**PROJECT DOCUMENTATION**

**EXPLORATORY DATA ANALYSIS & MACHINE LEARNING USING PYTHON**

|  |  |
| --- | --- |
| **TITLE:** | Exploratory Data Analysis and Match Outcome Prediction on IPL Data |
| **NAME:** | Arun Sundhar N |
| **COURSE:** | DA/DS, Offline |
| **BATCH:** | JUNE 2025 |

**TABLE OF CONTENT**

|  |  |
| --- | --- |
| 1. | **Introduction** |
| 2. | **Aim** |
| 3. | **Business Problem / Problem Statement** |
| 4. | **Project Workflow** |
| 5. | **Data Understanding** |
| 6. | **Data Cleaning** |
| 7a. | **Exploratory Data Analysis (EDA) - Univariate Analysis** |
| 7b. | **Bivariate Analysis** |
| 7c. | **Multivariate Analysis** |
| 8. | **Feature Engineering** |
| 9. | **Overall Insights from Analysis** |
| 10. | **Conclusion** |

## 1. Introduction

The Indian Premier League (IPL) is one of the most popular cricket tournaments worldwide. It generates a large amount of match data, including details about teams, toss decisions, venues, and match outcomes. Raw IPL datasets often contain missing values, duplicate entries, and inconsistencies that must be cleaned before meaningful insights can be extracted.

This project aims to clean, analyse, and model IPL data to identify patterns in match outcomes and evaluate machine learning models for predicting match winners.

## 2. Aim of the Project

* Clean and preprocess the IPL dataset.
* Explore match-level data visually using EDA.
* Identify patterns in toss decisions, team performance, and match outcomes.
* Engineer useful features to enhance predictive modelling.
* Build and compare multiple ML models to predict match winners.
* Select and save the best-performing model.

## 3. Problem Statement

Predicting cricket match outcomes is complex due to many influencing factors: toss decisions, teams, venues, and player performances. Teams, analysts, and betting markets seek insights into how these factors impact winning chances.

This project addresses:

* Does winning the toss significantly increase winning probability?
* Which teams dominate the IPL historically?
* Do venues and cities play a role in determining match results?
* Which machine learning models can best predict match winners from past data?

## 4. Project Workflow

1. Data Collection – Loaded the ipl\_matches.csv.xlsx dataset.

2. Data Cleaning – Handled missing values, dropped irrelevant columns, fixed data types.

3.Exploratory Data Analysis – Performed univariate, bivariate, and multivariate analysis.

4.Feature Engineering – Encoded categorical variables, derived new features (toss impact, home advantage).

5.Model Building – Trained Logistic Regression, Decision Tree, KNN, SVM, Random Forest, and XGBoost models.

6.Hyperparameter Tuning – Used GridSearchCV and RandomizedSearchCV for optimization.

7.Model Comparison – Evaluated models based on accuracy and confusion matrices.

8.Export – Saved best-performing model using joblib.

## 5. Data Understanding

Dataset contains the following columns:

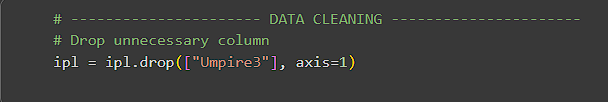
* **Season** – IPL season year
* **City** – City where match was played
* **Venue** – Stadium name
* **Team1** / **Team2** – Competing teams
* **Toss Winner** – Team that won the toss
* **Toss Decision** – Bat/Field decision after toss
* **Winner** – Team that won the match
* **Win by Runs** – Margin of victory (runs)
* **Win by Wickets** – Margin of victory (wickets)
* **Player of Match** – Best player of the match
* **Umpires** – Match officials

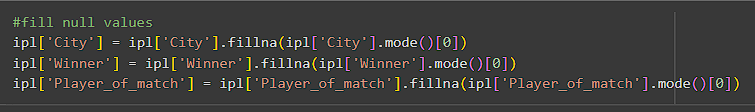
**Derived Features:**

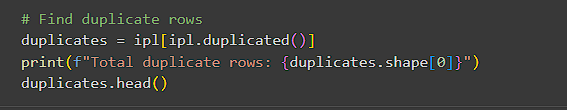
* **Team1\_encoded –** Numeric ID for the first team
* **Team2\_encoded** – Numeric ID for the second team
* **Toss\_winner\_encoded** – Which team won the toss
* **Toss\_decision\_encoded** – Toss decision (bat=1, field=0)
* **Toss\_match\_win** – Whether the toss-winning team also won the match (0/1)
* **Home\_team\_advantage** – Whether a team is playing at home (0/1)
* **Season\_encoded** – Season of the match as a number

## 6. Data Cleaning

* Dropped unnecessary column: Umpire1, Umpire2, Umpire3
* Filled missing values using **mode** (City, Winner, Player\_of\_match).
* Checked and removed duplicates.
* Converted categorical columns into numeric codes using Label Encoding.
* Created derived columns: Season Encoded, Toss Match Win (binary), Home Team Advantage.





