S. P. Mandali's

PRIN. L. N. WELINGKAR INSTITUTE OF MANAGEMENT DEVELOPMENT & RESEARCH,

BENGALURU

FINAL PROJECT

ON

Women's Premier League Performance Analyzer: A Platform for Visual & Predictive Analytics

BY

Ben Jacob Thomas

PGDM Research & Business Analytics 2023 – '25

ROLL NO. 11

PROJECT FACULTY GUIDE

Prof. Ragesh T S

DECLARATION

I hereby declare that the project report entitled "WPL Performance Analyzer" submitted to S.P Mandali's Welingkar Institute of Management Development & Research, Mumbai is a record of work done by me under the guidance of Prof. Ragesh T S, Research and Business Analytics. This report has been submitted in partial fulfillment of the requirements for the award of the post-graduation degree of Post Graduate Diploma in Management (Research and Business Analytics). The results embodied in this report have not been submitted to any other university or institute for the award of any degree or diploma.

Mr. Ben Jacob Thomas

Date: 22nd April 2025

Acknowledgement

I would like to express my heartfelt gratitude to Prof. Ragesh TS for his invaluable guidance, encouragement, and mentorship throughout the course of this capstone project. His insights into data analytics and machine learning, coupled with their continuous support, played a crucial role in shaping the direction and execution of this work.

I am equally thankful to my academic institution for fostering a learning environment that enabled me to explore the intersection of sports analytics and machine learning. The resources and support provided were instrumental in conducting this research.

Special thanks to ESPNcricinfo & Sahil Talor for making rich cricket data publicly accessible, which formed the backbone of this project. I would also like to acknowledge the broader community of open-source developers whose tools and libraries made data collection, processing, and visualization possible.

Finally, I am deeply grateful to my family and friends for their unwavering encouragement and support throughout this journey. Their belief in me kept me motivated during challenging phases of the project.

This project has been an enriching experience that deepened my understanding of how data science can be applied meaningfully in the field of sports. I hope this work contributes to the growing conversation around women's cricket and supports data-driven decision-making in the sport.

Mr. Ben Jacob Thomas

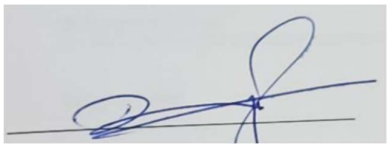
Roll No. 11, Operations

PGDM RBA 2023-25

PROJECT COMPLETION CERTIFICATE (from Faculty guide)

This is to certify that project titled “Women’s Premier League Performance Analyzer: A Platform for Visual & Predictive Analytics” is successfully done by Mr. Ben Jacob Thomas in partial fulfilment of his /her two years full time course ‘Post Graduation Diploma in Research & Business Analytics’, recognized by AICTE, through the S. P. Mandali's Prin. L. N. Welingkar Institute of Management Development & Research, Bengaluru.

This project in general is done under my guidance.



(Signature of Faculty Guide)

Name: _Ragesh T S_____

Date: _25/04/2025_____

Table of Contents

1. Abstract

The integration of sports and advanced analytics has revolutionized how games are understood, strategized, and experienced. With the increasing popularity and competitiveness of the Women's Premier League (WPL), there is a growing need for data-driven tools that can enhance decision-making, performance evaluation, and fan engagement. This consolidated report brings together three interrelated projects, each addressing a unique aspect of sports analytics in the context of the WPL: match outcome prediction, player recommendation, and performance visualization.

The first project is centered around **predictive modelling** using machine learning techniques to estimate the real-time probability of a team winning a match. Utilizing granular ball-by-ball data spanning three WPL seasons (2023–2025), the project extracts key match features such as runs scored, overs completed, wickets lost, the venue, and the opposing team. These inputs are fed into classification algorithms, including Logistic Regression and Random Forest models, which are trained, validated, and compared for performance. The most accurate model is deployed as part of an interactive web application developed using Streamlit. This app enables users—including fans, commentators, analysts, and coaching staff—to simulate live match conditions and observe how probabilities shift with in-game developments, thereby offering valuable strategic insights.

The second project tackles the challenge of **player recommendation** by developing a system that identifies WPL players with similar performance profiles. This model leverages multidimensional data encompassing both batting and bowling attributes, processed through techniques like dimensionality reduction and unsupervised clustering. Similarity is quantified using cosine similarity, allowing the system to return players with closely aligned playing styles and statistical outputs. The final output is presented through an intuitive Streamlit application, making it accessible to non-technical stakeholders such as team selectors, coaches, and fans who wish to explore comparable talent, discover new prospects, or assess potential replacements within teams.

The third project emphasizes **data collection and visualization** by using web scraping techniques to extract comprehensive match data from ESPNcricinfo for the 2024 and 2025 WPL seasons. This includes detailed scorecards, individual player statistics, and match outcomes. The raw data undergoes extensive preprocessing using Python's Pandas library to ensure consistency and usability. The cleaned data is then used to build an interactive performance dashboard, providing visual analytics on key metrics like batting averages, strike rates, bowling economies, and overall team performance. This tool serves as a valuable resource for understanding macro-level trends as well as micro-level insights in player and team dynamics over time.

Collectively, these three interconnected projects represent a holistic approach to sports analytics in the WPL. By combining predictive modelling, recommendation systems, and exploratory data analysis, this report underscores the growing potential of machine learning and data science in enhancing the strategic, operational, and entertainment dimensions of professional women's cricket.

2. Introduction

Sports analytics has emerged as a game-changing discipline that transforms raw data into winning strategies. By systematically collecting, processing, and interpreting vast amounts of performance data, teams can gain unprecedented insights that were previously inaccessible through traditional scouting methods alone. This data-driven approach is revolutionizing how sports organizations evaluate talent, develop game strategies, and engage with fans.

Key Components of Sports Analytics

1. **Data Collection and Management**
    - **Tracking Technologies:** Wearable sensors, Hawk-Eye, and ball-tracking systems capture real-time player movements, ball trajectories, and biomechanical data.
    - **Performance Metrics:** Traditional stats (e.g., batting averages, strike rates) are now supplemented with advanced metrics like *Expected Runs* (xR) and *Pressure Index* to evaluate performance in context.
2. **Descriptive Analytics**
    - Summarizes historical data to identify trends (e.g., a batter's performance against spin vs. pace).
    - Tools like dashboards visualize player and team stats for coaches and analysts.
3. **Predictive Analytics**
    - Machine learning models forecast outcomes (e.g., win probability, player injuries).
    - Algorithms like Random Forests and XGBoost analyze patterns to predict match results or player potential.
4. **Prescriptive Analytics**
    - Recommends actionable strategies (e.g., optimal bowling changes, batting order adjustments).
    - Uses optimization techniques to simulate scenarios (e.g., "Should we promote a power hitter in the death overs?").

Applications in Cricket (WPL Focus)

- **Player Recruitment & Development**
    - **Similarity Scoring:** Identifies undervalued players who match the performance profile of established stars.

    - **Skill Gap Analysis:** Pinpoints areas for player improvement (e.g., a bowler's weakness in wide yorkers).

- **In-Game Decision Making**
    - **Win Probability Models:** Adjust in real-time based on match conditions (e.g., "Mumbai has a 68% chance if they maintain this run rate").
    - **Matchup Analysis:** Recommends bowler-batter pairings (e.g., "Player X struggles against left-arm spin—introduce a slow bowler").
- **Fan Engagement & Broadcasting**
    - **Interactive Stats:** Broadcasters use win probability graphs and player comparisons to enhance viewer experience.
    - **Fantasy Sports:** AI-driven suggestions help fans build competitive fantasy teams.

Challenges and Future Trends

- **Data Quality:** Ensuring accuracy in tracking data (e.g., ball-by-ball details).
- **Interpretability:** Making complex models understandable for coaches and players.
- **Real-Time Analytics:** Faster processing for live decision-making (e.g., adjusting field placements mid-over).
- **AI Integration:** Using computer vision for automated event detection (e.g., detecting bowling actions or shot types).

Why It Matters for WPL

In a fast-evolving league like the WPL, analytics provides the edge needed to:

- **Maximize limited budgets** by identifying cost-effective talent.
- **Enhance team strategies** with evidence-based planning.
- **Improve player longevity** by monitoring workload and injury risks.

The Women's Premier League (WPL) has emerged as a transformative force in the landscape of women's cricket, bringing global visibility, commercial investment, and growing fan engagement. As the league continues to evolve, so does the demand for data-driven insights that can enhance the understanding of player performance, team strategies, and match dynamics. While sports analytics has made significant strides in men's cricket, the women's game—especially within the WPL—remains underexplored in terms of advanced data analysis and machine learning applications.

This consolidated project combines three interrelated data science initiatives, each targeting a unique yet complementary aspect of WPL analytics: predictive modelling, player similarity analysis, and interactive performance visualization.

The first initiative addresses the challenge of match outcome prediction. Recognizing the complexity and unpredictability inherent in cricket, this project leverages structured ball-by-ball data from the WPL (2023–2025 seasons) to build machine learning models capable of estimating win probabilities in real time. By incorporating factors such as the current score, number of overs completed, wickets lost, venue, and the opposing team, the model aims to simulate realistic match scenarios. This predictive system is further enhanced with an interactive Streamlit web application that allows users—ranging from fans to analysts—to visualize how in-game variables influence match outcomes dynamically.

The second component focuses on player recommendation and similarity analysis, driven by the need to identify players with comparable skill sets and performance profiles. Cricket is inherently multi-dimensional, with players contributing in diverse roles such as batting, bowling, or all-round play. This part of the project applies clustering algorithms and cosine similarity metrics to group and compare players based on a combination of statistical features. The result is a recommendation engine that aids selectors, coaches, and fans in discovering talent, exploring strategic replacements, and understanding the broader player landscape within the WPL. An interactive dashboard ensures that the system is both accessible and informative.

