**PROJECT DOCUMENTATION**

**NAME:** Sreejaa R G

**BATCH:** May 2025

**MODE:** Offline

**ROLL NO:** 16525CBRE34

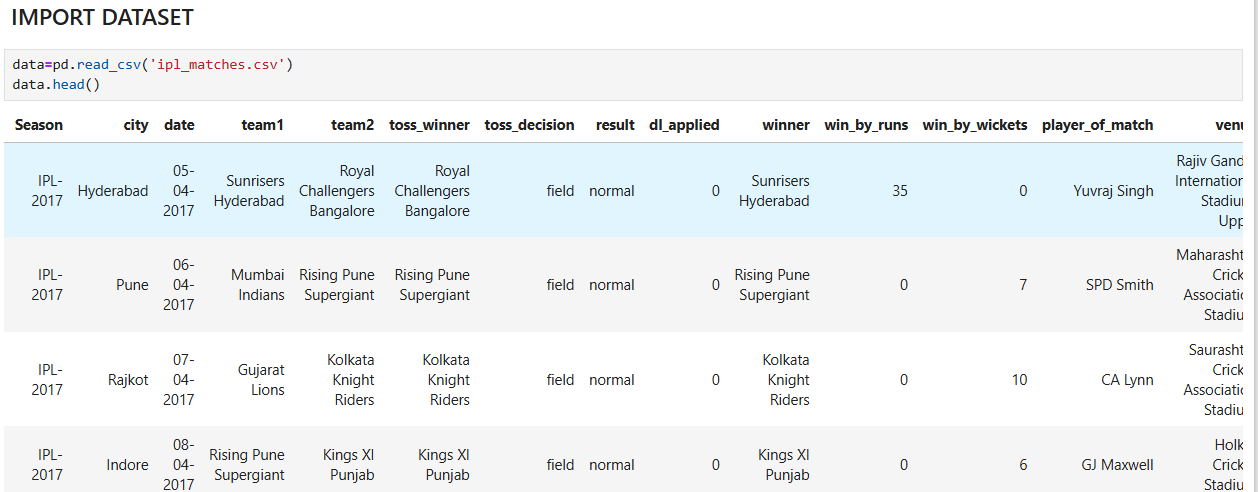
**TITLE:** IPL Match Winner Prediction

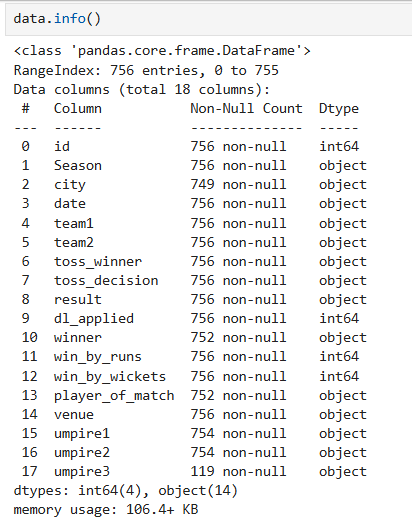
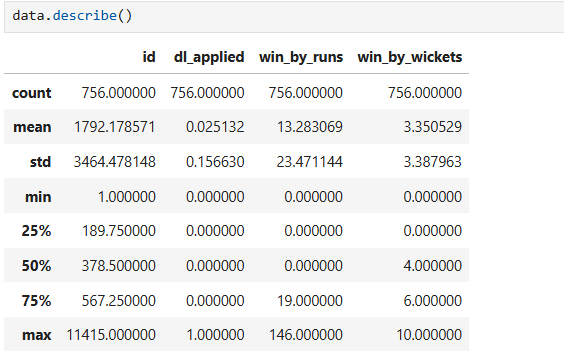
**1. Project Overview:**

The goal of this project is to build a machine learning model to predict the winner of an Indian Premier League (IPL) cricket match based on historical match data. This predictive model can be used for strategic planning, fan engagement, and sports analytics.