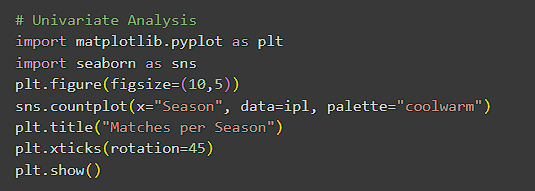


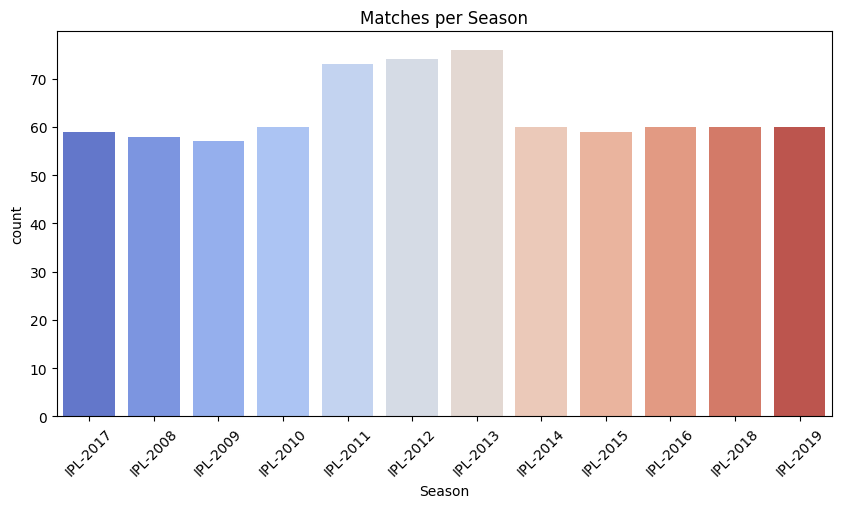
## 7. Exploratory Data Analysis

*7.1 Univariate Analysis:*

1. **Matches per Season (Countplot)** – Number of matches increased until 2013 and then stabilized.
2. **Top Players of the Match (Barplot)** – Players like Chris Gayle, AB de Villiers, and MS Dhoni dominate awards.
3. **Top Winning Teams (Barplot)** – Mumbai Indians and Chennai Super Kings emerge as the most successful franchises.

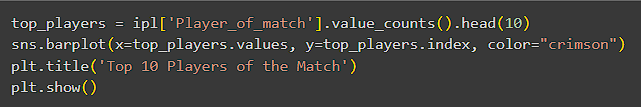
.

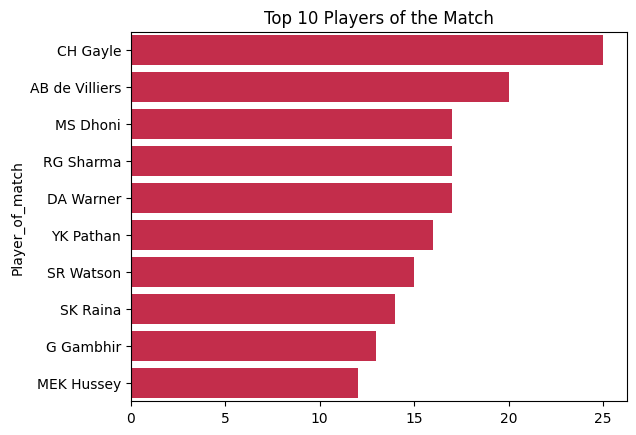




**Insights gained:**

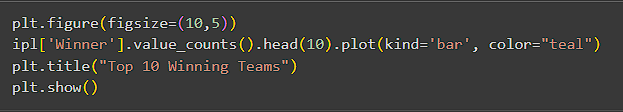
* Number of matches grew in early IPL years and stabilized after 2013.
* Confirms IPL’s expansion followed by consistent scheduling.