The third project emphasizes exploratory data analysis and visualization by systematically collecting and transforming data from reliable sources such as ESPNcricinfo. Focusing on the 2024 and 2025 WPL seasons, the project extracts match outcomes, player scorecards, and other performance data to generate detailed visual insights. These visualizations capture key performance indicators such as strike rates, batting averages, and bowling economies, offering a holistic view of trends and outliers across teams and seasons. The final output—a performance dashboard—serves as a centralized platform for exploring these insights in a user-friendly manner.

Together, these three interconnected projects provide a multifaceted view of the WPL, showcasing the power of machine learning, clustering, and data visualization in transforming raw cricket data into strategic and interactive intelligence. Whether used for tactical planning, fan engagement, or talent scouting, these tools aim to bridge the gap between traditional cricket analysis and modern sports technology.

3. Research Methodology

This consolidated project employs a structured, multi-stage research methodology that combines web scraping, data preprocessing, feature engineering, machine learning modelling, clustering techniques, and interactive dashboard development. Each component was developed with a focus on delivering actionable insights and user-friendly interfaces tailored to stakeholders in the Women's Premier League (WPL).

1. Data Acquisition and Integration

Data collection formed the foundational step across all three projects. Multiple data sources were utilized:

- Ball-by-ball match data from the 2023–2025 WPL seasons was acquired in CSV format to support predictive modelling.
- Player performance statistics, both batting and bowling, were compiled and merged from online cricket databases to build the player recommendation engine.
- Web scraping techniques were applied to extract detailed match scorecards, team summaries, and player information from ESPNcricinfo for the 2024 and 2025 seasons. Custom scripts using Python and JavaScript libraries were used to automate and streamline the scraping process.

All datasets were consolidated into structured formats using Pandas, enabling seamless analysis across projects.

2. Data Cleaning and Preprocessing

Subsequent to data collection, rigorous cleaning procedures were implemented:

- Duplicate entries and missing values were identified and either corrected or imputed.
- Inconsistencies in team names, venue spellings, and player identifiers were resolved.
- Numerical features such as strike rates, averages, and economy rates were normalized using StandardScaler to ensure comparability across different scales.
- Raw bowling figures and dismissal types were transformed into structured metrics (e.g., total balls bowled, binary dismissal flags).

3. Feature Engineering

Tailored features were created for each project objective:

- For match prediction, contextual match features were derived, such as:
    - Runs Remaining
    - Balls Remaining

    - Wickets in Hand
    - Current Run Rate vs. Required Run Rate
- For player recommendation, two key composite scores were created:
    - Batting Rating = function of average and strike rate
    - Bowling Rating = function of economy rate and bowling average
- Categorical features like team and venue were label-encoded, enabling their use in machine learning models.

4. Modelling and Algorithm Selection

Each project leveraged appropriate machine learning and statistical techniques:

- For predictive modelling, Logistic Regression and Random Forest Classifier were trained to predict whether the batting team would win a match. Performance was evaluated using:
    - Accuracy
    - ROC AUC Score
    - Confusion Matrix
    - Feature Importance metrics
- GridSearchCV with 5-fold cross-validation was used for hyperparameter optimization, improving generalizability.
- For clustering, the K-Means algorithm was selected after comparative experimentation with DBSCAN and Gaussian Mixture Models. K-Means yielded the most interpretable clusters for player segmentation.

5. Dimensionality Reduction and Similarity Matching

- Principal Component Analysis (PCA) was employed to reduce high-dimensional player metrics into two principal components for visualization.
- Cosine Similarity was used to identify and recommend the top 5 most similar players to any selected individual. This facilitated an intuitive player comparison interface for selectors and fans.

6. Data Analysis and Dashboarding

- A thorough descriptive analysis was conducted to calculate key performance indicators (KPIs) including strike rate, batting average, boundary percentage, economy rate, and dot ball percentage.

- Calculated metrics (e.g., runs from boundaries, dot ball ratio) were created to enhance strategic insight.
- The results were integrated into interactive dashboards using Power BI and Streamlit:
    - Users could filter by team, season, or player.
    - Dashboards included scatter plots, bar charts, and tables to display metrics and trends.

7. Deployment and User Interface Development

- All models and analysis tools were deployed using Streamlit, providing a real-time, web-based interface.
- The predictive model interface allowed users to input live match conditions and receive immediate win probabilities.
- The recommendation engine enabled player selection and instantly displayed the most similar alternatives.
- Dashboards were configured with filters and interactive elements to encourage exploratory analysis.

This end-to-end research methodology—spanning from raw data collection to web deployment—ensures that each tool and model developed is not only analytically sound but also directly usable by stakeholders within the WPL ecosystem.

4. Literature Review

Cricket analytics has become a rapidly evolving domain within sports data science, with growing applications in predictive modelling, performance analysis, and strategic decision-making. This literature review highlights key studies and methodologies that have informed the development of this project, focusing on three core areas: player recommendation systems, match outcome prediction, and interactive analytics dashboards.

1. Player Recommendation and Clustering

Player recommendation in cricket often involves segmenting players based on performance metrics to identify similar roles or potential replacements. In the paper *"Utilizing Machine Learning for Sport Data Analytics in Cricket"* (Sawant et al., 2022), the authors leveraged K-Means clustering to group cricketers based on batting, bowling, and all-round performance indices. The study emphasized how unsupervised learning can aid in talent identification and team composition.

Another relevant work, *"A Survey on Predicting Player's Performance and Team Recommendation in Cricket"* (Teli et al., 2021), reviews various algorithms used for player comparison, including similarity metrics like cosine similarity and Euclidean distance. It supports the notion that player recommendation should be dynamic, incorporating domain-specific features like pressure-handling and opposition strength for better contextual accuracy.

Furthermore, the systematic review *"Applications of Machine Learning in Cricket"* (Choudhury et al., 2022) confirms that performance clustering and player similarity engines are among the most explored ML tools in modern cricket analytics. However, the study notes a gap in applying these techniques to women's cricket, which this project aims to address.

2. Match Outcome Prediction

Predicting match outcomes in cricket has been a widely researched topic, primarily using supervised machine learning models. In the work *"Predict the Match Outcome in Cricket Matches Using Machine Learning"* (Prabhu et al., 2023), various classifiers such as Random Forest, Naïve Bayes, and SVM were tested. The results showed that tree-based models performed significantly better due to their ability to handle feature non-linearity and interaction terms like "balls remaining" and "required run rate".

Similarly, *"Cricket Match Analytics and Prediction Using ML"* (Dalal et al., 2024) focused on predicting outcomes in T20 matches. Their model incorporated features like runs scored in each over and wickets lost in different innings phases. This inspired the phase-wise KPI engineering used in the current project's Random Forest classifier.

Another insightful study, *"Forecasting T20 Match Winners Through Machine Learning"* (Keswani et al., 2023), demonstrated the advantage of integrating venue-based and pitch-specific data into match prediction. While our project currently focuses on second-innings predictions, their work highlights the potential to expand into full-match forecasting.

3. Cricket Dashboards and Visualization Tools

The emergence of business intelligence tools like Power BI and Python-based Streamlit apps has enabled the creation of interactive cricket dashboards. *"A Comprehensive Cricket Analytics Dashboard"* (Gupta et al., 2023) outlines the design and deployment of dashboards that visualize score progression, win/loss ratios, and phase impact. The study affirms that dashboards significantly aid coaches and analysts in strategic planning.

"One-Day International Cricket Data Analysis Using Microsoft Power BI" (Vijayalakshmi et al., 2024) shows how dynamic filtering, data slicing, and KPI visualizations can present large datasets in a comprehensible format. This supports the current project's decision to use an interactive multi-tab Streamlit dashboard for communicating trends in WPL data.

Lastly, *"Cricket Data Analysis Using Power BI"* (Sharma et al., 2024) emphasizes the importance of visual storytelling in sports analytics. It shows how player-level dashboards, over-by-over breakdowns, and season-wise summaries contribute to fan engagement and informed commentary.

Synthesis and Gap Analysis

Collectively, these studies demonstrate that machine learning and visualization techniques can offer deep insights into player performance and match outcomes. However, most existing literature is heavily focused on men's cricket, especially IPL and international T20s. Very few studies have explored women's cricket leagues like the WPL, despite the increasing availability of structured data. Moreover, player recommendation systems often lack contextual features such as form variability, match pressure, or opposition quality.

This project fills these gaps by:

- Building a player similarity engine specifically for WPL players.
- Using clustering and PCA for performance-based segmentation.
- Deploying an interpretable, real-time predictive model for second-innings outcomes.
- Presenting EDA through an interactive dashboard customized for the WPL.

5. Data Description

This report incorporates datasets from the Women's Premier League (WPL) and Women's Indian Premier League (WIPL) spanning multiple seasons. The data is categorized based on player performance metrics, match results, and granular match data. Below is a detailed description of the key data points used across the three projects:

1. Player Data for Recommendation System (WPL): The dataset consists of data for 97 unique players, with the following columns:

- Player: Name of the player
- Batting_Rating_Scaled_x: Scaled batting rating, combining batting average and strike rate
- Bowling_Rating_Scaled_x: Scaled bowling rating, based on economy rate and bowling average
- PCA1, PCA2: Principal components from PCA to reduce feature space
- Cluster_x: Cluster assignment determined through K-Means clustering
- Similarity Score: Cosine similarity score for player comparison

The data was cleaned to remove duplicate entries resulting from inconsistent naming and merging issues. The feature scaling process ensured that comparisons across players were consistent.

