IPL Data Analysis & 2025 Winner Prediction Report

1. Introduction

Cricket analytics has evolved into an essential part of modern sports strategy. The Indian Premier League (IPL) stands as one of the world's most followed Twenty20 cricket leagues. In this analysis, we use historical data from 2008 to 2024 (covering both match summaries and ball-by-ball deliveries) to extract insights on team performance, player contributions, and seasonal trends. Ultimately, an ensemble model combining Random Forest, XGBoost, and a neural network (MLP) was developed to predict match outcomes and forecast the IPL 2025 champion.

---

2. Data Overview & Preprocessing

Datasets Used:

- **matches.csv (1,095 rows, 20 columns):** Contains match-level information such as season, teams, venue, toss details, results, and key statistics.
- **deliveries.csv (260,920 rows, 17 columns):** Provides ball-by-ball data including batting and bowling details, runs, extras, and wicket information.

Preprocessing Steps:

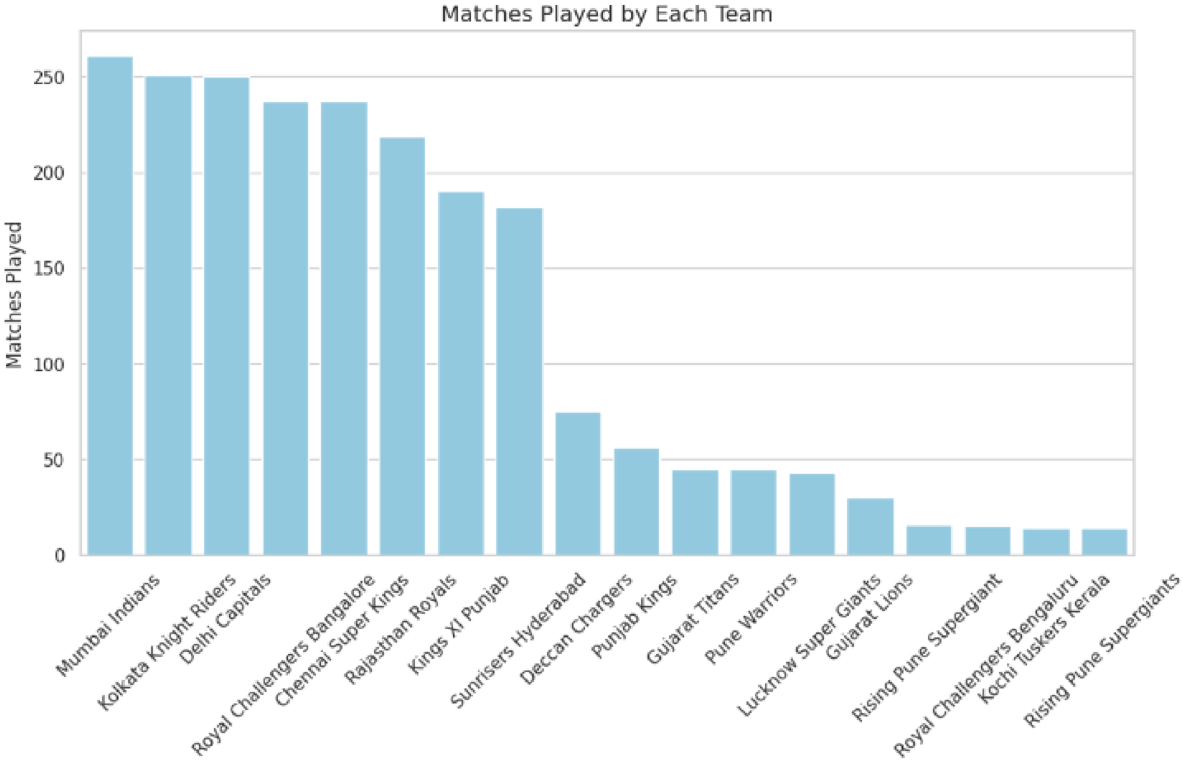
- Converted date columns to datetime format and sorted the matches chronologically.
- Handled missing values by dropping rows with missing critical data (e.g., winner, team1, team2, batsman, bowler).
- Standardized team names (e.g., replaced "Delhi Daredevils" with "Delhi Capitals").
- Renamed columns in the deliveries dataset for consistency (e.g., “batter” to “batsman”).

