IPL Match Winner Prediction Using Machine Learning

1. Project Overview

This project aims to develop a machine learning model that predicts the winner of Indian Premier League (IPL) cricket matches using pre-match data. With the growing popularity and commercialization of T20 cricket, especially the IPL, accurate match predictions are valuable for enhancing strategic gameplay, fan engagement, and betting market insights.

The prediction system is trained on historical match data spanning 12 seasons (2008–2019) and includes various features such as team names, venues, toss outcomes, and match results. The primary focus is on data-driven modeling and extracting actionable insights that contribute to winning strategies.

2. Objective

The core objective of this project is to:

- Build a predictive model that forecasts the match winner before the game begins.
- Provide data-driven insights to assist in strategic planning.
- Explore and evaluate different machine learning algorithms for classification.
- Understand the relationship between match features and match outcomes.

By achieving these objectives, the model can assist analysts, coaches, and stakeholders in making informed decisions

3. Dataset Description

- **Timeframe**: 2008 to 2019 IPL seasons.
- **Total Matches**: Approximately 700+ matches.
- **Attributes**:
    - season: Year of the match

- city: Location of the match
- date: Match date
- team1, team2: Competing teams
- toss_winner: Team that won the toss
- toss_decision: Decision made (bat or field)
- result: Normal/Draw/Tie/No Result
- dl_applied: Duckworth-Lewis method applied (yes/no)
- winner: Match winner
- win_by_runs, win_by_wickets: Margin of win
- player_of_match: Best performer
- venue: Stadium
- umpire1, umpire2: Match officials

	id	Season	city	date	team1	team2	toss_winner	toss_decision	result	dl_applied	winner	win_by_runs	win_by_wickets	player_of_match
0	1	IPL-2017	Hyderabad	05-04-2017	Sunrisers Hyderabad	Royal Challengers Bangalore	Royal Challengers Bangalore	field	normal	0	Sunrisers Hyderabad	35	0	Yuvraj Singh
1	2	IPL-2017	Pune	06-04-2017	Mumbai Indians	Rising Pune Supergiant	Rising Pune Supergiant	field	normal	0	Rising Pune Supergiant	0	7	SPD Smith
2	3	IPL-2017	Rajkot	07-04-2017	Gujarat Lions	Kolkata Knight Riders	Kolkata Knight Riders	field	normal	0	Kolkata Knight Riders	0	10	CA Lynn
3	4	IPL-2017	Indore	08-04-2017	Rising Pune Supergiant	Kings XI Punjab	Kings XI Punjab	field	normal	0	Kings XI Punjab	0	6	GJ Maxwell
4	5	IPL-2017	Bangalore	08-04-2017	Royal Challengers Bangalore	Delhi Daredevils	Royal Challengers Bangalore	bat	normal	0	Royal Challengers Bangalore	15	0	KM Jadhav

4. Data Cleaning and Preprocessing

a. Data Cleaning

- **Dropped Columns**: umpire1, umpire2, and other metadata not impacting match outcomes were removed.

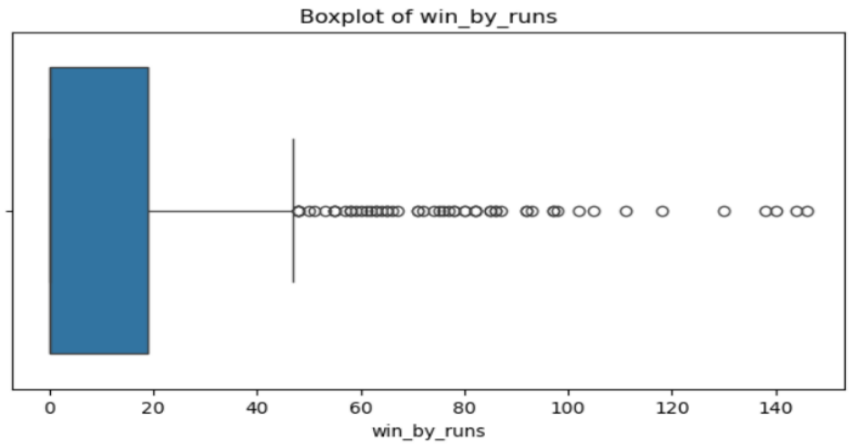
- **Missing Values**: Dropped null values in winner and city column since less than 5% of data was only missing.