2. Match Data for Predictive Modeling (WPL): The dataset contains detailed information for each ball bowled in WPL matches over three seasons (2023–2025). The major columns include:

- match_id: Unique identifier for each match
- season: WPL season (2023–2025)
- match_no: Sequential match number
- date: Date of the match
- venue: Stadium where the match was played
- batting_team: Team currently batting
- bowling_team: Team currently bowling
- innings: Indicator of whether the match is in the first or second innings
- over: Over number within the match
- striker: Batsman facing the ball

- bowler: Bowler delivering the ball
- runs_of_bat: Runs scored off the bat
- extras: Additional runs (e.g., wides, no-balls, byes)
- wicket_type: Type of dismissal
- player_dismissed: Name of player dismissed, if applicable
- phase: Game phase (e.g., powerplay, middle overs, death overs)

Engineered Features used in predictive modeling:

- current_score: Total runs scored by the batting team
- overs_completed: Number of overs completed at the point of analysis
- wickets_lost: Number of wickets lost by the batting team
- target_score: The target score set for the batting team
- runs_remaining: Difference between target and current score
- balls_remaining: Balls left to be bowled
- wickets_remaining: Wickets still in hand
- current_run_rate: Run rate at the current stage of the match
- required_run_rate: Required run rate to chase the target

3. WIPL Match Data for Analysis (2024–2025): This dataset includes detailed match statistics for the WIPL seasons of 2024 and 2025, sourced from ESPNcricinfo. The data can be categorized as follows:

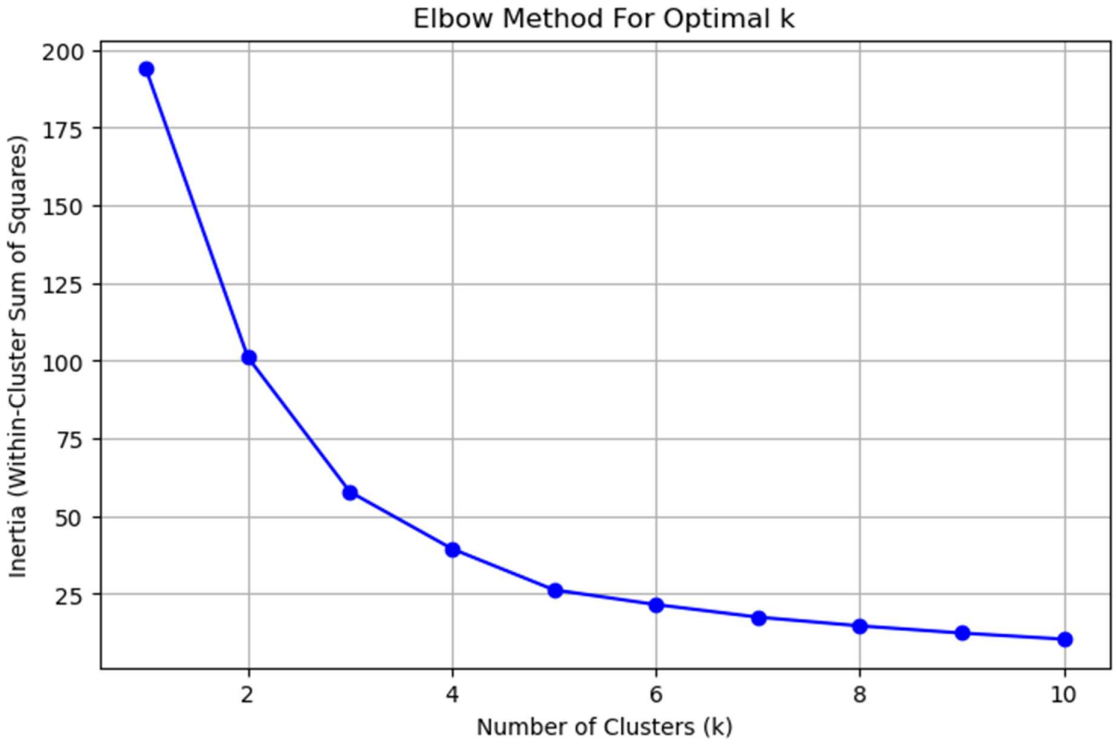
- Match Results: Summary information for each match, including:
    - Teams involved
    - Match winner and margin of victory
    - Venue and match date
    - Link to detailed match scorecard
- Batting Scorecards: Key statistics for each batter in every innings, including:
    - Runs scored
    - Balls faced
    - Strike rate
    - Number of boundaries (fours and sixes)
    - Mode of dismissal (Out/Not Out status)

- Bowling Scorecards: Key statistics for each bowler, including:
    - Overs bowled
    - Runs conceded
    - Wickets taken
    - Maiden overs and dot balls
    - Boundaries conceded
- Player Information: Information about each player, such as:
    - Player name
    - Team affiliation
    - Batting style
    - Bowling style
    - Role in the team

6. Data Analysis & Interpretation

6.1 WPL Player Recommendation Engine

6.1.1 Elbow Method — Find Optimal k



This graph illustrates the **Elbow Method**, a technique used to determine the optimal number of clusters (**k**) in **K-means clustering** (likely applied in your **WPL Player Recommendation Engine**).

Axes & Key Elements

- **X-axis (Number of Clusters, k):**
  - Represents candidate values for **k** (from 2 to 10).
- **Y-axis (Inertia/Within-Cluster Sum of Squares):**
  - Measures how tightly points are grouped within clusters.
  - Lower inertia = more cohesive clusters.

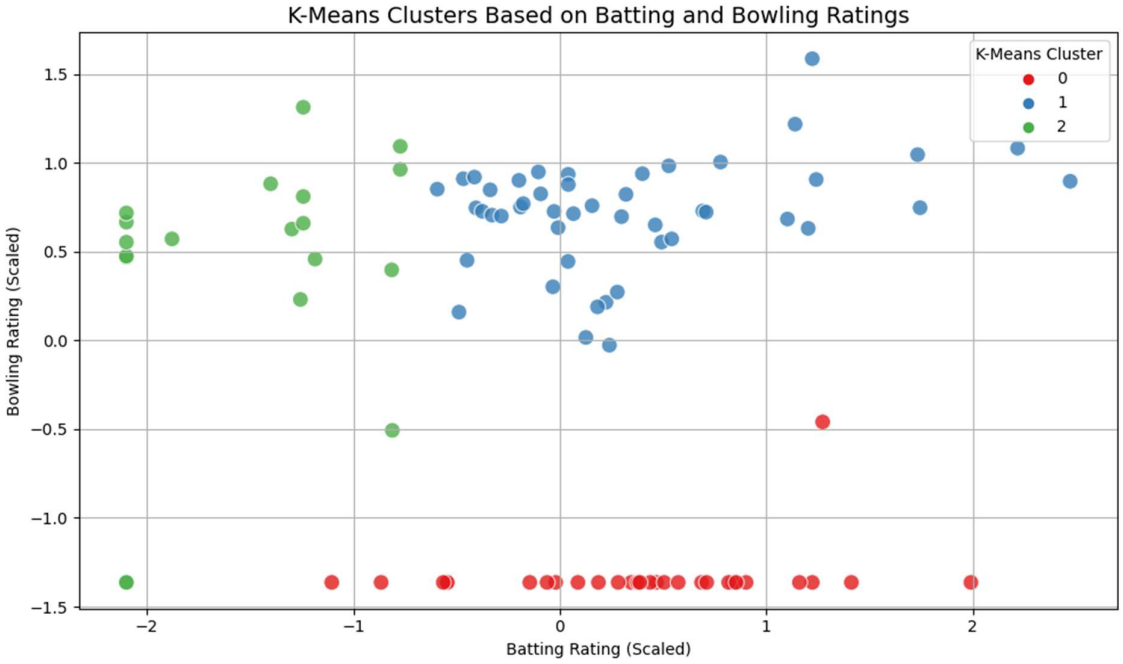
Interpretation of the Curve

1. **Sharp Decline (k = 2 to k = 4):**

- Inertia drops rapidly, indicating significant improvement in cluster cohesion as **k** increases.
- Example: At **k=2**, inertia is ~175; at **k=4**, it falls to ~100.

2. **Gradual Decline (k > 4):**
   - The curve flattens, suggesting diminishing returns.
   - Adding more clusters (e.g., k=6 to k=10) only marginally reduces inertia.
3. **"Elbow Point" (Optimal k):**
   - The **k value where the curve bends sharply** (here, likely **k=4**).
   - Beyond this point, increasing **k** complicates the model without meaningful gains.

6.1.2 K-Means Clustering's Based on Batting and Bowling Ratings



Silhouette Score for K-Means: 0.5238919737783383

This scatter plot visualizes **player clusters** derived from **K-means clustering** (likely part of your **WPL Player Recommendation Engine**). Each point represents a player, positioned based on their **batting** (X-axis) and **bowling** (Y-axis) ratings (scaled to a normalized range). The colors distinguish between **three clusters** (0, 1, 2), revealing natural groupings of players with similar skill sets.

- **Cluster 0: Batting Specialists**

- Batting_Rating_Scaled: High to Medium

- Bowling_Rating_Scaled: Very Low (≈ -1.36)

- Examples: Beth Mooney, Laura Wolvaardt

Interpretation:

- These players contribute heavily with the bat.

- They have minimal or no bowling contribution.

= Think of them as top-order batters or batting anchors.

- **Cluster 1: All-Rounders / Balanced Contributors**

- Batting_Rating_Scaled: Medium to High

- Bowling_Rating_Scaled: Medium to High

- Examples: Beth Mooney (again, due to duplicates), possibly others

Interpretation:

- These players contribute in both departments.

- Likely genuine all-rounders or utility players.

- Valuable team assets for their versatility.