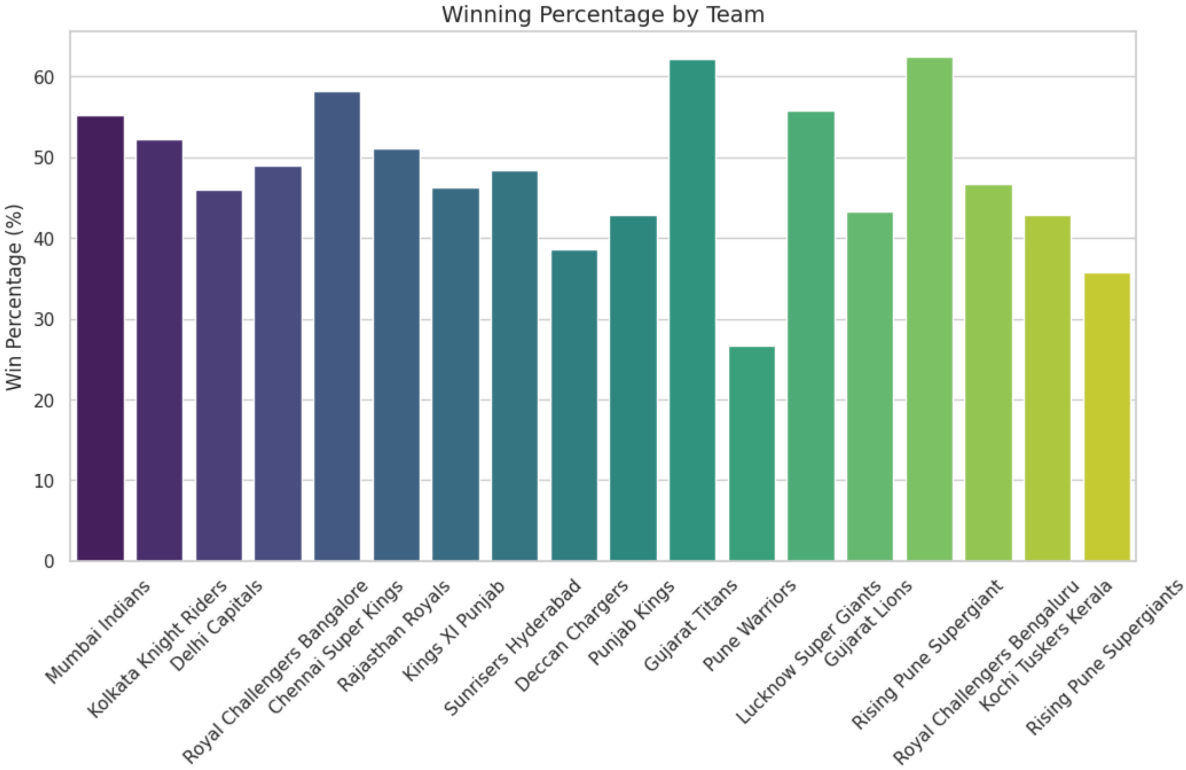
---

3. Exploratory Data Analysis (EDA)

3.1 Team Performance Analysis

- **Matches & Win Percentages:**
  The analysis computed the total matches played and winning percentages for each team. For example, Mumbai Indians, Chennai Super Kings, and Kolkata Knight Riders emerged as top performers based on wins.
- **Visualization:**
  Bar plots were generated to visualize both the total number of matches played and the win percentages across teams.