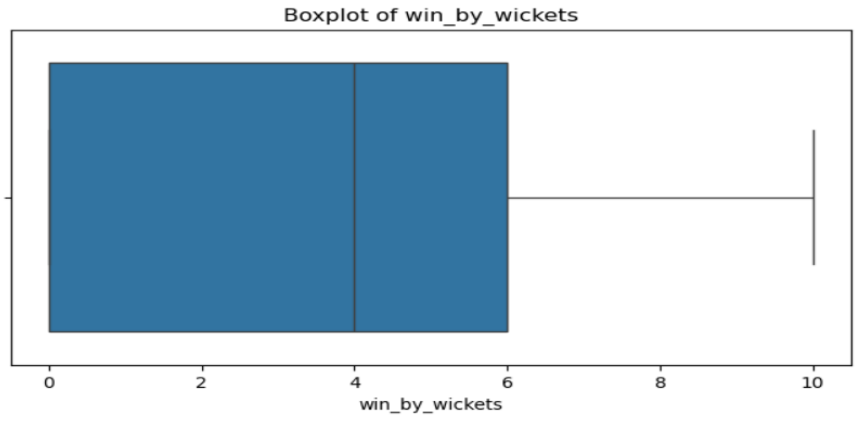
b. Standardization

- **Team Name Mapping**: Merged inconsistently named teams like 'Delhi Daredevils' and 'Delhi Capitals', etc.,
- **Categorical Normalization**: Standardized values across team1, team2, and venue.

c. Treating outliers

- Analyzed the **'wins by runs'** and **'wins by wickets'** columns for outliers.
- Found that a few matches had exceptionally high win margins in the 'wins by runs' column, which appeared as outliers.
- The 'wins by wickets' column showed a more balanced distribution, with no significant outliers.
- However, these outliers were not removed, as such results are possible in actual matches and not due to data entry errors.





d. Feature Encoding

- Applied **Label Encoding** to transform categorical variables into numerical labels, which are necessary for many machine learning models.

5. Exploratory Data Analysis (EDA)

a. Team Performance Insights

- **Mumbai Indians (MI)** and **Chennai Super Kings (CSK)** are the most consistent and successful franchises.

b. Toss Analysis

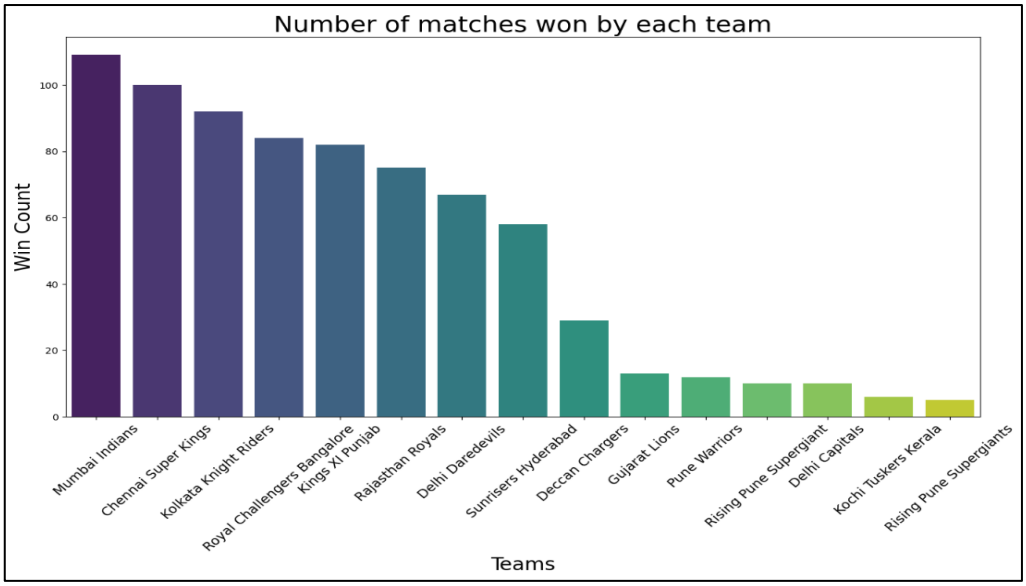
- A majority of toss-winning teams choose to bowl first.
- A positive correlation exists between toss wins and match wins.