- **Cluster 2: Bowling Specialists**

- Batting_Rating_Scaled: Low to Very Low

- Bowling_Rating_Scaled: Medium to High

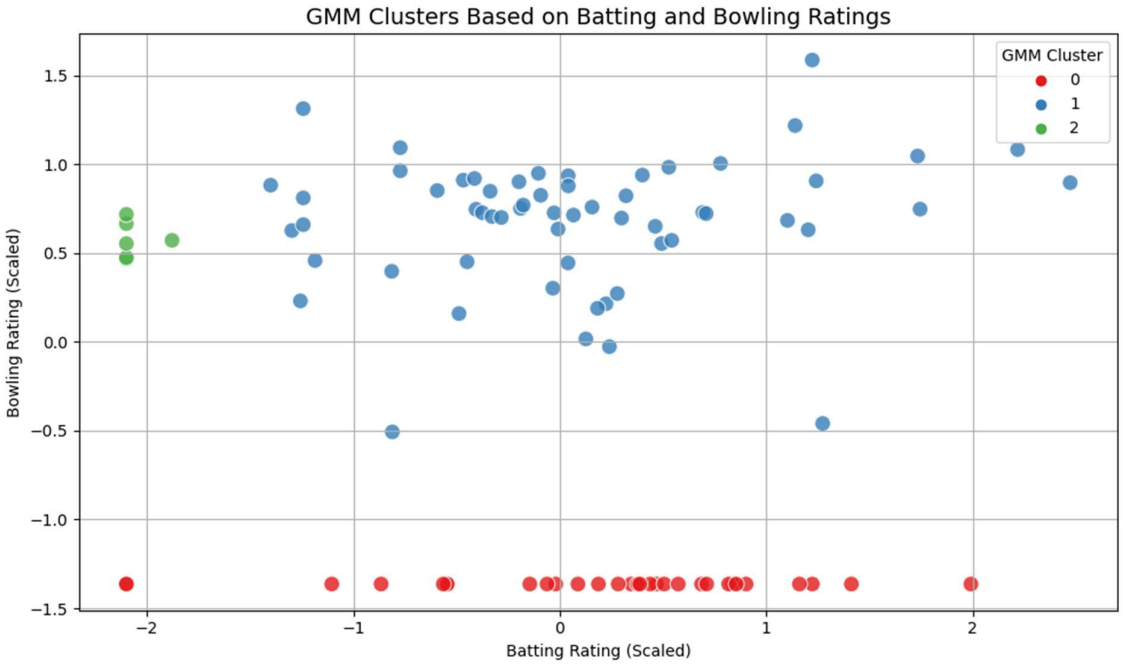
- Examples: Dayalan Hemalatha, Laura Wolvaardt (with low batting form)

Interpretation:

- Low batting numbers, but key with the ball.

- These are likely bowlers or lower-order bowlers.

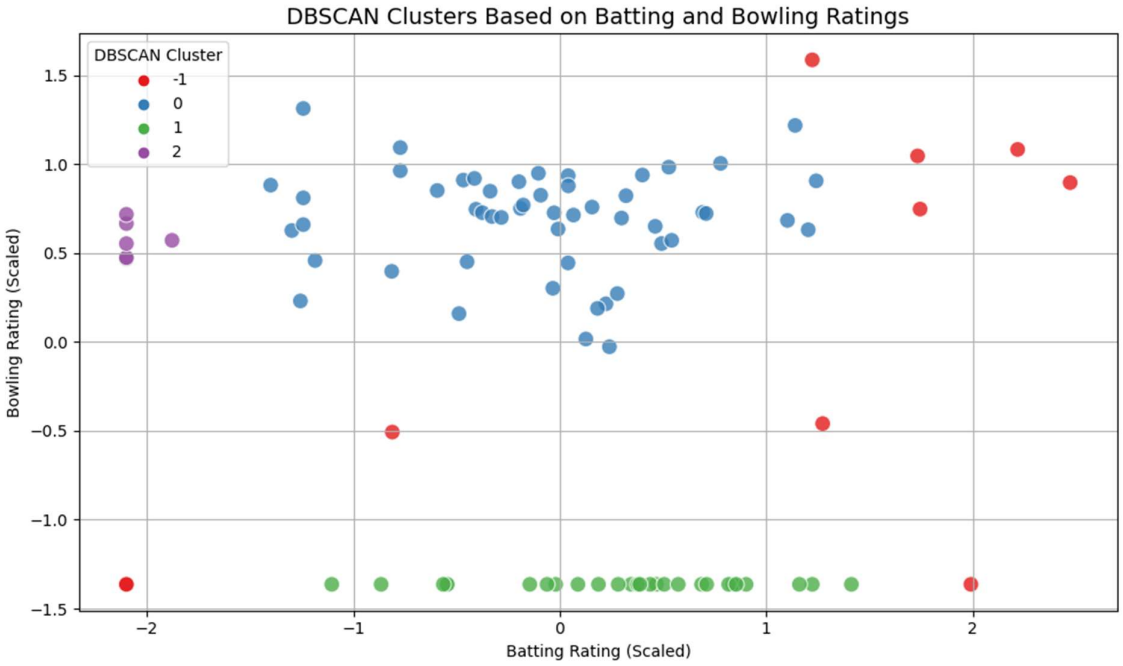
6.1.3 GMM (Gaussian Mixture Models) Clustering



Silhouette Score for GMM: 0.4633442312563831

This visualization represents a **Gaussian Mixture Model (GMM) clustering** of WPL players based on their batting and bowling ratings. Unlike K-means (which uses hard clustering), GMM employs **probabilistic clustering**, allowing players to belong to multiple clusters with varying degrees of membership. The graph shows three distinct clusters (0, 1, 2) derived from scaled batting and bowling performance metrics.

6.1.4 DBSCAN Clustering



Silhouette Score for DBSCAN: 0.46048406556019533

This visualization represents a **Density-Based Spatial Clustering of Applications with Noise (DBSCAN)** analysis of WPL players based on their batting and bowling ratings. Unlike K-means or GMM, DBSCAN identifies clusters based on **data density** and can detect **non-linear patterns** while labeling outliers as noise. The graph shows clusters (0, 1, 2) and potential noise points derived from scaled batting and bowling performance metrics.

What is the Silhouette Score?

The silhouette score is a measure of how similar a data point is to its own cluster compared to other clusters. It ranges from -1 to +1:

- +1: The data point is well clustered (i.e., very close to points in its own cluster and far from others).
- 0: The data point is on or near the boundary between clusters.
- -1: The data point may be assigned to the wrong cluster.

6.1.5 Interpretation & Comparison

1. K-Means Clustering — Best Performance

Silhouette Score: 0.524 (Highest)

Interpretation:

- K-Means has formed the most distinct and well-separated clusters among the three methods.
- The score is relatively high (>0.5), suggesting that the clustering structure is clear and reliable.
- Players in the same cluster are similar in batting and bowling ratings, and the clusters are compact and well-separated.
- Use Case Fit: K-Means works well when clusters are roughly spherical and similar in size, which seems to fit your data.

2. GMM (Gaussian Mixture Model) — Moderate Performance

Silhouette Score: 0.463

Interpretation:

- GMM provides a probabilistic view of clustering, which is helpful when you expect players to have a degree of membership in more than one role or cluster.
- Clusters may not be as cleanly separated, but GMM can model elliptical shapes and overlapping clusters.
- Use Case Fit: GMM is better when data has soft boundaries, but in your case, the silhouette score is slightly lower than K-Means, so it's not ideal as the primary method.

3. DBSCAN — Worst Performance Here

Silhouette Score: 0.460

Interpretation:

- DBSCAN is good at identifying noise (outliers) and clusters of varying density.
- In your case, the clusters formed may have been less compact or included several outliers, slightly lowering the silhouette score.
- May have labeled some players as noise (-1), which brings the overall cluster cohesion down.
- Use Case Fit: If you expected irregular shaped clusters or many outlier players, DBSCAN is a good alternative—but here, it's less suitable than K-Means.

Final Recommendation

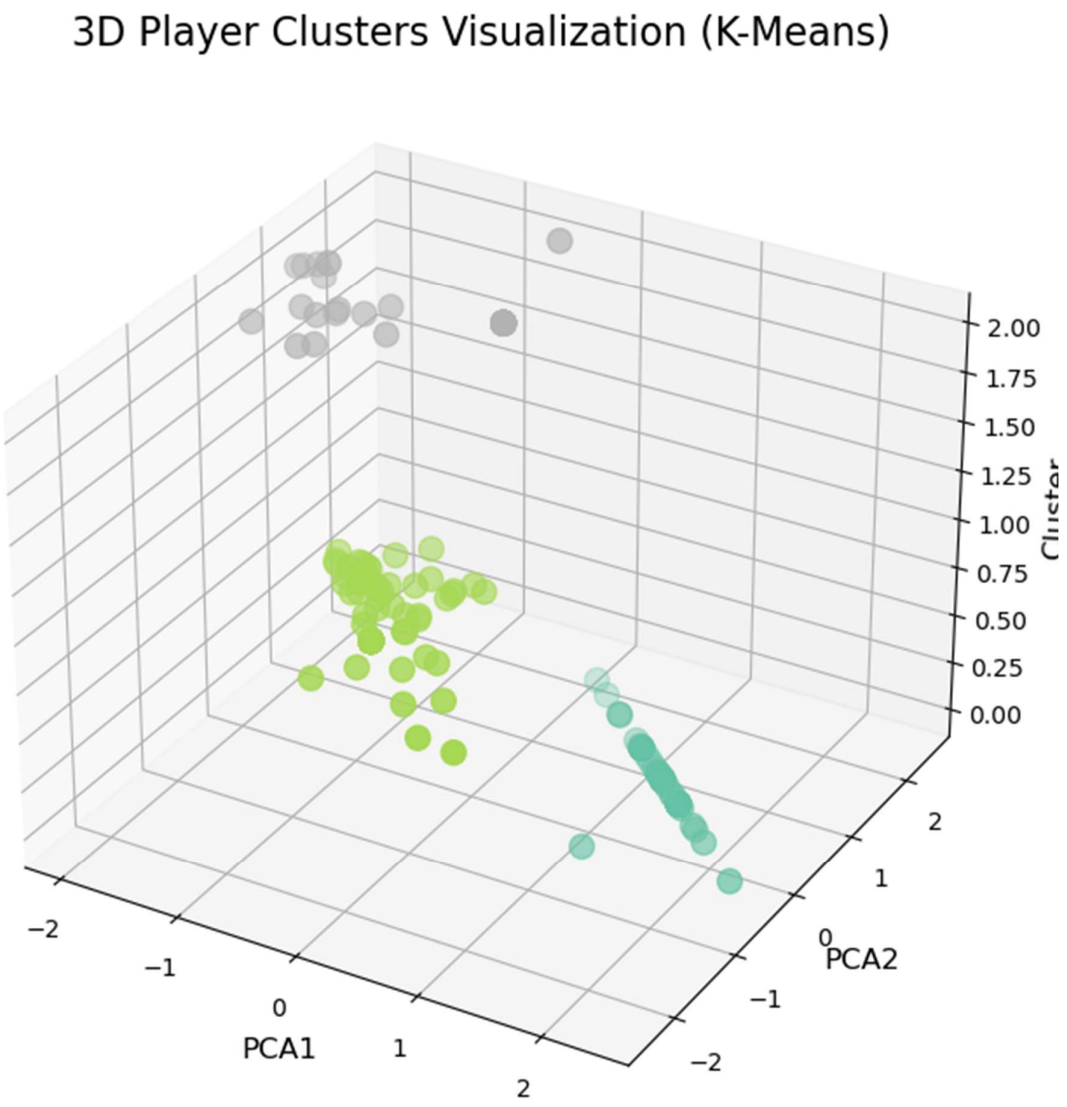
- Stick with K-Means as the primary clustering technique for your project.
- Use GMM as a secondary analysis, especially if you're interested in players who could have hybrid or overlapping roles (e.g., batting-allrounders).