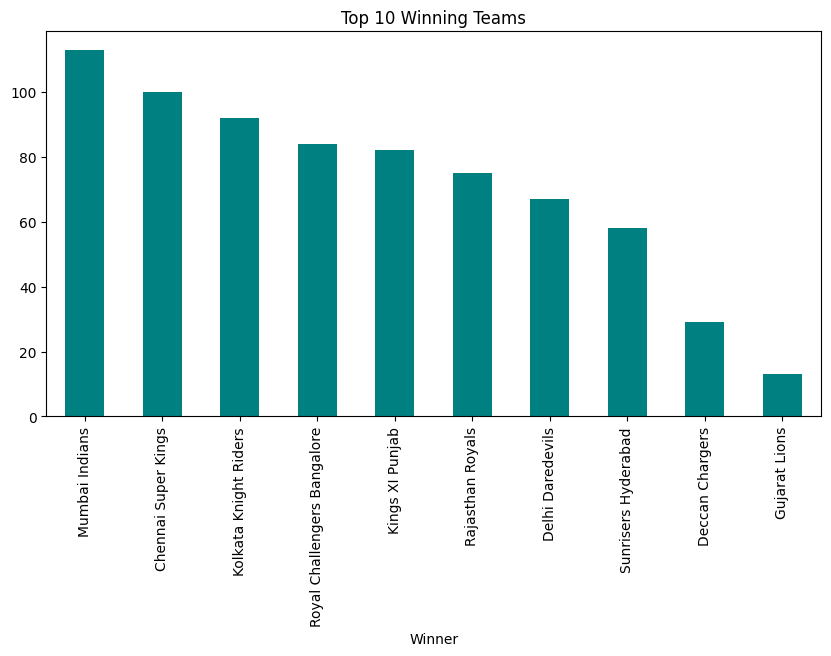




**Insights gained:**

* **Chris Gayle, AB de Villiers, and MS Dhoni** dominate the list.
* A few players consistently deliver match-winning performances**.**



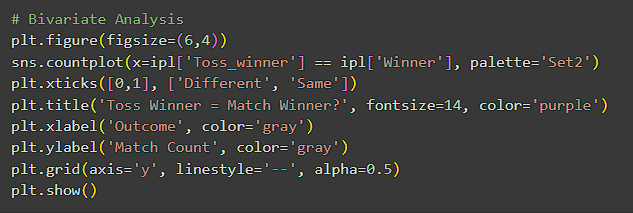


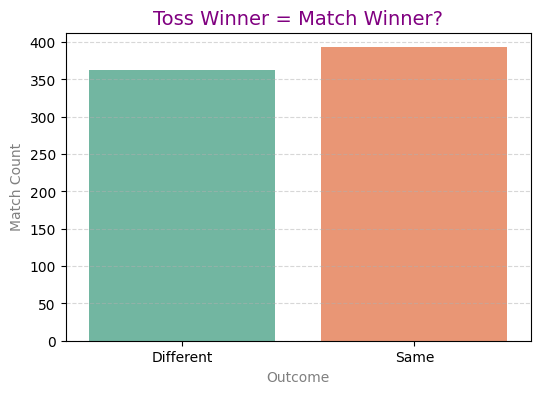
**Insights gained:**

* **Mumbai Indians (MI) and Chennai Super Kings (CSK)** lead in overall wins.
* Their consistency makes them the most successful IPL franchises.

*7.2 Bivariate Analysis*

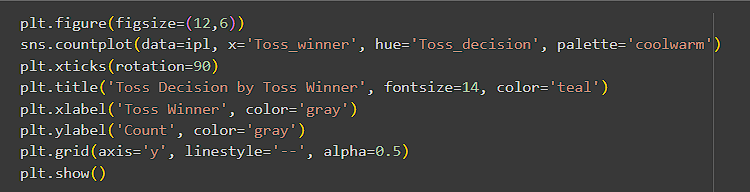
* **Toss Winner vs Match Winner (Countplot)** – Toss winners have a slight advantage, but not decisive (~50%+ win).
* **Toss Decision by Toss Winner (Countplot)** – Teams prefer fielding first more often.
* **Average Win by Runs vs Toss Decision (Barplot)** – Teams batting first win by larger run margins, while fielding teams win by wickets.