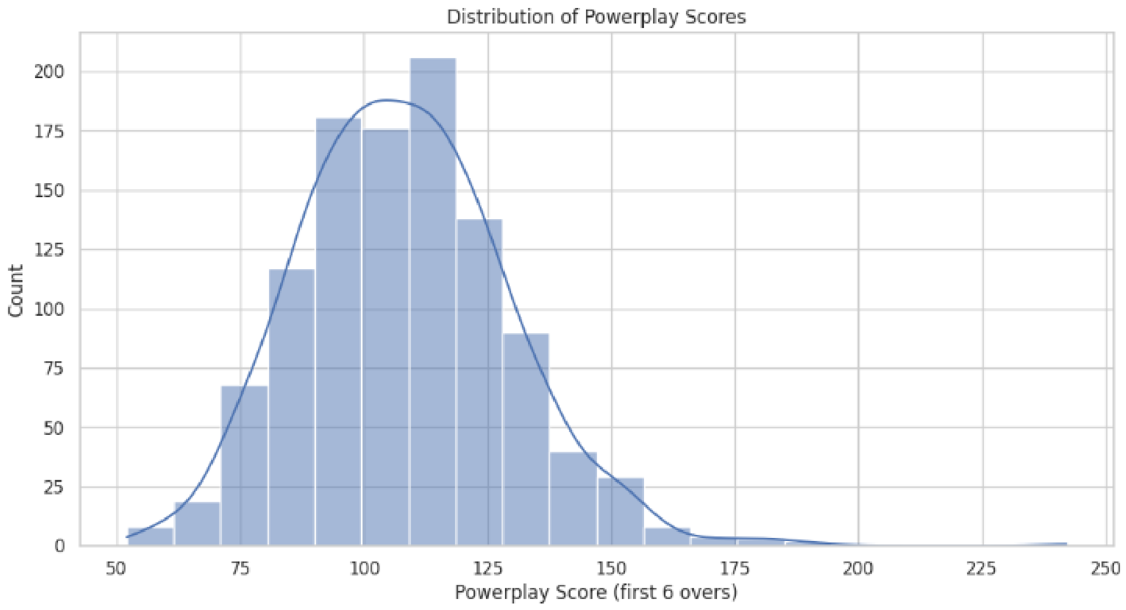
3.2 Batting & Bowling Analysis

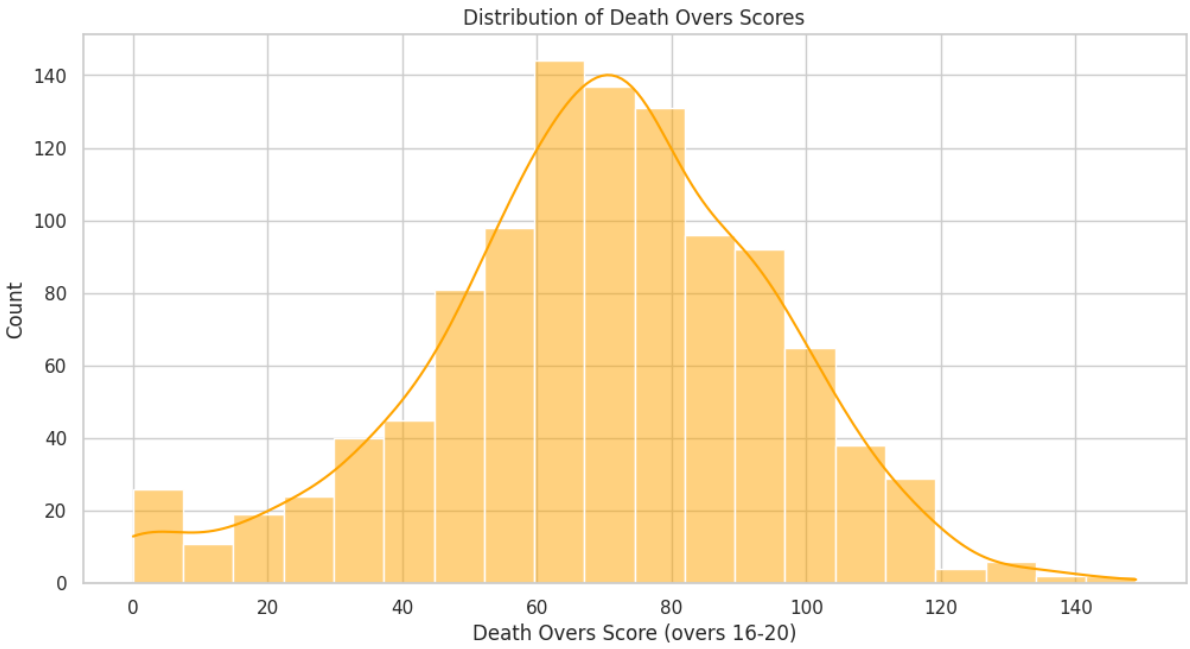
- **Run Rate & Economy Rate:**
  Using the ball-by-ball deliveries data, the run rate (total runs per over) for batting teams and the economy rate (runs conceded per over) for bowling teams were calculated.