**2.Data Description:**

| **Column Name** | **Description** |
| --- | --- |
| id | Unique identifier for each match |
| season | The IPL season (year) in which the match was played |
| date | Date of the match |
| venue | Stadium or ground where the match was played |
| city | City in which the match was held |
| team1 | First competing team |
| team2 | Second competing team |
| toss\_winner | Team that won the toss |
| toss\_decision | Decision taken after toss (bat/field) |
| result | Result type (normal, tie, no result) |
| winner | Winning team of the match |
| win\_by\_runs | Margin of victory if decided by runs |
| win\_by\_wickets | Margin of victory if decided by wickets |
| player\_of\_match | Best performing player of the match |
| umpire1, umpire2 | Names of the match umpires |

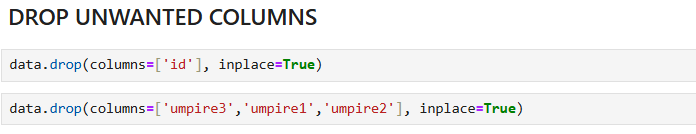


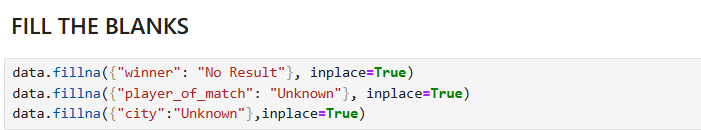
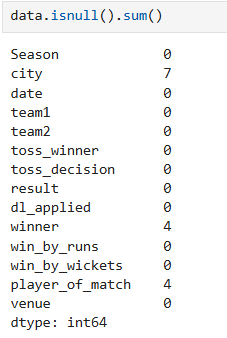


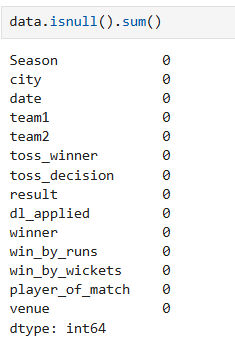
**3. Data Preprocessing:**

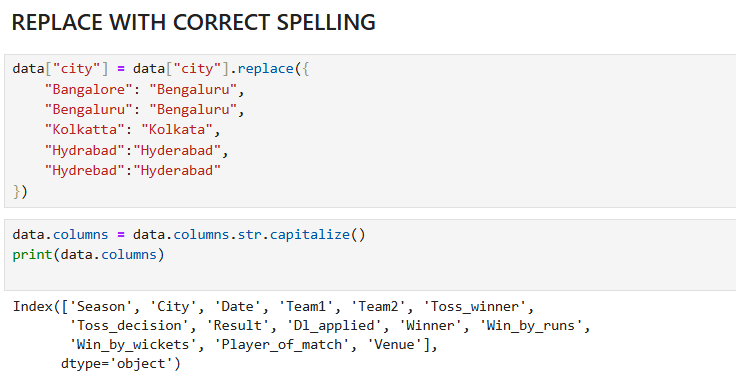
* **Handling Missing Values**:
  + Checked for null values in columns such as winner, player\_of\_match, and umpire fields.
  + Replaced missing values with "Unknown" (categorical) or 0 (numerical, e.g., win\_by\_runs).
* **Removing Duplicates**:
  + Verified and dropped duplicate records from both matches and ball-by-ball datasets.
* **Standardizing Team Names**:
  + Unified old team names (e.g., Delhi Daredevils → Delhi Capitals, Kings XI Punjab → Punjab Kings) to avoid inconsistencies.