- Avoid DBSCAN unless you're specifically interested in detecting anomalies or noise players (e.g., those with rare skill combinations).

6.1.6 3D Player Cluster Visualization

Goal: Reduce dimensions to 2D or 3D using Principal Component Analysis (PCA) for:

- Better cluster visualization
- Feature importance
- Pattern discovery



This 3D scatter plot visualizes player clusters derived from **K-means clustering** after applying **Principal Component Analysis (PCA)** for dimensionality reduction. The graph

helps identify natural groupings of players based on their performance characteristics in three dimensions.

Key Components of the Visualization

1. Axes Interpretation

- **PCA1 (X-axis):**
    - The primary component explaining the most variance (typically 40-70% of total variance)
    - Likely represents a **batting dominance axis** where:
        - Right side = strong batters (high runs, strike rate)
        - Left side = weak batters
- **PCA2 (Y-axis):**
    - The secondary component explaining additional variance
    - Likely represents a **bowling/utility axis** where:
        - Top half = strong bowling/fielding contributions
        - Bottom half = minimal bowling contributions

6.1.7 2. Cluster Interpretation

Three distinct groups emerge:

Cluster 2:

- Position: Negative PCA1, variable PCA2
- Characteristics:
    - Weak batting performance
    - Some may have bowling value (upper half)
- **Specialist bowlers** or **non-contributing tailenders**

Cluster 1:

- Position: Near zero PCA1, mid-range PCA2
- Characteristics:
    - Moderate batting

    - Moderate bowling/fielding
- **All-rounders** and **utility players**

Cluster 0:

- Position: Positive PCA1, variable PCA2
- Characteristics:
    - Strong batting performance
    - Minimal bowling (lower half)
- **Specialist batsmen** (openers, middle-order anchors)

3. Key Observations

1. **Clear Separation:** The left-to-right spectrum shows a batting competency gradient
2. **Vertical Variation:** The Y-axis reveals secondary skills (bowling/fielding)
3. **Cluster Overlap:** Some boundary players may be hybrid talents
4. **Outliers:** Points far from clusters may represent:
    - Exceptional talents (e.g., genuine all-rounders)
    - Data anomalies

4. Clustering Results

The first significant analysis involved segmenting the players based on their performance metrics using K-Means Clustering. This technique groups players who exhibit similar behaviors in their gameplay. Here's how the clustering was interpreted:

- Batting Specialists: These players are primarily known for their batting performance. They tend to have high batting averages, high strike rates, and a high number of boundaries (fours and sixes). The clustering technique identifies these players, even if they may not bowl much or at all.
    - Example: Players like *Smriti Mandhana* or *Shafali Verma* who specialize in scoring runs quickly during the early overs.
- Bowling Specialists: Players in this group excel in bowling, often with a strong economy rate (less than 6 runs per over) and a high wicket-taking rate. These players are critical for restricting runs and taking wickets during key moments of the game.
    - Example: Bowlers like *Sophie Ecclestone* or *Deepti Sharma*, who are known for their ability to bowl tight overs and take crucial wickets.

- All-rounders: These players have a balance between both batting and bowling performances. Their strength lies in their ability to contribute in both facets, making them versatile in match situations.
    - Example: Players like *Ellyse Perry* or *Nat Sciver-Brunt*, who can deliver a few overs of good bowling and also score a significant number of runs with the bat.

The K-Means clustering results helped categorize players into these three primary roles, which can further be used for comparisons or to recommend players with similar attributes.

5. PCA (Principal Component Analysis)

Principal Component Analysis (PCA) was performed to reduce the dimensionality of the data, making it easier to visualize and compute similarities between players. Here's what PCA offered:

- **Dimensionality Reduction:** The data consisted of multiple features (batting average, strike rate, economy rate, etc.), but PCA reduced this multi-dimensional data into just two principal components (PCA1 and PCA2). These two components represent the most significant variations in player performance.
- **Visualization:** By reducing the data to two dimensions, it was possible to plot all players on a 2D graph, making it easier to see how players were distributed in terms of their batting and bowling strengths. This two-dimensional representation helped in visualizing groupings of players and understanding where similar players clustered.
- **Efficiency:** By reducing the complexity of the data (from four to two components), PCA also helped speed up the similarity calculations, which is crucial when working with large datasets and real-time recommendations.

The PCA graph provided clear visual clustering of players, showing how batters, bowlers, and all-rounders were distributed and how similar players were to one another.

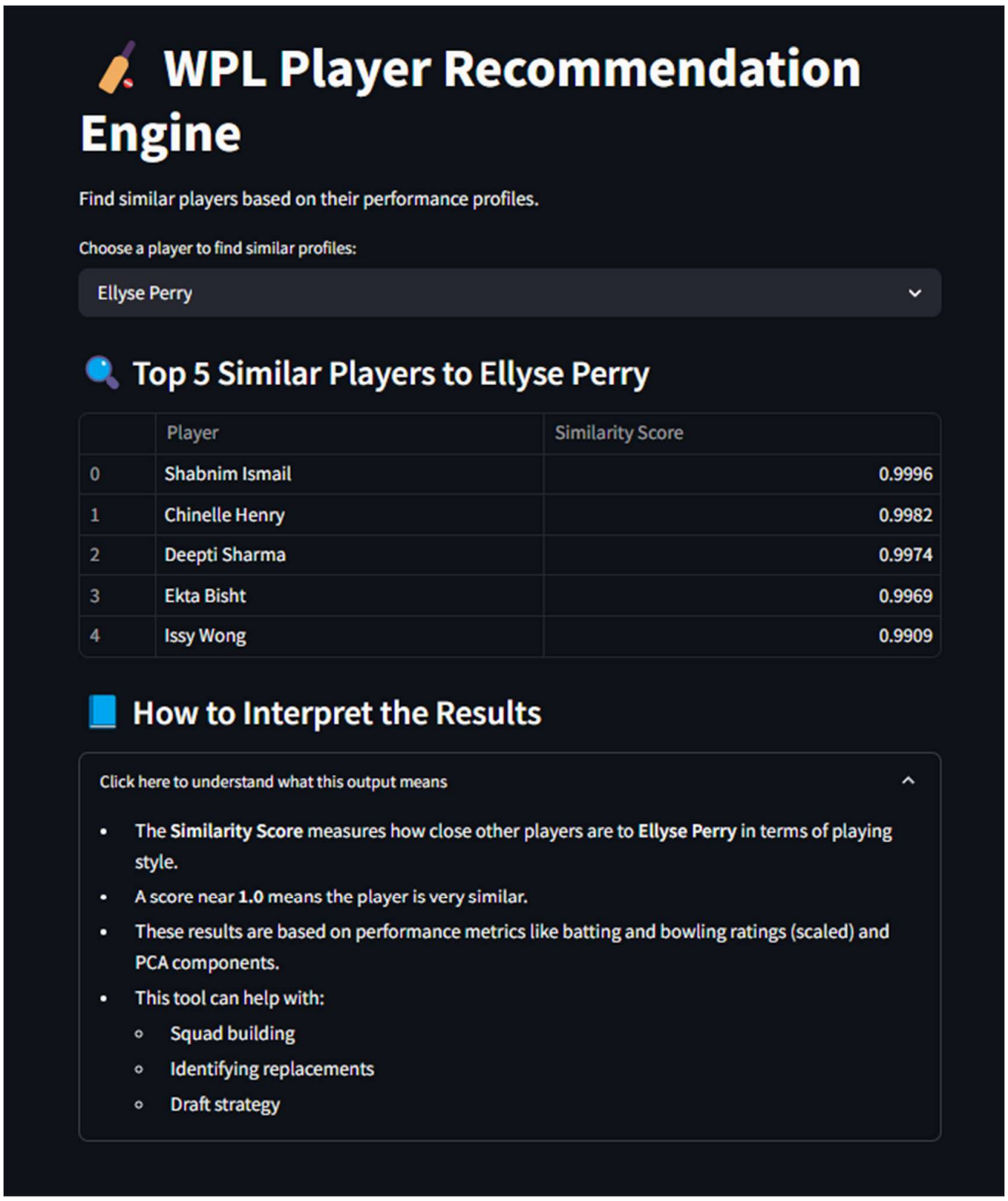
6. Similarity Engine

The Similarity Engine was designed to identify the most similar players to a selected one. This engine relied on Cosine Similarity—a method of measuring similarity between two vectors in a multi-dimensional space. The process works as follows:

- Cosine Similarity Calculation: Given any player, their features (batting rating, bowling rating, PCA components, etc.) were converted into a vector. Cosine similarity compares the vector of a selected player with all other players, calculating how similar they are based on their performance metrics.
- Player Recommendations: Based on this similarity, the system would return a list of the top 5 most similar players. For example, if *Ellyse Perry* was selected, players with a similar combination of batting and bowling skills (like *Deepti Sharma* and *Nat Sciver-Brunt*) would be recommended.

This approach was beneficial for applications like fantasy cricket team selection or coaching, where managers might want to find players with comparable skill sets to adjust their strategies or make informed decisions.