- **Batting:** Run rates were derived by summing runs and dividing by the number of overs faced.
- **Bowling:** Economy rates were computed similarly from total runs conceded.

- **Boundaries and Scores:**
The analysis counted the total number of boundaries (4s and 6s) per team. Additionally, the highest and lowest match scores were identified from the aggregated deliveries.

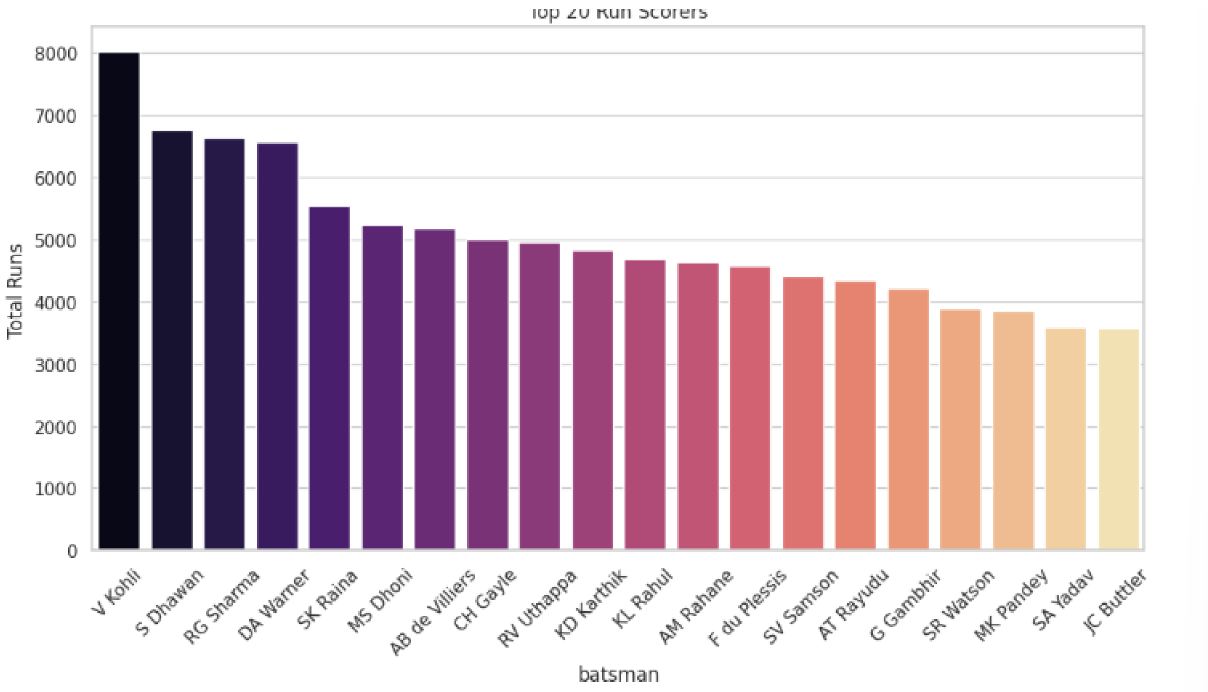
- **Powerplay and Death Overs:**
Functions were implemented to compute scores during the powerplay (first 6 overs) and death overs (overs 16–20). Histograms of these scores helped illustrate scoring distributions.



