IPL Match Analysis

Submitted by:
Prabhdeep Singh (2024010078)
Shamandeep Singh (2024010095)

MCA 2nd Year

Submitted to:
Dr. Anjula Mehto
Assistant Professor

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department
Thapar Institute of Engineering and Technology,
Patiala
November 2025

TABLE OF CONTENTS

Introduction or Project Overview

One of the most competitive cricket leagues is the Indian Premier League (IPL), which gives us a lot of structured match data every season. By looking at this data, we can figure out what makes a team strong, how players usually do, how venues change over time, and what makes a match-winning squad. This project analyses extensive IPL match data to look at past patterns and develop models that can forecast the outcomes of matches and the scores of innings. The analysis includes everything from preprocessing and exploratory data analysis to complex visualisations and machine learning. The whole process is meant to enable cricket experts, fans, and strategists get a better understanding of the IPL ecosystem.

1. Data Cleaning & Preprocessing

- Loaded and inspected dataset shape, columns, and data types
- Identified missing values and generated a missing-data heatmap
- Removed duplicate entries
- Converted date columns into proper datetime format
- Standardized categorical values and prepared columns for analysis
- Handled inconsistent entries across seasons and teams
- Encoded categorical features for ML models
- Scaled numerical features to improve model performance

2. Exploratory Data Analysis (EDA)

Visually analyzed team performance, toss influence, match results, season trends, and venue behavior.

Key Insights Identified:

- **Most successful teams** by wins across seasons
- **Teams with highest toss-win counts**
- **Season-wise performance patterns**
- **Venues with highest match counts and scoring patterns**
- **Home vs. away performance differences**
- **Run distribution patterns:** powerplay, middle overs, death overs
- **Player-impact metrics** such as top run scorers, wicket takers, and match winners

3. Deep Cricket Insights Extracted

Season-by-Season Analysis

- Trends in team dominance
- High-scoring vs. low-scoring seasons
- Emerging patterns in match outcomes

- **Venue Analysis**
- Batting-friendly vs. bowling-friendly stadiums
- Toss decisions based on venue behavior
- Average first-innings and second-innings scores

Phases of Play Analysis

- Powerplay scoring vs. wickets lost
- Middle-overs stability
- Death-overs acceleration patterns

•Head-to-Head (H2H) Rivalries
•Win/loss patterns between key rival teams
•Dominant matchups and long-term trends

4. Machine Learning Models

Built predictive models to understand outcome drivers and future predictions.

Model 1: Match Winner Prediction
•Input features: toss result, venue, team composition, historical stats, etc.
•Model: Random Forest / (or model used in your notebook)
•Outputs: predicted winning team
•Evaluated using accuracy, confusion matrix, and classification metrics

•Model 2: First Innings Score Prediction
•Predicts runs based on venue, team strength, overs completed, wickets lost
•Visualizations include feature importance and error metrics

5. Statistical Analysis

•Identified significant predictors using correlation matrices
•Chi-Square tests for categorical influence
•ANOVA tests for numerical variables
•Revealed features with direct match-winning relationships

6. Final Deliverables & Insights

•A fully cleaned and structured IPL dataset
•Comprehensive dashboards with team, player, and venue insights
•High-quality visualizations for trends and patterns
•Predictive ML models for match winner and score estimation
•Actionable cricket insights useful for:

    •Analysts
    •Fantasy league players
    •Coaches and strategists
    •Fans interested in deep match breakdowns

Problem Statement

The Indian Premier League (IPL) generates vast amounts of match, player, and team data every season. However, this data is often underutilized because it is scattered, complex, and difficult to interpret without systematic analysis. Teams, analysts, and enthusiasts require deeper insights into match outcomes, scoring patterns, venue behavior, and performance trends to make informed decisions.

The main challenge is to **identify the factors that influence match results, scoring trends, and team performance**, and to build models that can use historical data to make reliable predictions. This includes predicting match winners, estimating first-innings scores, understanding team strengths, analyzing rivalries, and uncovering patterns across different venues and seasons.

Additionally, organizations or analysts need a **centralized dashboard** where these insights can be visualized interactively, allowing real-time exploration of IPL data without manual analysis.