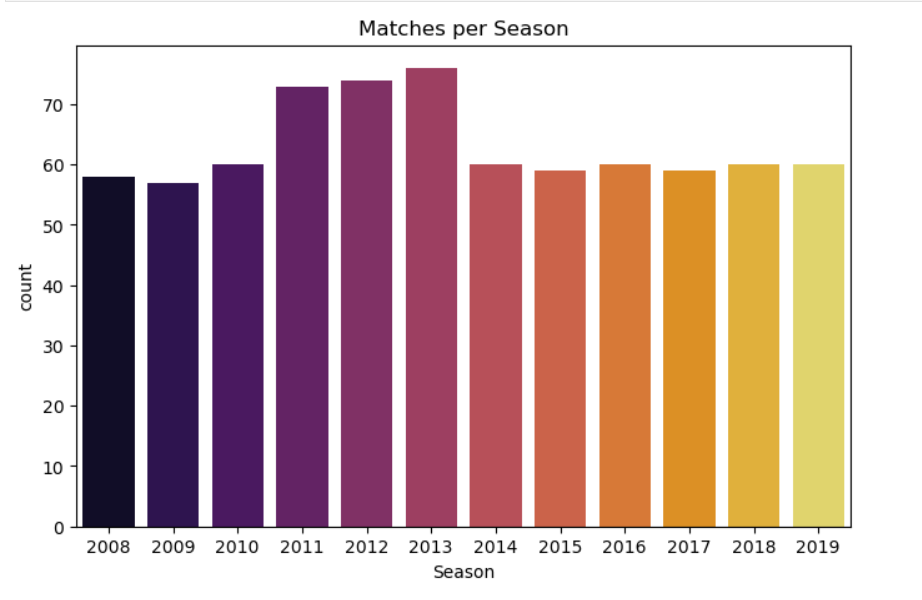


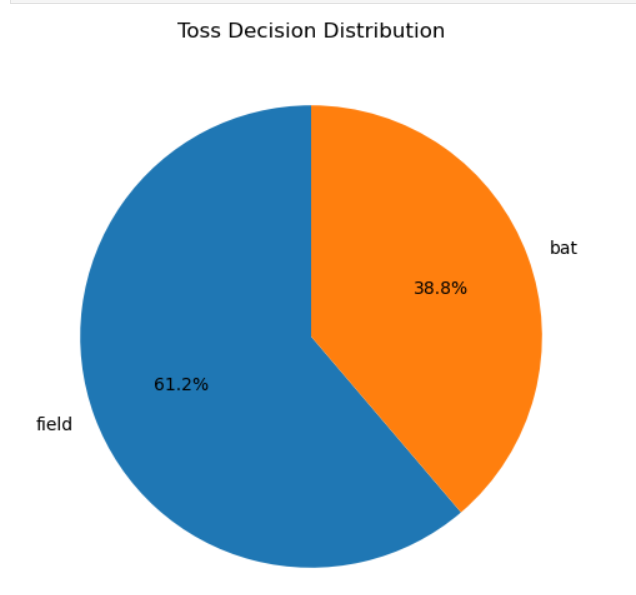


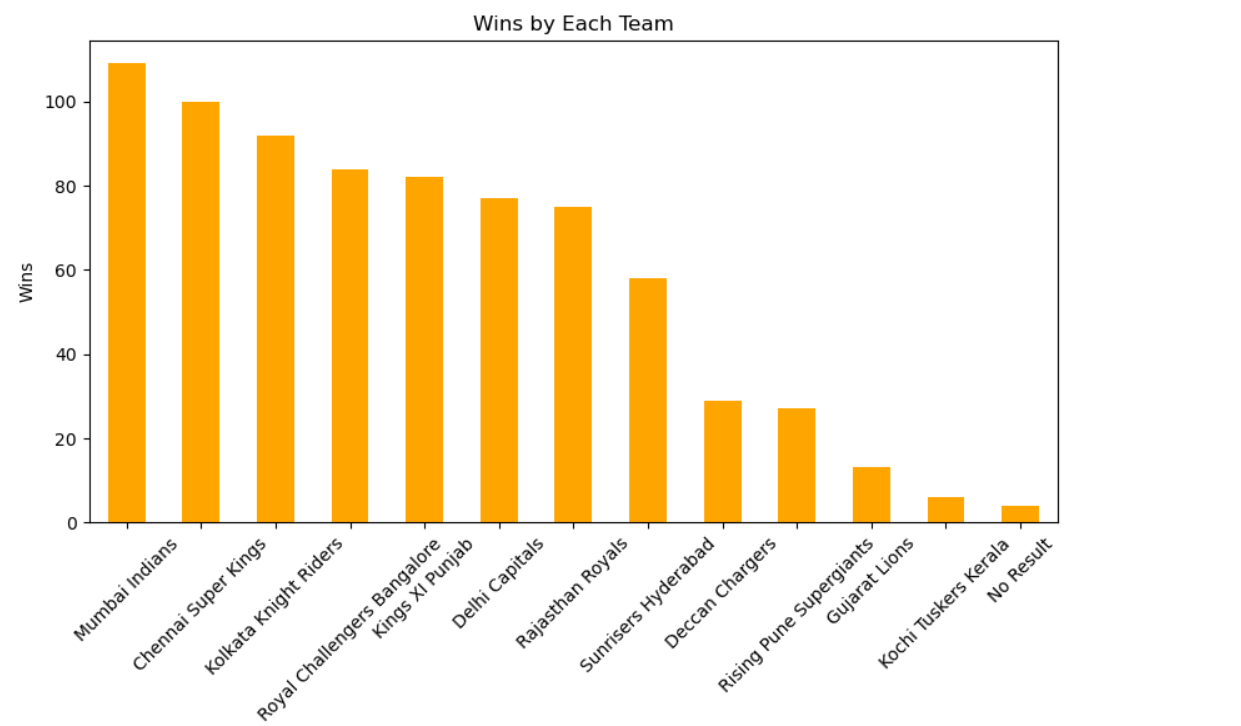


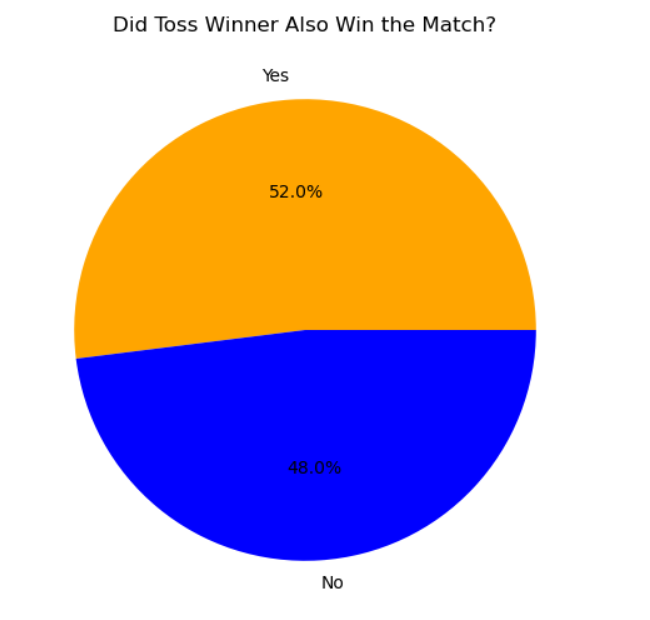
**4.Exploratory Data Analysis (EDA):**

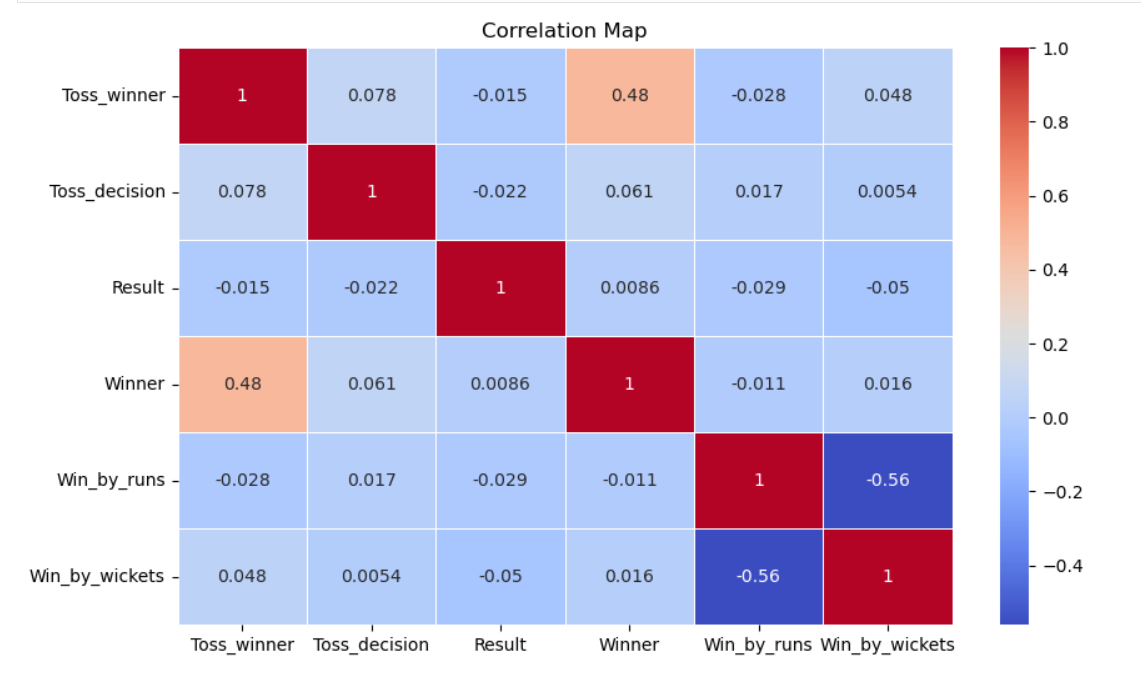
Exploratory Data Analysis (EDA) is the process of examining and visualizing the IPL dataset to uncover patterns, trends, and insights. It helps understand team performances, player contributions, and match outcomes before any predictive modeling.