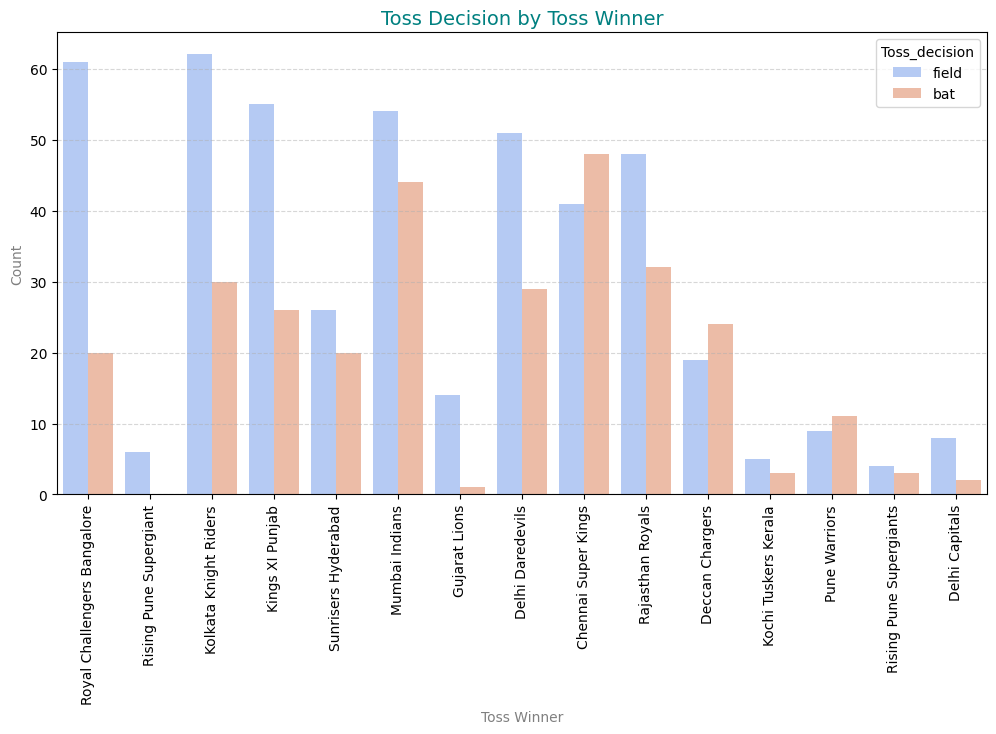




**Insights gained:**

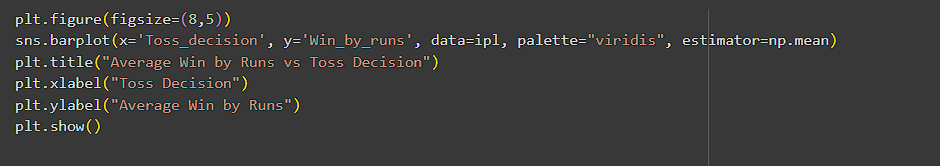
* **Toss winners** win slightly more often (~50%+).
* Toss provides an edge but does **not guarantee** victory.

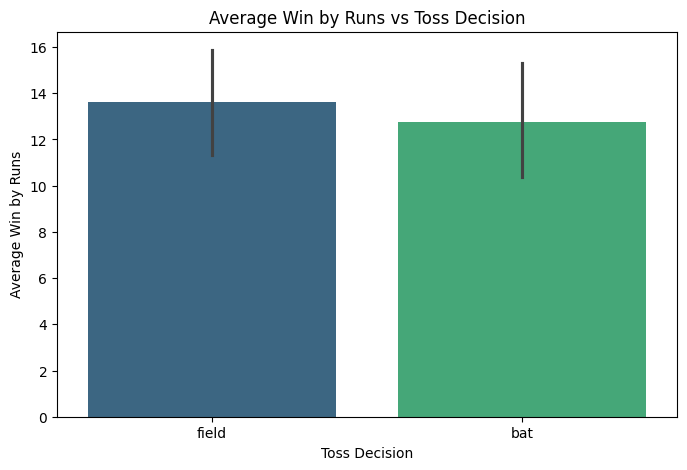




**Insights gained:**

* **Toss-winning** teams usually choose to **field** first.
* Reflects the chasing advantage trend in IPL.