This project aims to solve these challenges by:

- Cleaning, preprocessing, and structuring the IPL dataset
- Performing deep exploratory analysis to extract meaningful insights
- Building machine learning models that can predict match outcomes and scores
- Identifying statistically significant performance factors
- Creating a dashboard that provides intuitive visualizations and predictions

By addressing these needs, the project provides a complete analytical framework that transforms raw IPL data into actionable insights and predictive intelligence.

Overview of the Dataset used

The dataset used in this project is a detailed ball-by-ball record of **Indian Premier League (IPL)** matches, covering multiple seasons starting from **2008**. It provides granular information for every ball bowled, along with match-level metadata, team performance statistics, player actions, and scoring breakdowns.
This rich dataset forms the foundation for deep cricket analytics, predictive modeling, and exploratory insights across seasons, venues, teams, and match situations.

Dataset Size

•**Total Rows: 131,970+** entries (each row = one ball delivered)
•**Total Columns: 50+** features capturing match, team, and player details
•High-volume, structured sports data suitable for both EDA and machine learning.

Key Feature Categories

The dataset contains multiple types of variables. Below is a structured breakdown:

1. Match Information

Data providing match-level context:

•match_id
•date
•match_type (T20)
•event_name (Indian Premier League)
•event_match_no
•stage (Playoffs, league stage, etc.)
•match_number

2. Team Details

Information about teams involved in each delivery:

•batting_team
•bowling_team
•team1, team2 (where available)
•winner (for match outcome prediction)

3. Innings & Over Details

Variables describing game progression:

•innings
•over
•ball
•balls_per_over
•overs (cumulative overs)

4. Player-Level Information

Tracks individual batting and bowling performance:

•batter
•bowler
•non_striker
new_batter
•next_batter

5. Delivery Outcomes

These columns capture the exact result of each ball:

- runs_off_bat
- extras
- wides, noballs, byes, legbyes, penalty
- total_runs (sum of bat + extras)
- bowler_wicket
- striker_out
- method (dismissal method)

6. Team-Level Accumulators

Cumulative match progression metrics:

- team_runs
- team_balls
- team_wicket
- batting_partners
- batter_runs, batter_balls

These features help compute run rates, powerplay trends, scoring phases, and partnership strengths.

7. Contextual Metadata

Extra columns useful for feature engineering:

- match_id
- event_match_no
- stage
- method (Duckworth–Lewis, normal result, etc.)

Project Workflow

The project follows a complete end-to-end data science pipeline, starting from data ingestion and cleaning, moving through exploratory analysis and modeling, and ending with visualizations and predictive insights. The workflow ensures that raw IPL match data is transformed into meaningful cricket intelligence and actionable predictions.

1. **Data Collection & Understanding**
   - Loaded the **IPL ball-by-ball dataset** containing 131k+ records and 50+ columns.
   - Inspected dataset shape, columns, data types, and initial structure.
   - Identified numerical, categorical, date-time, and match-level metadata.
   - Formed initial understanding of key cricket variables: overs, innings, runs, wickets, teams, and players.

2. **Data Cleaning & Preprocessing**
   - Removed duplicates and irrelevant fields.
   - Handled missing values in numerical and categorical columns.
   - Converted date and match identifiers into proper formats.
   - Fixed inconsistent team or player names (if present).
   - Standardized categorical values for uniformity.
   - Created derived fields such as:
     - cumulative innings score
     - phase of play (powerplay/middle/death overs)
     - partnership runs
     - match progression metrics

3. **Exploratory Data Analysis (EDA)**

   Performed deep exploratory visual analysis to understand cricket insights.

   Key EDA Steps

   - Visualized **team performance** across seasons.
   - Analyzed **toss impact** on match outcomes.
   - Explored **venue behavior** (batting-friendly vs bowling-friendly).
   - Studied **run distributions** across overs and innings.
   - Generated **season-wise scoring patterns**.
   - Compared **head-to-head (H2H) rivalries** between major teams.
   - Identified:
     - high-scoring grounds
     - dominant teams
     - phase-wise scoring (Powerplay, Middle Overs, Death Overs)

4. **Feature Engineering**

   Created meaningful ML-ready features from raw data:
   - Total runs per over
   - Wickets fallen up to each ball
   - Run rate progression
   - Batting phase classification
   - Derived match context features (venue, toss decision, etc.)
   - One-hot encoding of teams & venues

•Scaling numerical features

These engineered features significantly improved model accuracy.

