, with meaningful features and a well-defined target variable.
2. **Model Performance**: A Random Forest Classifier achieved an accuracy of 86%, with balanced precision, recall, and F1-score. This demonstrates the model's reliability in predicting whether sixes contributed to more than half of a player's total runs.
3. **Insights**: By leveraging features like balls faced, fours, sixes, and strike rate, the project highlights the significant impact of boundary statistics on player performance classification.
4. **Visualizations**: Using interactive tools like Plot enhanced the presentation of performance metrics, making the results more interpretable and engaging.