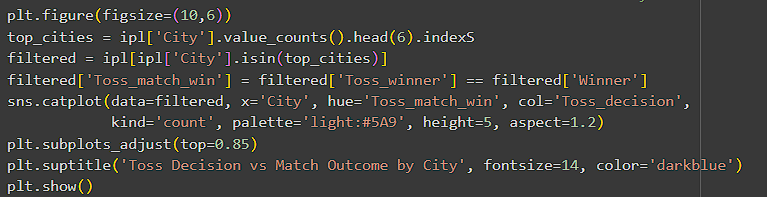
**

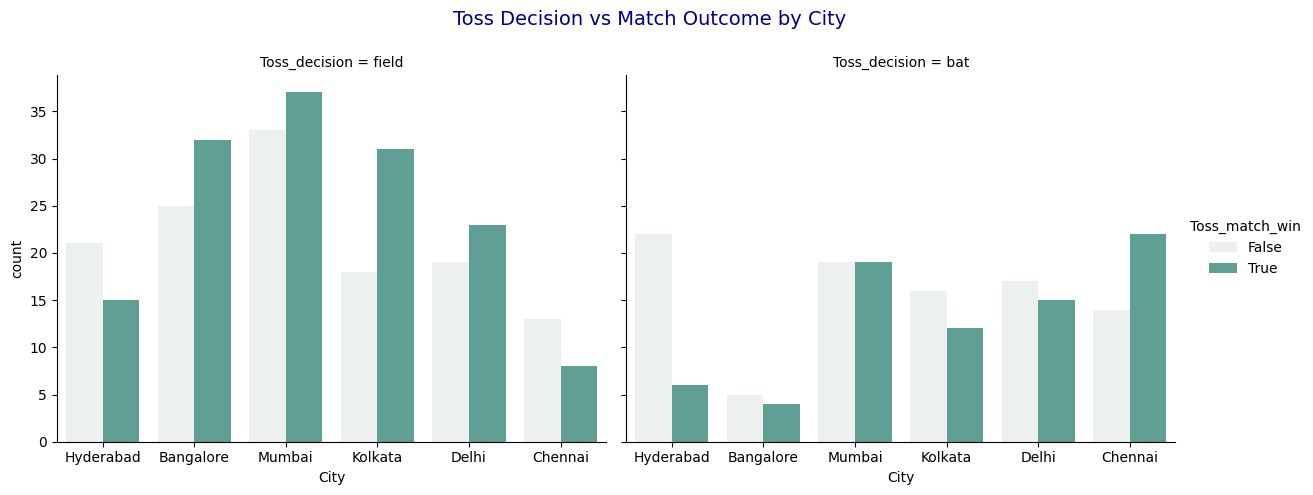
**Insights gained:**

* Teams that **bat first** win by larger **run margins**.
* Teams that **field first** usually win by **wickets while chasing**.

*7.3 Multivariate Analysis*

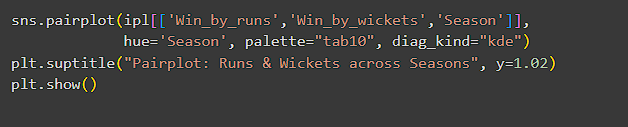
* **Toss Decision vs Match Outcome by City** – Cities like Mumbai and Bengaluru show stronger chasing success, proving venue conditions impact outcomes.
* **Pairplot (Win by Runs, Win by Wickets, Season)** – Win by runs has decreased in later seasons as chasing improved, while wicket wins remain stable (~6–8).
* **Correlation Heatmap** – Win by runs and win by wickets are negatively correlated, while season has weak correlation with match outcomes.

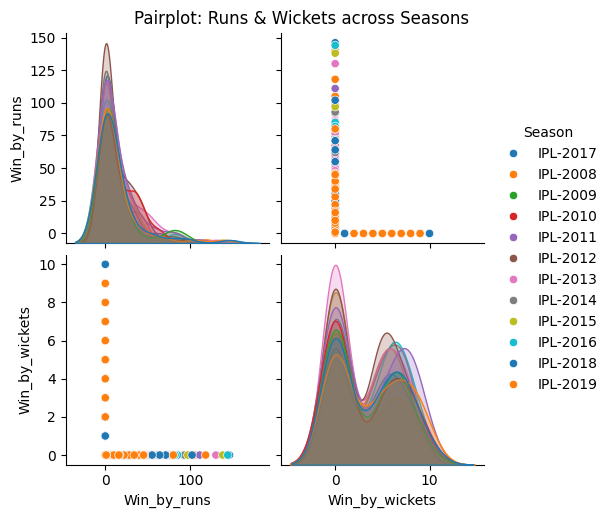




**Insights gained:**

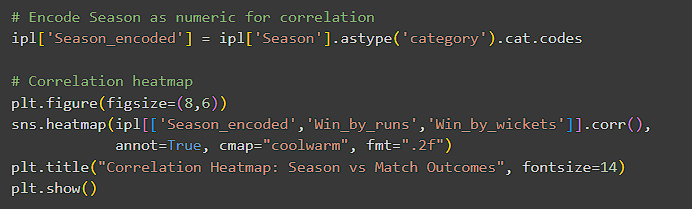
* Cities like **Mumbai and Bengaluru** show stronger chasing success.
* **Venue** conditions play a significant role in match outcomes.