5. Machine Learning Modeling

Model 1: Match Winner Prediction

•Input features: toss result, teams, venue, scoring trends, etc.
•Algorithms used: Random Forest / other models tested.
•Evaluated using accuracy, confusion matrix, F1-score.
•Feature importance extracted to identify key outcome drivers.

Model 2: First Innings Score Prediction

•Predicts score based on overs, wickets, venue, team strength.
•Used regression approaches.
•Measured performance using MAE, RMSE, and R^2.
•Visualized error distribution and model fit.

6. Statistical Analysis

•Performed **Correlation Analysis** to identify impactful features.
•Conducted **Chi-Square Tests** on categorical variables.
•Performed **ANOVA** to check numerical variable significance.
•Validated which features strongly influence match outcomes.

7. Data Visualization & Insights

Created intuitive and cricket-focused plots:
•Team win distributions
•Toss vs match outcome comparison
•Venue-wise performance plots
•Season scoring trends
•Run rate progression charts
•H2H rivalry visualization
•Over-by-over scoring analysis

Results

The project turned the raw, ball-by-ball IPL dataset into a full analytical and predictive system that could find significant cricket insights. After cleaning and structuring over 131,000 deliveries over 50+ variables, the data became consistent, enhanced, and suitable for exploratory analysis and machine learning. The preprocessing stage made sure that missing numbers, inconsistent formats, and noisy entries were all fixed. New features like cumulative scores, run rates, phase classifications, and wickets progression made the analysis more detailed.

We found a lot of patterns in team performance, season trends, venue behaviour, and scoring phases by doing a lot of exploratory data analysis. It was easy to see which teams had been dominant for a long time, and the way scores changed from season to season showed how IPL cricket had changed over the years, with run rates continuously rising. Venue study showed that some venues were better for batting and others were better for bowling. Some stadiums always had high first-innings totals. The breakdown of overs by phase showed that powerplay overs are still the most unpredictable, middle overs are still the rebuilding phase, and death overs are still the time when scoring speeds up the greatest. Rivalry analysis showed that some encounters have always been one-sided, which supports long-standing competitive patterns in the league.

The machine learning part made it possible to make very good predictions. The match winner prediction model, which included techniques like Random Forest, was quite accurate and easy to understand. The most important things that came out were the venue, the toss decisions, the team strength indications, and the early-over performance. The first-innings score prediction model also worked well, properly predicting scores based on wickets, overs, run-rate growth, and the attributes of the venue. MAE and RMSE, two error metrics, showed that the model was stable and useful for making predictions.

Statistical analysis confirmed these observations even more. Correlation matrices, Chi-Square tests, and ANOVA tests validated that both categorical variables (e.g., venue and team) and numerical variables (e.g., wickets and run rate) substantially influence match outcomes and innings totals. These statistical results helped back up the patterns found during EDA and the factors that the machine learning models pointed out.

Conclusion

This project successfully demonstrates how IPL cricket data can be transformed from raw ball-by-ball records into meaningful insights and predictive intelligence. Through comprehensive cleaning, preprocessing, and exploratory analysis, the study uncovers key patterns related to team performance, scoring behavior across different match phases, venue influence, and long-term seasonal trends. These insights illustrate how strategic factors such as powerplay performance, death-over acceleration, and venue conditions consistently shape match outcomes, providing a deeper understanding of the game’s dynamics.

The machine learning models developed—focused on match winner prediction and first-innings score estimation—further enhance the analytical value of the project, delivering reliable and interpretable results. Supported by statistical validation and integrated into an interactive Streamlit dashboard, the system provides users with a practical, data-driven platform for real-time analysis. Overall, the project successfully combines analytics, visualization, and predictive modeling into a unified framework that benefits analysts, strategists, fantasy players, and cricket enthusiasts seeking informed decision-making based on IPL data.

GitHub Link

- https://github.com/PrabhdeepJassal/MCA_308/blob/main/project-labeval.ipynb
- https://mca308prabhshamanlabeval.streamlit.app