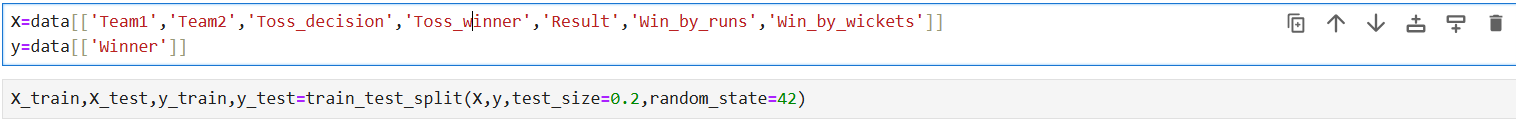








**5.Feature Selection:**



**Model Selection:**

**Logistic Regression** – Simple baseline for binary/multiclass classification.

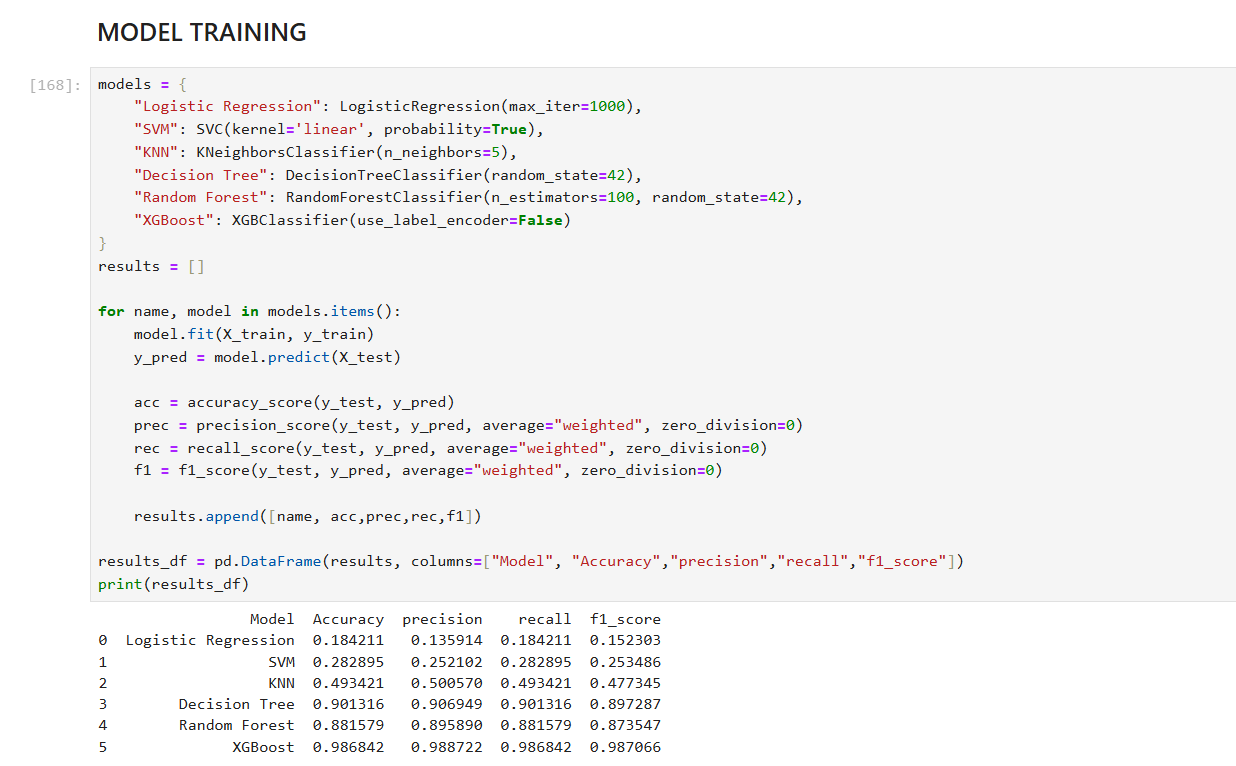
**Decision Tree Classifier** – Captures non-linear relationships between features.