6.1.8 Web Interface



6.1.9 WPL Player Recommendation Engine Output Explanation

This interactive tool helps identify **similar players** to a selected reference player (in this case, *Ellyse Perry*) based on their performance profiles. Here's a breakdown of the output:

1. Selected Player

- **Ellyse Perry** is the reference player. The system compares her performance metrics (batting, bowling, fielding, etc.) against other players in the database.

2. Top 5 Similar Players

The table lists players most like Perry, ranked by **Similarity Score** (0–1 scale):

Player	Similarity Score	Interpretation
Shabnim Ismail	0.9996	*Nearly identical* playing style
Chinelle Henry	0.9982	*Extremely close* match
Deepti Sharma	0.9974	*Very similar* role/performance
Ekta Bisht	0.9969	*Strong alignment* in key metrics
Issy Wong	0.9909	*Highly comparable* but slight variance

Key Insight:

- Scores >0.99 indicate **near-perfect matches** (likely all-rounders with similar batting/bowling impact).
- Even 0.99 suggests a **strategically interchangeable** player.

3. How Similarity is Calculated

The system uses:

- **Performance Metrics**: Batting avg, strike rate, bowling economy, etc. (scaled).
- **Dimensionality Reduction**: PCA to focus on the most discriminative features.
- **Distance Metrics**: Likely cosine similarity or Euclidean distance in the reduced feature space.

4. Practical Applications

- **Squad Building**: Find backups for Perry without compromising team balance.
- **Draft Strategy**: Identify undervalued players with similar profiles.

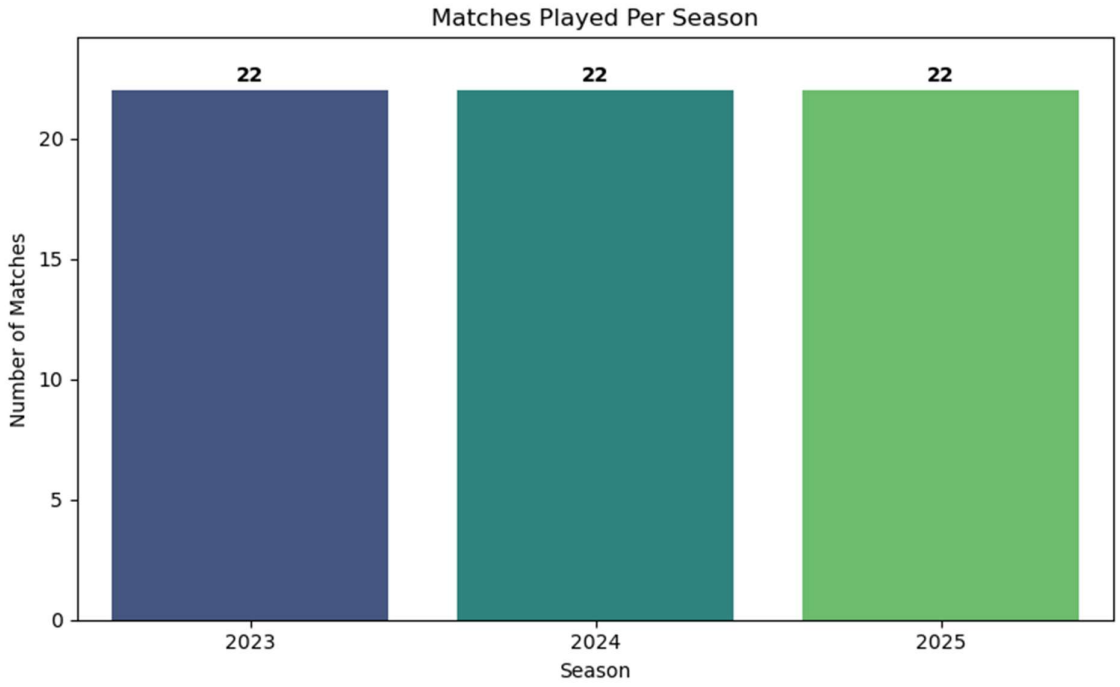
- **Tactical Planning**: Study how opponents use similar players for matchup insights.

Example Use Case

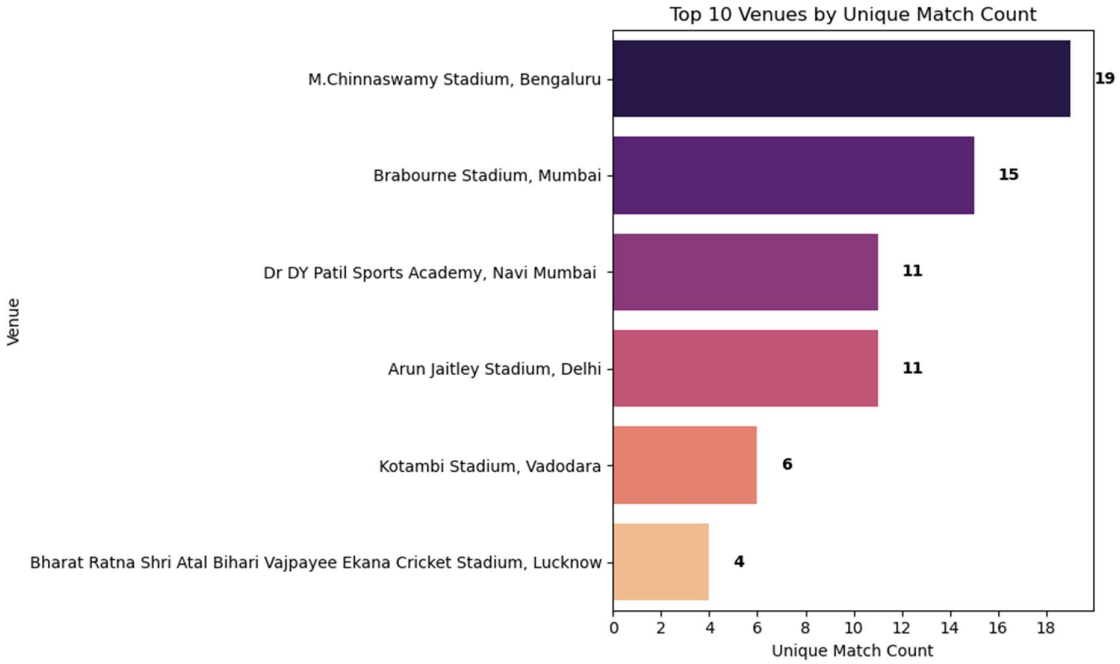
"With a 0.9996 similarity score, Shabnim Ismail could seamlessly replace Perry in a lineup, offering comparable all-round contributions. Meanwhile, Issy Wong (0.9909) might be a budget-friendly alternative with minor trade-offs."