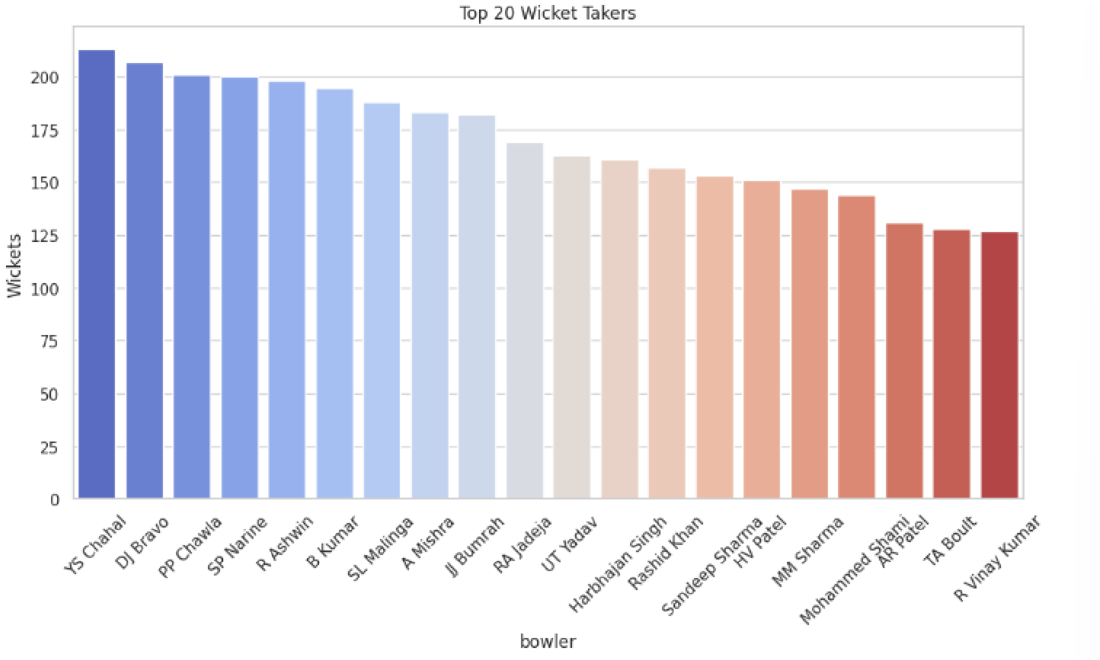


3.3 Player Performance Analysis

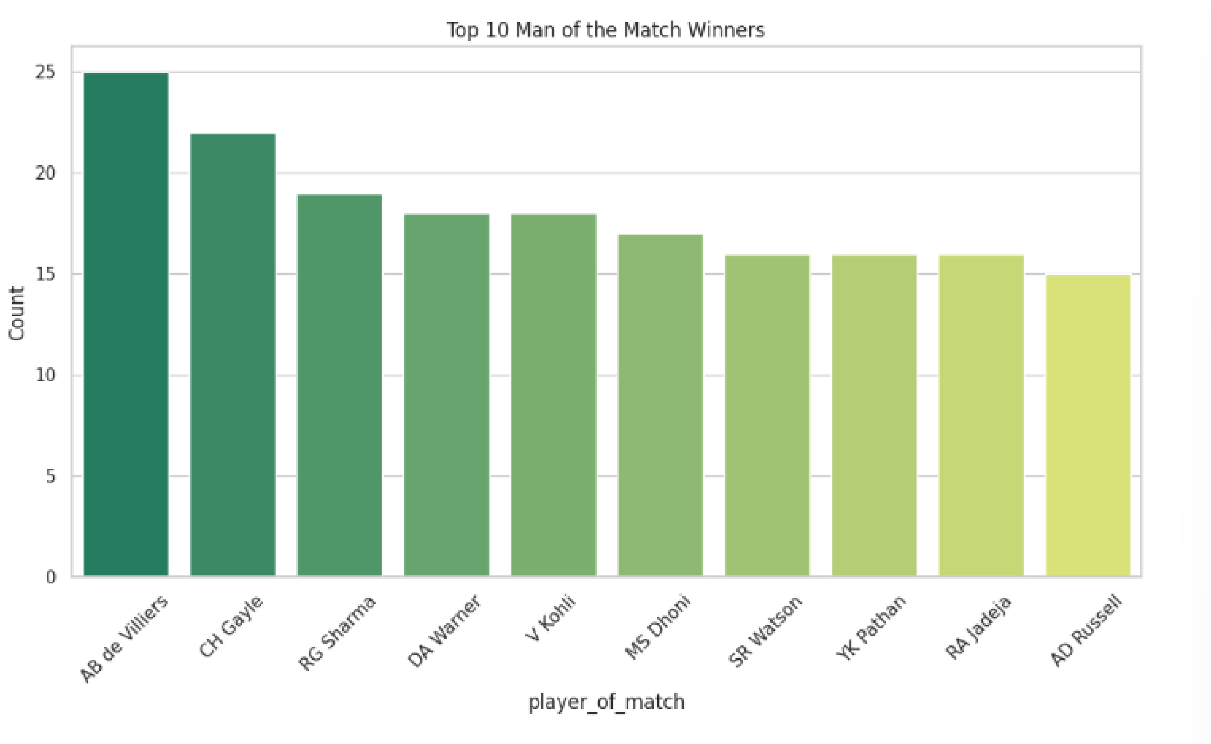
- **Top Run-Scorers & Batting Metrics:**
  The top 20 run-scorers were identified from the deliveries data. Batting averages and strike rates for these players were computed using runs, balls faced, and dismissals. A scatter