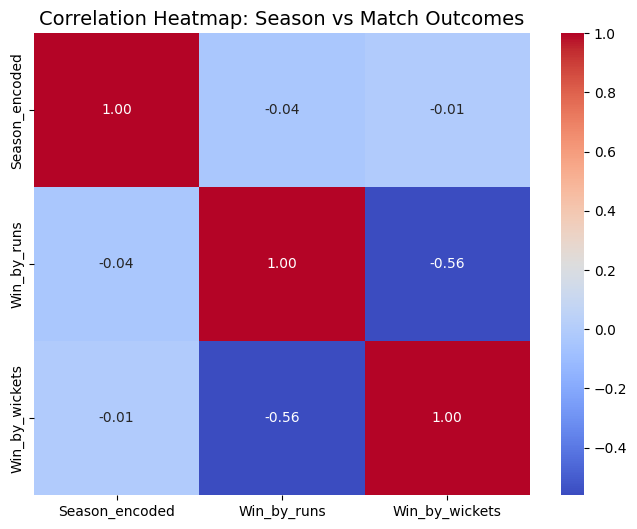




**Insights gained:**

* **Win by Runs** has decreased over seasons as chasing has improved.
* **Wicket wins** remain stable (~6–8 wickets), showing consistency in chasing.





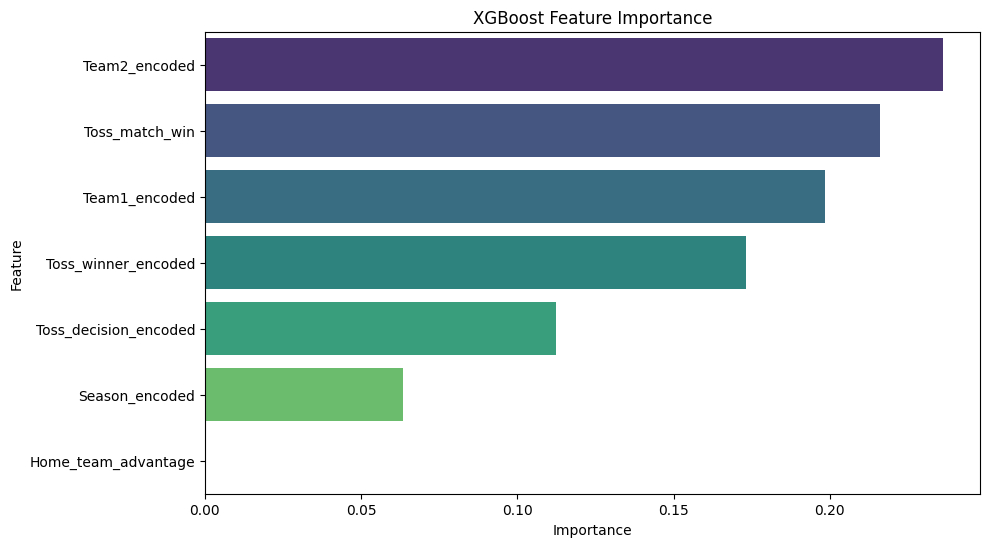
**Insights gained:**

* **Win by Runs and Win by Wickets** are strongly **negatively correlated.**
* **Season** shows **weak correlation,** meaning team strength and toss decisions matter more.