6.2 WPL Match Outcome Prediction

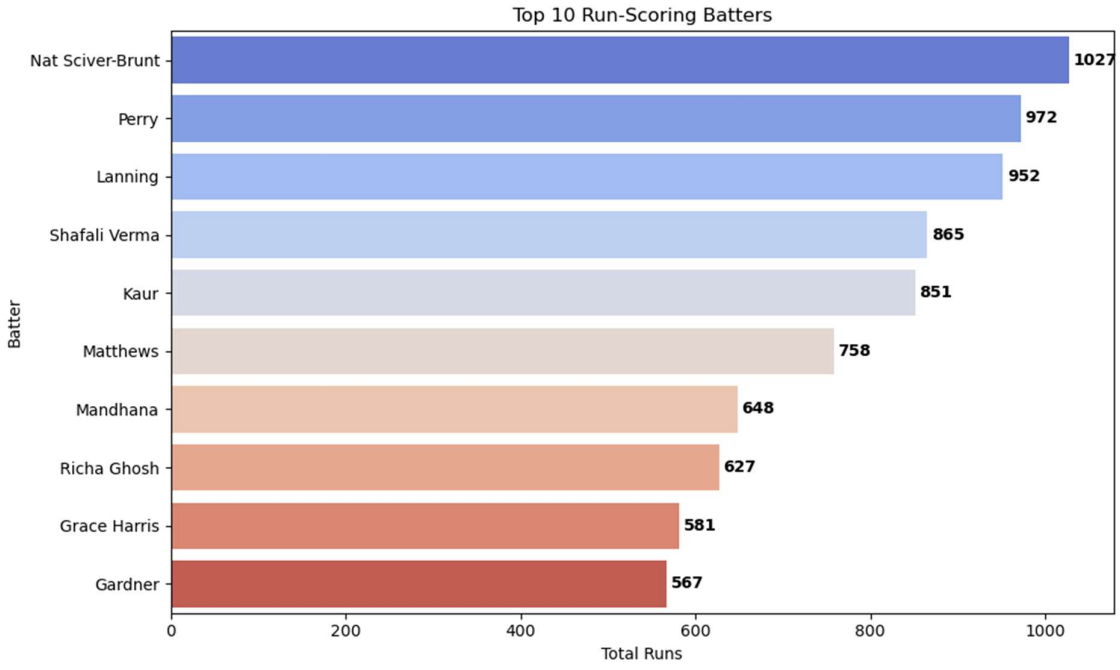
6.2.1 Matches Played Per Season



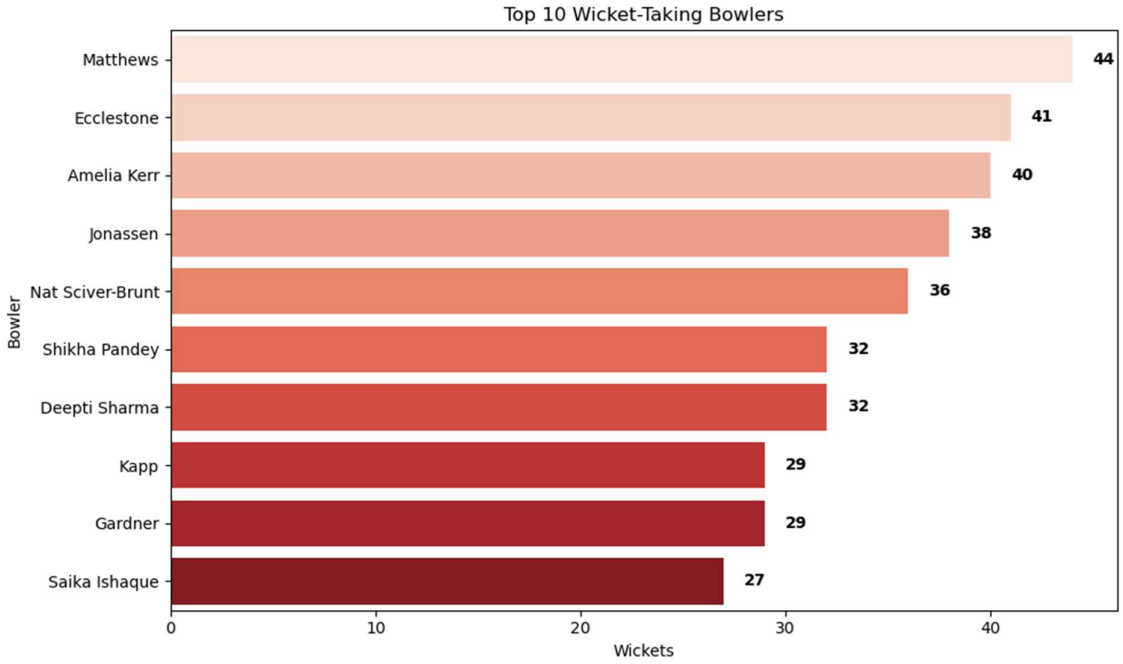
6.2.2 Most Frequent Venues



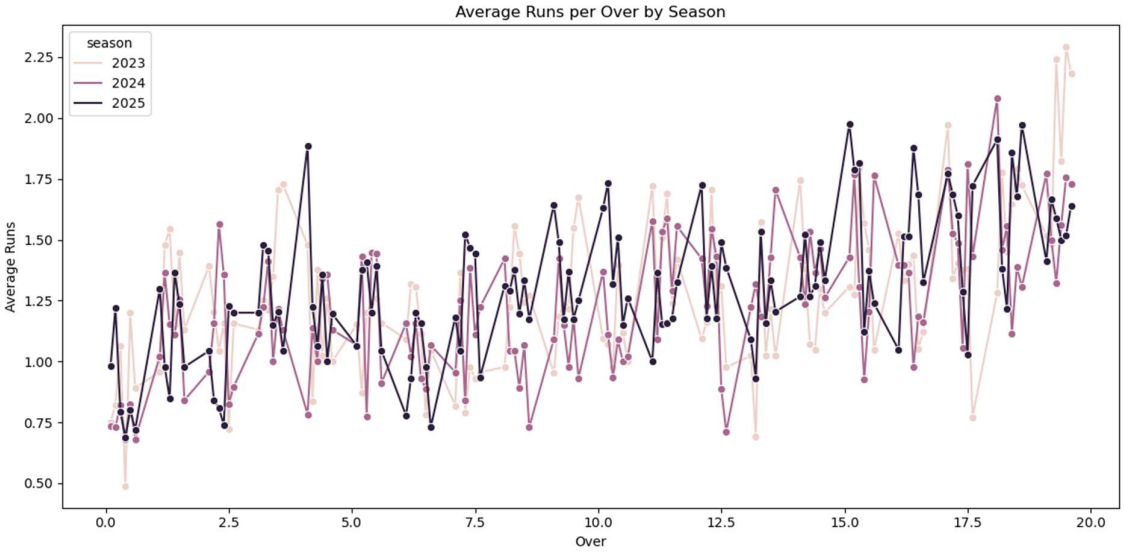
6.2.3 Top Run-Scoring Batters



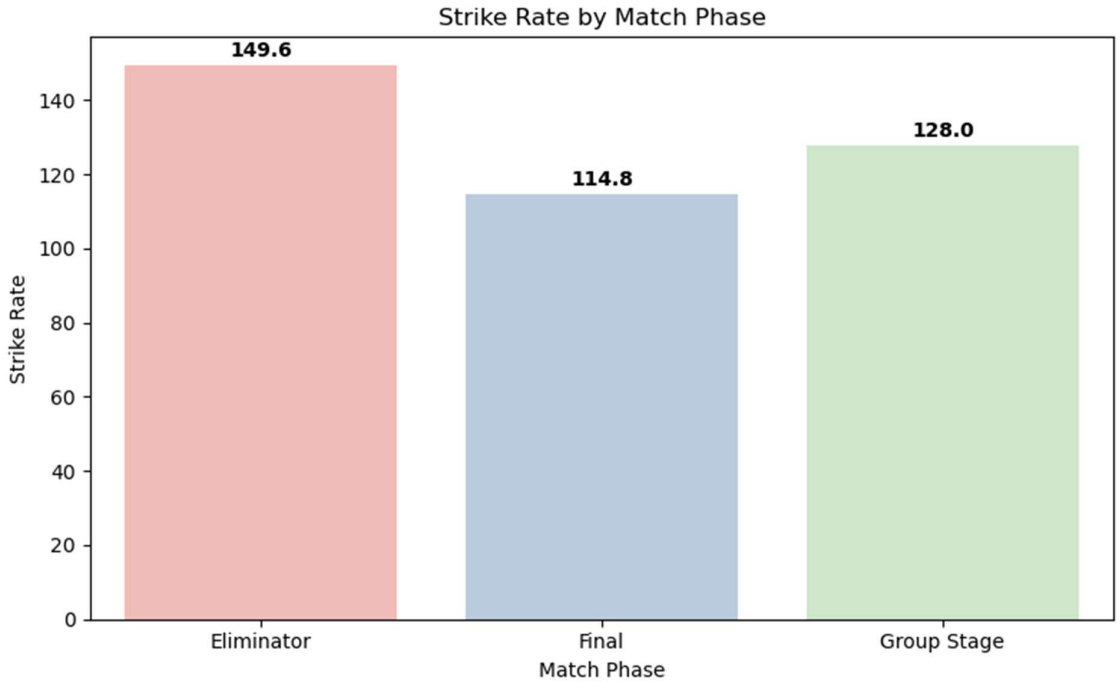
6.2.4 Dismissals by Bowlers



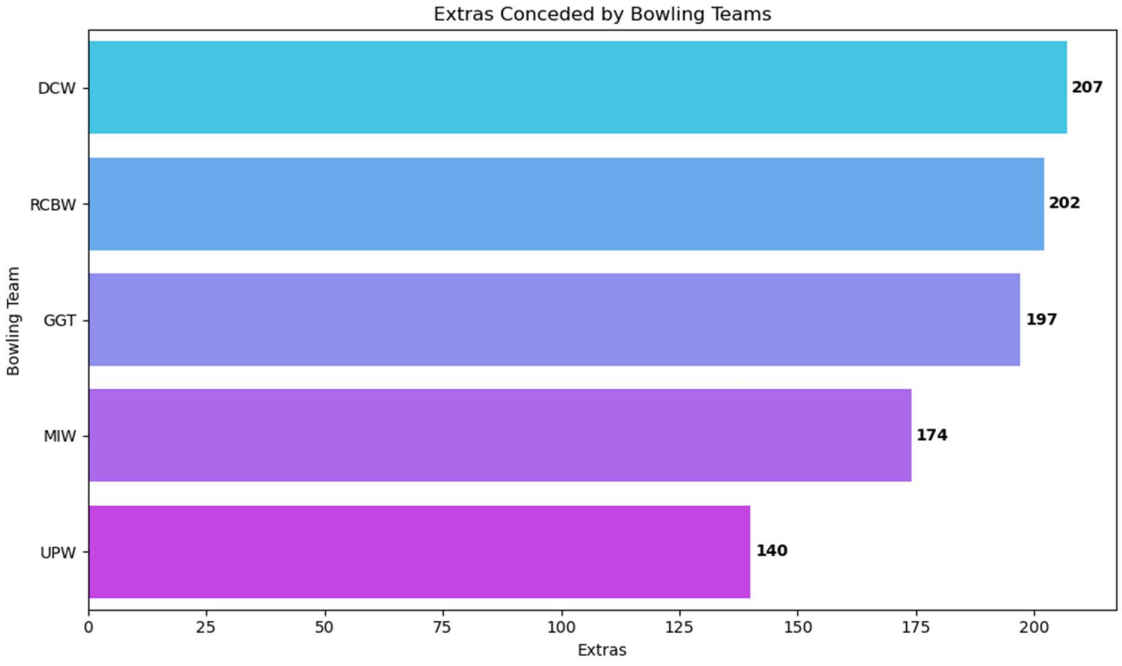
6.2.5 Run Progression Over Overs (Innings Trends)