plot visualized the trade-off between batting average and strike rate.



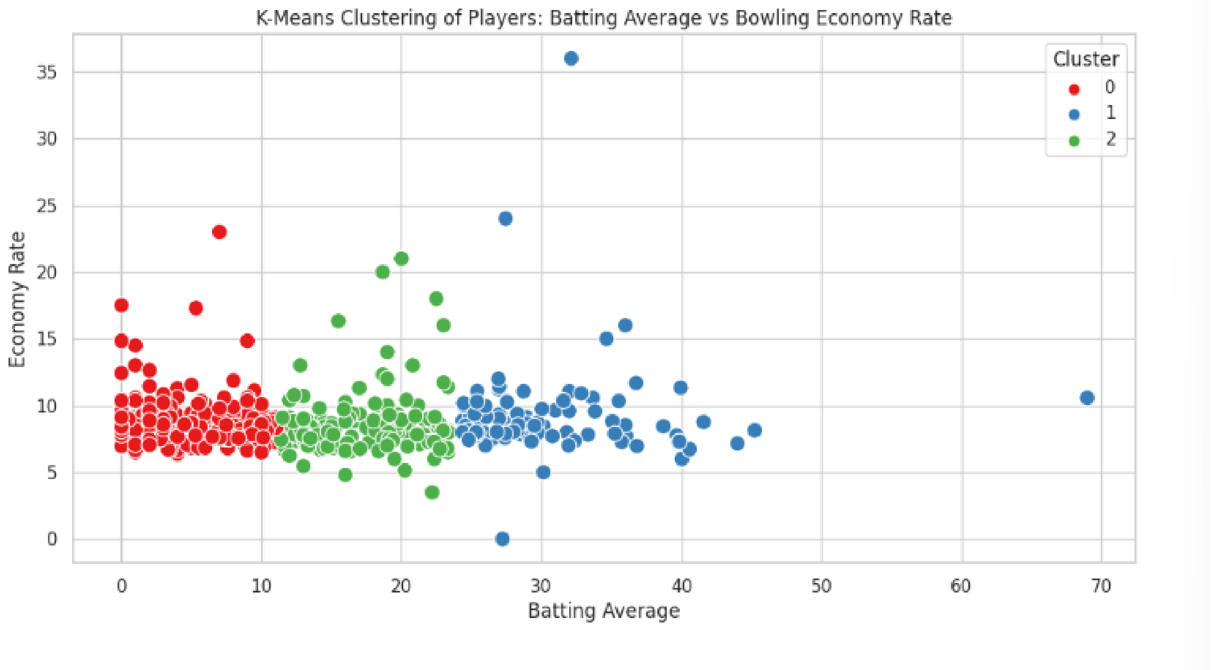
- **Wicket-Takers:**
  The top wicket-takers were determined by filtering deliveries with dismissals and counting wickets per bowler.
  - For instance, players like YS Chahal, DJ Bravo, and others featured prominently.