**8. Feature Engineering**

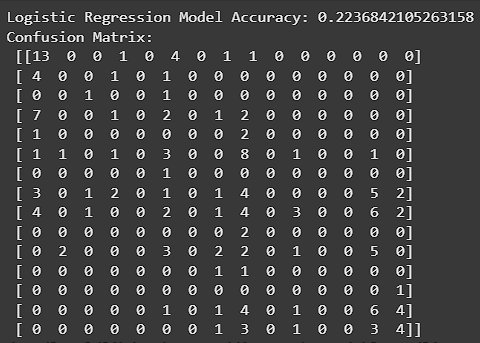
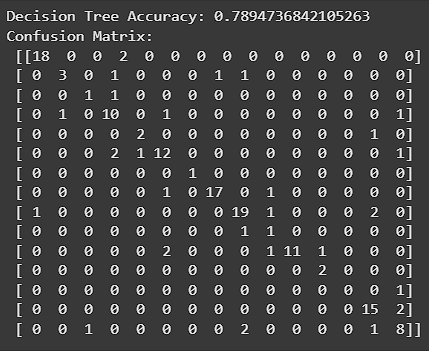
* Encoded Toss Decision (bat=1, field=0).
* Encoded Team1, Team2, Toss Winner, and Match Winner.
* Derived Toss Impact (1 if toss winner = match winner, else 0).
* Created Home Team Advantage column.
* Encoded Season numerically.

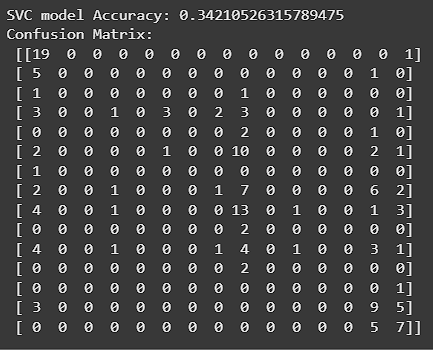
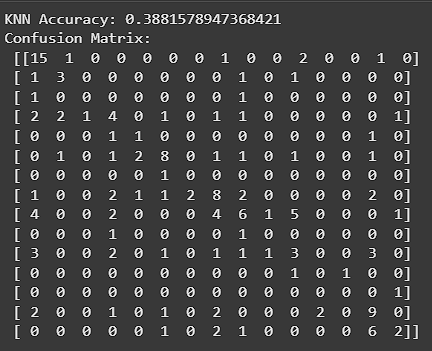
**FEATURE IMPORTANCE:**

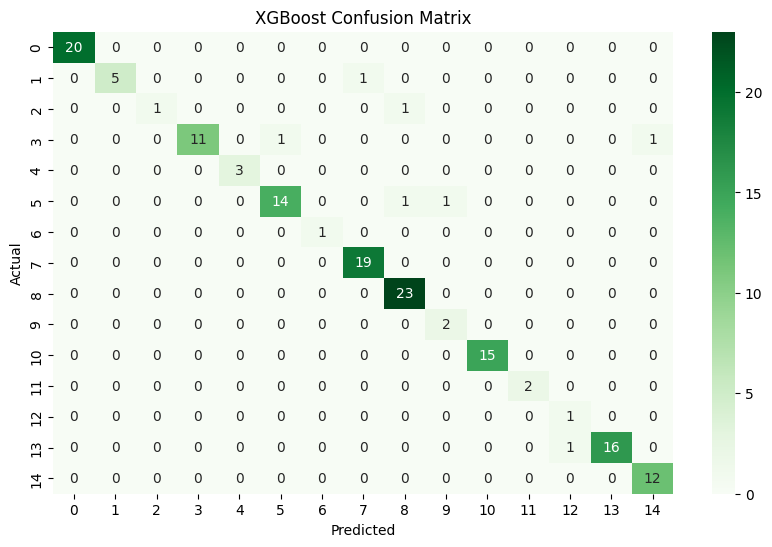
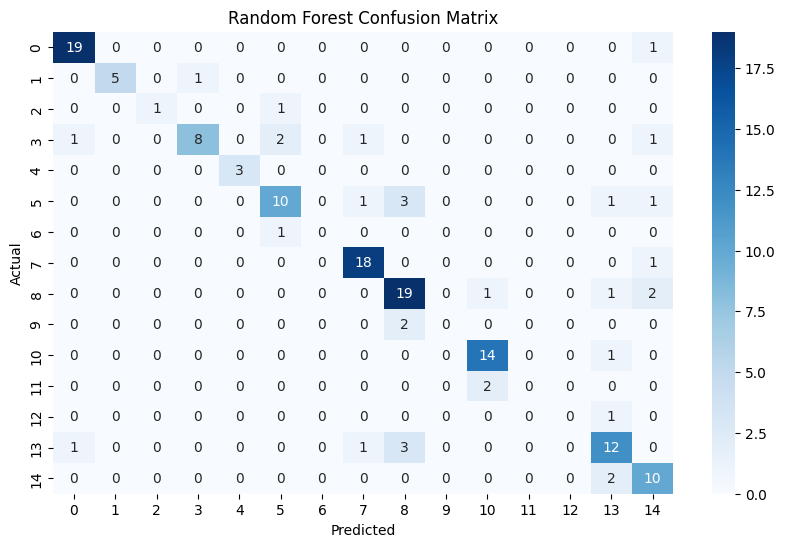
****