**Random Forest Classifier** – Ensemble model improving accuracy and reducing overfitting.

**Support Vector Machine (SVM)** – Effective for classification with high-dimensional data.

**K-Nearest Neighbors (KNN)** – Based on similarity between feature vectors.

**XGBoost Classifier** – Gradient boosting algorithm for high performance on tabular data.



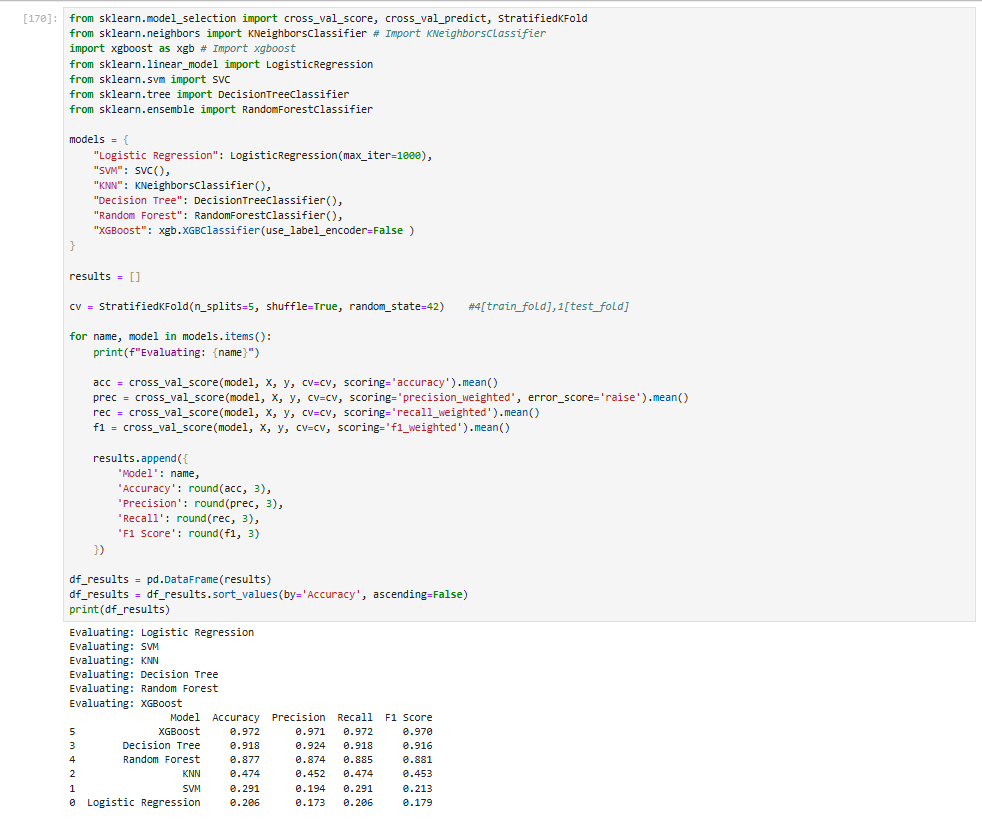
**6. Model Tuning & Evaluation:**

**Hyperparameter Tuning**: Adjust model parameters to improve performance.

* Decision Tree: Max depth, min samples split, criterion (gini/entropy).
* Random Forest: Number of trees (n\_estimators), max depth, min samples leaf.
* XGBoost: Learning rate, number of estimators, max depth, gamma.