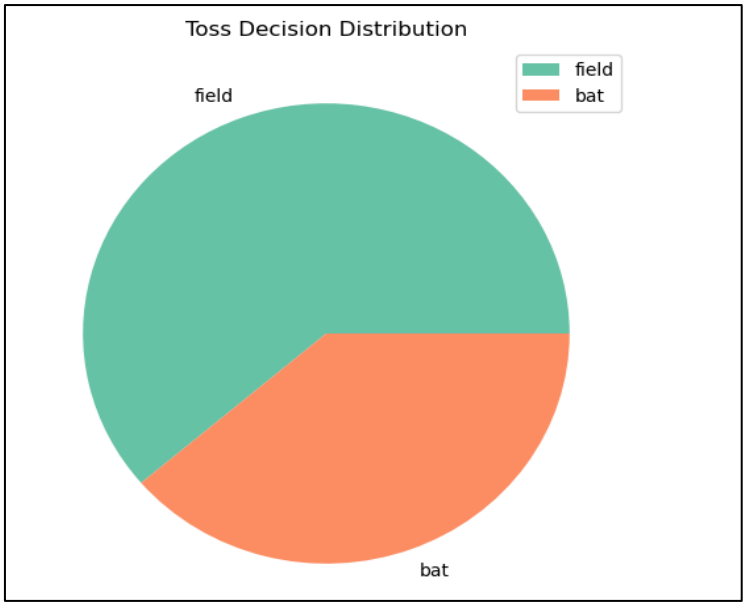
c. Venue Analysis

- Teams perform better in familiar stadiums indicating a strong **home advantage**.
- Some venues are high-scoring, which influences team strategies.

d. Distribution Checks

- Analyzed class imbalance in match outcomes.
- Verified data distributions for numerical features (win_by_runs, win_by_wickets).





6. Machine Learning Models Used

Models Implemented:

- **Logistic Regression**: Baseline linear model for binary classification.
- **K-Nearest Neighbors (KNN)**: Distance-based algorithm for prediction based on proximity.
- **Support Vector Machine (SVM)**: Utilized for non-linear decision boundary separation.
- **Decision Tree**: Tree-based model for interpretable rules.
- **Random Forest**: Bagging ensemble of multiple trees to reduce overfitting.
- **XGBoost**: Gradient boosting algorithm optimized for speed and accuracy.

Model Selection Process

- Initial experiments with simpler models to establish baseline performance.
- Used **RandomizedSearchCV** for efficient hyperparameter optimization.
- Evaluated models based on **multiple classification metrics**.

7. Model Training and Evaluation

Evaluation Metrics:

- **Accuracy**: Ratio of correct predictions to total predictions.
- **Precision**: Ability to correctly identify positive outcomes.
- **Recall**: Model's ability to detect all positive samples.
- **F1 Score**: Harmonic mean of precision and recall.
- **Confusion Matrix**: Breakdown of correct and incorrect classifications.

8. Model Performance Comparison

Model	Accuracy
Logistic Regression	28%
SVM	40%
KNN	38%
Decision Tree	86%
Random Forest	79%
XGBoost	**97%**

XGBoost emerged as the best performer due to its robustness, ability to handle overfitting, and effectiveness with categorical features.

	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.288591	0.269180	0.288591	0.267255
1	SVM	0.409396	0.392448	0.409396	0.393086
2	KNN	0.382550	0.406974	0.382550	0.381607
3	Decision Tree	0.865772	0.879224	0.865772	0.867514
4	Random Forest	0.798658	0.800724	0.798658	0.797143
5	XGBoost	0.973154	0.974071	0.973154	0.972527

9. Conclusion

This project demonstrates that machine learning, especially advanced ensemble models like XGBoost, can be successfully applied to predict sports outcomes such as IPL match winners. The analysis revealed that toss decisions, team matchups, and venues significantly influence outcomes.

The study also emphasizes the importance of data preprocessing and model tuning in achieving high prediction accuracy.

10. Future Work

a. Data Enrichment

- **Player Form**: Include player stats like recent performance and averages.
- **Weather and Pitch Data**: Include environmental factors impacting matches.
- **Injury Reports**: Monitor changes in team strength.

b. Real-time Prediction System

- Design and deploy a **web/mobile application** to provide live predictions before and during matches.
- Integrate **APIs for live data feeds**.

c. Advanced Modeling

- Implement **Ensemble Learning** techniques (Stacking, Voting).
- Explore **Neural Networks** and **LSTM** models for time-series match trends.

11. Tools and Technologies

- **Python**: Main programming language
- **Pandas/Numpy**: Data manipulation
- **Matplotlib/Seaborn**: Visualization
- **Scikit-learn**: ML models and preprocessing

Prepared by: SUBHIKSHA P

Batch: DADS November

Date: 16 May 2025