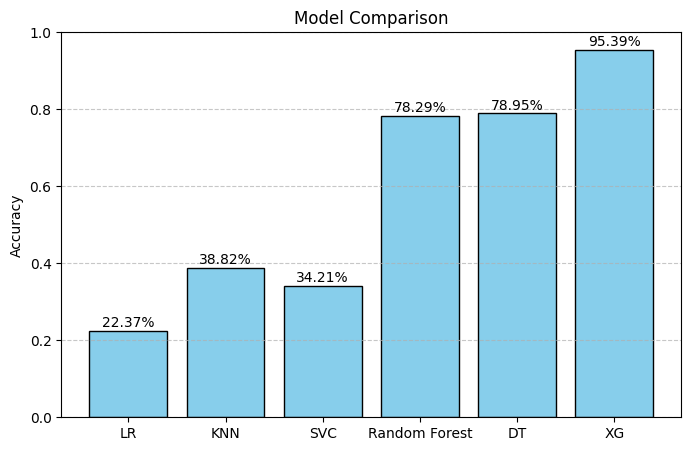
## 9. Model Building & Evaluation

* Multiple models were trained: **Logistic Regression, Decision Tree, SVC, KNN, Random Forest, XGBoost**.
* The Accuracy of each models are:





|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Model** | **Accuracy (Test)** | **CV Accuracy** | | --- | --- | --- | | **Logistic Regression** | 0.25 (25%) | 0.21 (21%) | | **Decision Tree** | 0.79 (79%) | 0.71 (71%) | | **KNN** | 0.36 (36%) | 0.37 (37%) | | **SVC** | 0.34 (34%) | 0.33 (33%) | | **Random Forest** | 0.75 (75%) | 0.67 (67%) | | **XGBoost** | 0.96 (96%) | 0.89 (89%) | | **XGBoost (Tuned)** | 0.93–0.95 (93–95%) | ~0.92–0.93 (CV) | |

* Random Forest and XGBoost performed best among models.
* Comparison chart confirmed XGBoost as the best performer after tuning.
* Overfitting was minimal, as train/test accuracies were close.
* The final tuned XGBoost was exported as ipl\_predictions\_best\_model.pkl.

## 10. Overall Insights from Analysis

* **Toss Impact** – Winning the toss increases win probability but is not decisive.
* **Team Performance** – Mumbai Indians and CSK are the most successful teams.
* **Venue Impact** – Certain cities favor chasing teams more.
* **EDA Insight** – Wins are usually by small margins (runs/wickets), Feature importance analysis also confirmed that toss decisions and team encodings strongly affect outcomes.
* **Modeling Insight** – Random Forest and XGBoost outperform simpler models, A comparative accuracy chart showed Random Forest and XGBoost as the best performers, while Logistic Regression, Decision Tree, KNN, and SVC had lower accuracy. Hyperparameter tuning further boosted XGBoost performance.
* **Overfitting Issue** – Overfitting check (train vs test accuracy) showed only a small gap for XGBoost, confirming the model generalizes well.

## 11. Conclusion

This project successfully demonstrated how IPL match data can be cleaned, analysed, and modelled for predictions. Key takeaways include:

* Mumbai Indians and CSK dominate IPL history in terms of wins.
* Toss decisions matter but are not the only deciding factor.
* Venues and city conditions influence strategies.
* Machine learning models like Random Forest and XGBoost provide the best predictive performance.

**Future Scopes:**

* Including player-level performance data (batting/bowling stats) could improve predictions.
* Real-time prediction models could be integrated into dashboards for analysts.

The final tuned XGBoost model was exported and can be used to predict IPL match winners, supporting decision-making for analysts and franchises.

**Business Impact**:  
Insights can help analysts, teams, and franchises understand winning strategies, optimize team decisions, and design predictive systems for future IPL seasons.

* Helps franchises refine toss & venue strategies.
* Provides analysts insights into win patterns.
* Can support broadcasters/fans with prediction-based engagement.