- **Individual Scores & Awards:**
  The top individual scores in a match were extracted and the count of "Player of the Match" awards was summarized.



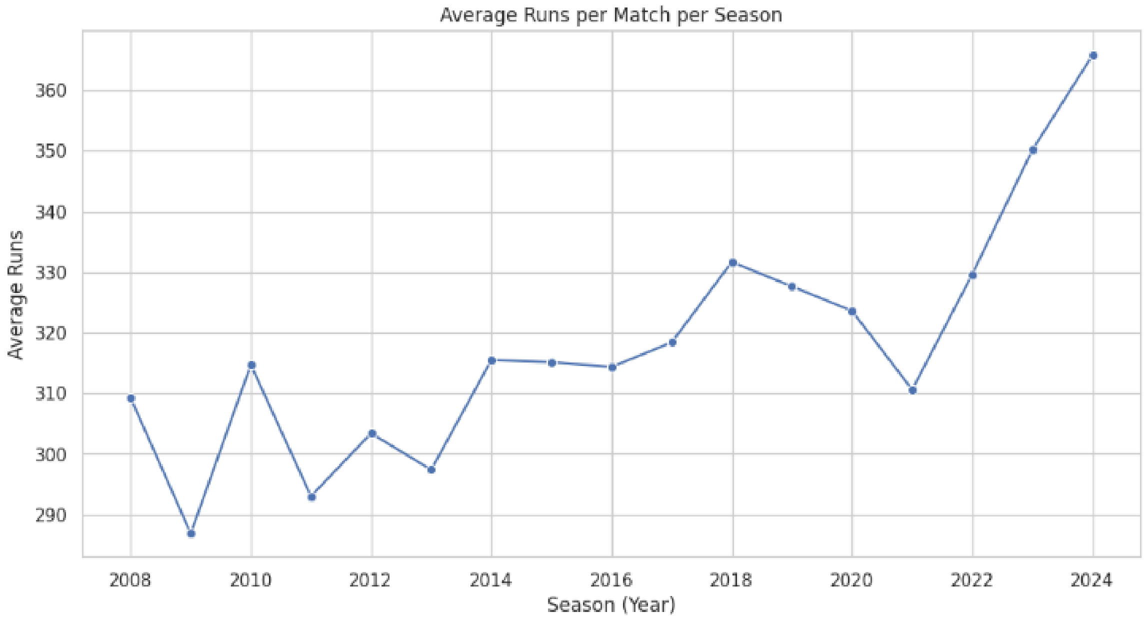
- **Clustering Analysis:**
  K-Means clustering was applied to a merged dataset (batting and bowling metrics) to identify player profiles (batsman, bowler, all-rounder) based on batting average and bowling economy rate.



- **Scoring Breakdown:**
  A detailed breakdown of scoring (singles, doubles, boundaries) was performed for additional insights.

3.4 Seasonal Analysis

- **Average Runs per Match:**
  By merging match and delivery data, the average total runs per match were calculated for each season, showing trends over time.



- **High-Scoring Matches:**
  The analysis identified 1,057 matches with target scores equal to or above 200 runs, indicating the evolving nature of high-scoring encounters.
- **Cap Analysis:**
  - **Orange Cap:** For each season, the highest run-scorer (Orange Cap holder) was identified. For example, in 2008, SE Marsh led the tally, while in 2024, V Kohli topped the list.
  - **Purple Cap:** Similarly, the top wicket-taker (Purple Cap holder) was determined for each season, with notable performances by players like Sohail Tanvir (2008) and HV Patel (2024).
- **Top Bowlers:**
  For each season, the top 10 bowlers were listed based on wicket counts, providing a clear view of seasonal bowling performances.