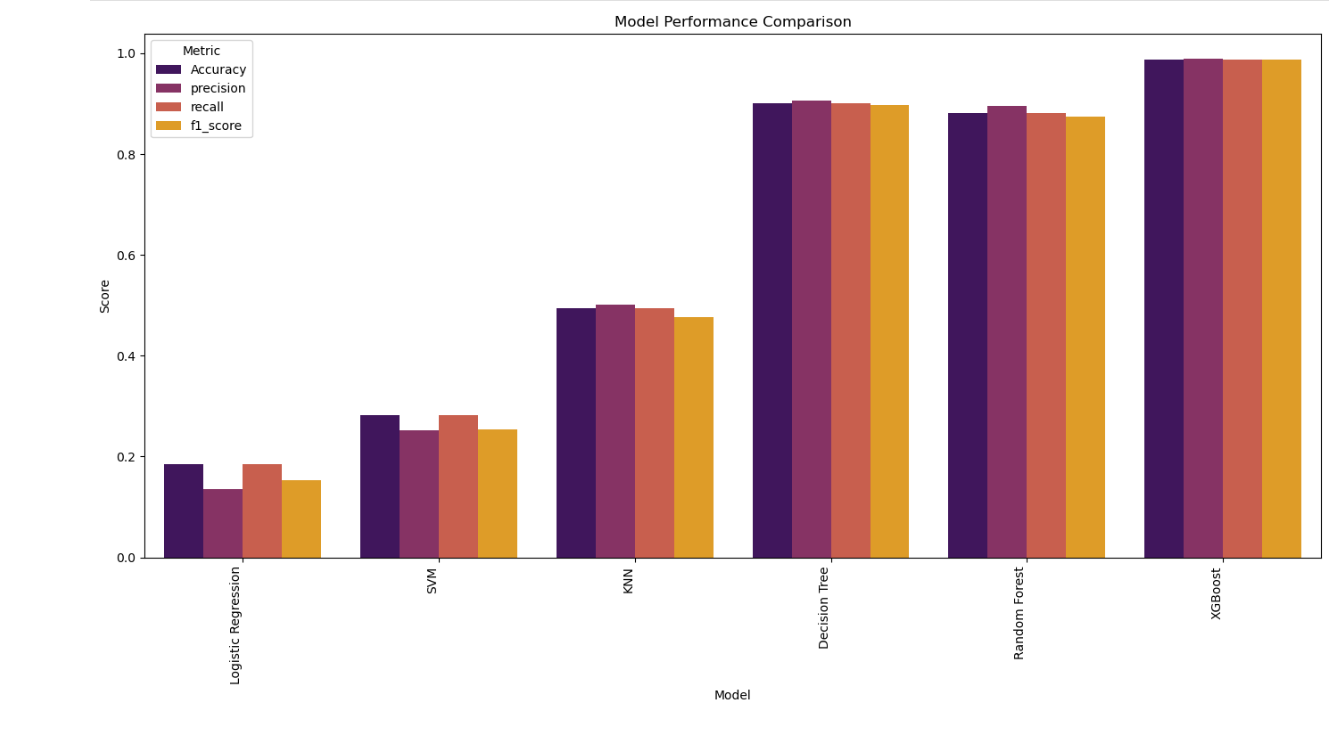
**Techniques Used**:

* Grid Search – Exhaustive search over parameter combinations.
* Random Search – Randomly sample parameters to reduce computation.
* Cross-Validation (k-fold) – Evaluate model stability across different data splits.



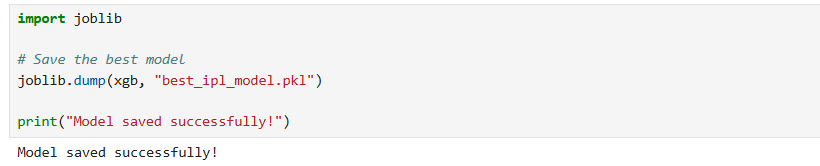
**Model Comparison:**

* Compare all models based on evaluation metrics



**Save the final model:**

save the final trained model in Python using pickle or joblib. This is typically done after training and tuning the best model ( XGBoost)



**Conclusion & Insights:**

The IPL dataset provides comprehensive information on matches, teams, and player performances across seasons. Analyzing this data helps uncover trends such as team strengths, top-performing players, and the influence of toss or venue on match outcomes. By applying preprocessing, exploratory analysis, and predictive modeling, meaningful insights can be drawn, and match outcomes can be predicted with reasonable accuracy. The dataset serves as a valuable resource for sports analytics, strategy planning, and data-driven decision-making in cricket.

* Mumbai Indians and Chennai Super Kings are the most consistent teams across seasons.
* Toss decisions can influence match outcomes, but team strength is more important.
* Top-performing players often win multiple Player of the Match awards, showing consistency under pressure.
* Most matches are closely contested, with occasional high-margin victories.
* Venue and season trends affect team performance, indicating strategic advantages.