---

4. Feature Engineering & Winner Prediction Model

4.1 Feature Engineering

In addition to the basic match features (team1, team2, toss_winner, toss_decision, venue), additional historical performance metrics were engineered:

- **Overall Win Percentages:**
  For each team, historical win percentages were computed and mapped to the matches data.
- **Toss Win Advantage:**
  A feature was derived representing the win percentage of the toss-winning team (if the toss winner was one of the playing teams).

4.2 Model Development

An ensemble model was built by combining three classifiers:

- **Random Forest Classifier**
- **XGBoost Classifier**
- **MLP Classifier (Neural Network)**

These base models were integrated using a VotingClassifier. The model was trained on historical match data, where categorical features were label-encoded and numerical features included the engineered win percentages and toss advantage.

4.3 Model Performance

The ensemble model was evaluated using a classification report with the following key metrics:

- **Overall Accuracy:** ~50%
- **Detailed metrics:**
  Precision, recall, and F1-scores were reported per class, indicating the variability of model performance across different match outcomes.

Note: The moderate accuracy reflects the complexity and variability inherent in cricket match outcomes, which can be influenced by numerous unpredictable factors.

---

5. Prediction for IPL 2025

5.1 Enhanced Simulation

Rather than using a simple dummy schedule, the simulation for the 2025 season incorporated the enriched feature set:

- **Simulated Schedule:**
  A dummy schedule for 60 matches was created. For each match, the basic features along with the historical win percentages and toss win advantage were included.
- **Prediction Aggregation:**
  The ensemble model predicted the outcome of each match in the simulated schedule. By aggregating match-level predictions, the team with the most predicted wins was forecast as the champion.

5.2 Prediction Outcome

- **Sample Predictions:**
  For individual matches, predicted winners ranged across various teams.
- **Championship Forecast:**
  The aggregated predictions led to **Chennai Super Kings** being identified as the predicted champion for the IPL 2025 season.

---

6. Discussion & Conclusion

Findings

- **Team & Player Insights:**
  Comprehensive EDA revealed trends in team performance, seasonal scoring, and key player contributions. Both batting and bowling metrics showed notable evolution over the years.
- **Model Insights:**
  Incorporating historical performance metrics (win percentages and toss advantage) enriched the prediction model. Although the ensemble model achieved an overall accuracy of approximately 49.08%, it provided valuable insights into potential match outcomes.
- **2025 Prediction:**
  The forecast based on simulated enriched features predicted **Chennai Super Kings** as the champion. This outcome is in line with historical trends where consistent performance has been a hallmark of successful IPL franchises.

Strengths

- **Comprehensive EDA:**
  The analysis covered multiple facets—from team performance and player metrics to seasonal trends—providing robust insights.
- **Feature Engineering:**
  Inclusion of domain-specific metrics (win percentages and toss advantage) helped contextualize match outcomes.
- **Ensemble Approach:**
  Combining multiple classifiers helped leverage the strengths of different models for more robust predictions.

Limitations

- **Model Accuracy:**
  With an overall accuracy of ~49%, the model indicates room for improvement. Real-world match outcomes are influenced by dynamic factors (e.g., player form, injuries, weather) that are challenging to capture fully.
- **Simulation Assumptions:**
  The simulated schedule for 2025 assumes historical performance trends continue, which may not account for evolving team dynamics and external conditions.

- **Data Granularity:**
  While the analysis is detailed, further granular features (e.g., pitch conditions, player fitness data) could potentially enhance predictive performance.

Conclusion

This report demonstrates how advanced data analytics and ensemble machine learning methods can be applied to sports data to generate actionable insights and forecasts. By integrating comprehensive EDA with domain-specific feature engineering, the analysis provides a data-driven prediction for the IPL 2025 champion. Despite inherent limitations, the methodology serves as a strong foundation for further refinement and real-world applications in cricket analytics.

References

- IPL official match data and statistics.
- Domain-specific literature on cricket analytics and machine learning applications in sports.