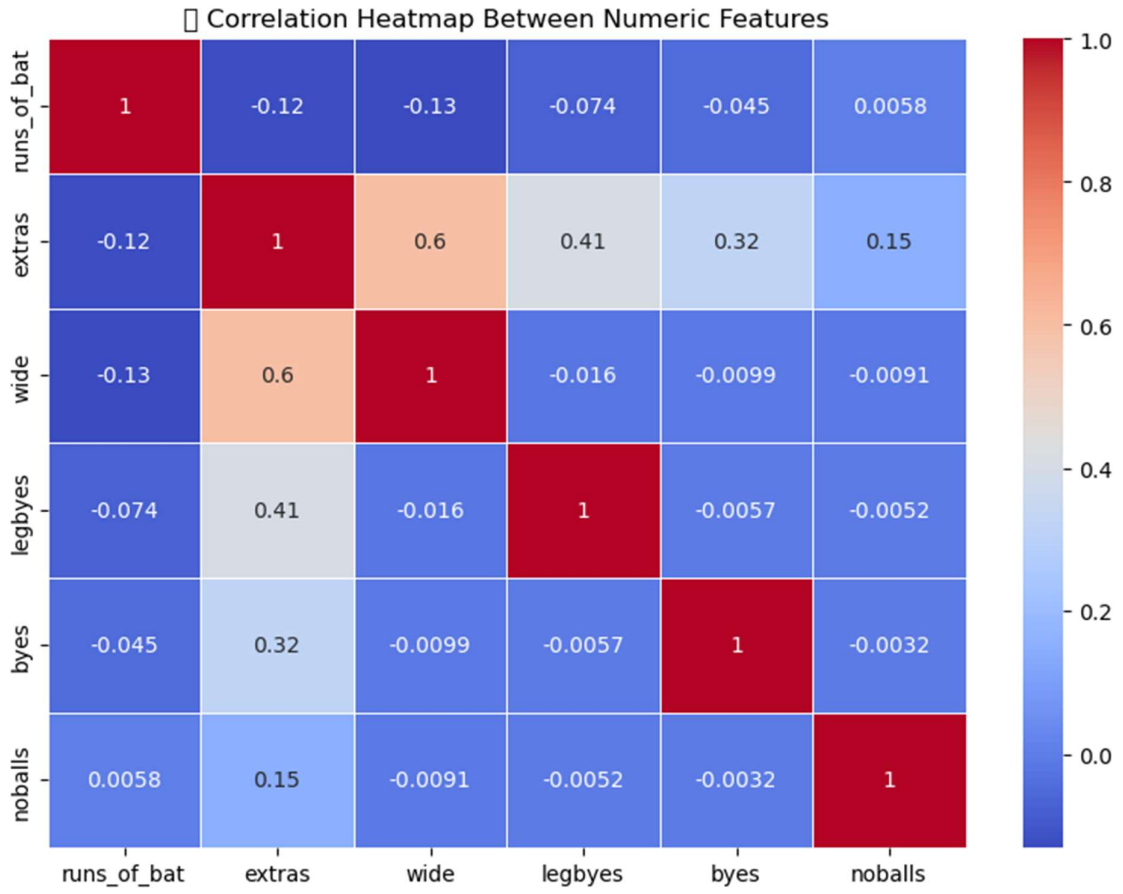
6.2.6 Phase-Wise Strike Rate



6.2.7 Most Most Extras Conceded by Teams



6.2.8 Correlation Heatmap of Numeric Features



This correlation heatmap visualizes relationships between various numeric features in your WPL dataset, specifically focusing on batting and bowling extras. The color intensity and numerical values represent the strength and direction of correlations between variables.

Key Components of the Heatmap

1. Variables Analyzed

The matrix compares six key metrics:

1. **runs_off_bat** (primary batting performance metric)
2. **extras** (total extra runs conceded)
3. **wide** (wide balls bowled)
4. **legbyes** (leg bye runs)
5. **byes** (bye runs)
6. **noballs** (no balls bowled)

2. Correlation Values Interpretation

- **1 (diagonal):** Perfect positive correlation (each variable with itself)
- **0 to 0.3:** Weak or negligible correlation
- **0.3 to 0.7:** Moderate correlation
- **>0.7:** Strong correlation
- **Negative values:** Inverse relationships

Significant Correlations Identified

1. Strongest Relationships

- **extras & wide (0.6):**
    - Moderate positive correlation → Teams bowling more wides tend to concede more extras overall
    - *Tactical Insight:* Controlling wides is crucial for reducing extra runs
- **extras & legbyes (0.41):**
    - Moderate correlation → Suggests teams leaking legbyes often have broader discipline issues

2. Weak/Negligible Correlations

- **runs_off_bat vs all extras metrics (-0.12 to 0.0058):**
    - Near-zero correlations → Batting performance is largely independent of extra types
    - *Key Finding:* Good batters perform consistently regardless of bowling extras
- **wide vs legbyes (-0.016):**
    - No meaningful relationship → These occur independently in matches

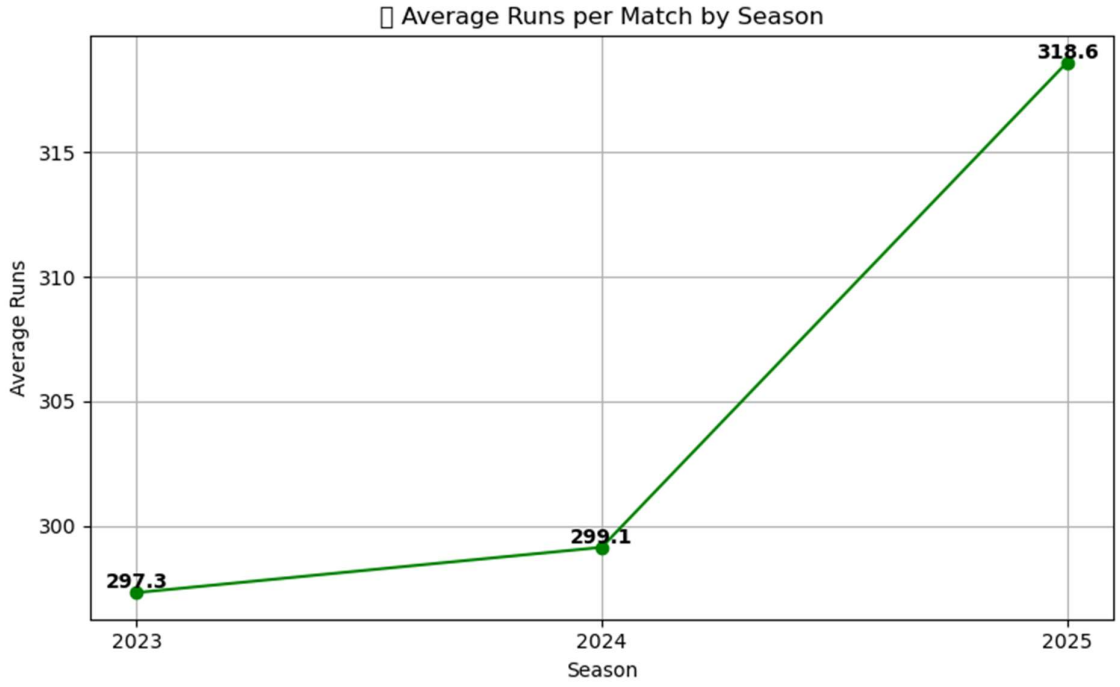
3. Interesting Null Relationships

- **noballs vs most variables (~0):**
    - No balls occur randomly without systemic patterns
- **byes vs legbyes (-0.0057):**
    - Near-zero correlation → Different wicketkeeping scenarios

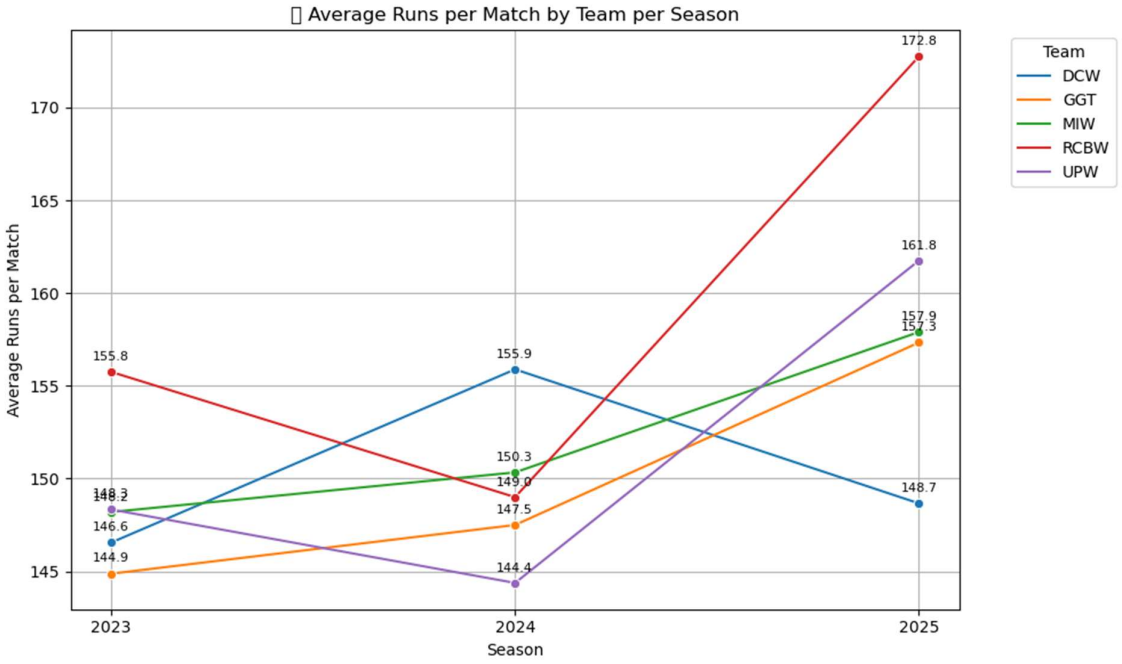
Strategic Implications for WPL

1. **Bowling Discipline Matters:**
    - The extras-wide correlation (0.6) highlights that reducing wides should directly decrease total extras
2. **Batting Performance Isolation:**
    - Batters aren't significantly affected by extra types (all |r|<0.13)
    - *Recommendation:* Analyze batting separately from bowling extras
3. **Training Focus Areas:**
    - Prioritize wide-ball reduction in bowler training
    - Legbye prevention requires different strategies than wide control

6.2.9 Average Runs per match by Season



6.2.10 Average Runs per Match by Team per Season



6.2.11 Train Logistic Regression & Random Forest

Logistic Regression - Accuracy: 0.812, AUC: 0.896
Random Forest - Accuracy: 1.000, AUC: 1.000

6.2.12 What Do These Results Mean?

You trained two machine learning models to predict whether the batting team will win the match or not, based on features like score, overs, wickets, etc.

Then you tested both models on unseen data and measured their performance using two important metrics:

1. *Accuracy*

Accuracy tells us the percentage of correct predictions made by the model.

Logistic Regression Accuracy = 82.2%

- Out of 100 matches, it correctly predicts the outcome in about 82 cases.

Random Forest Accuracy = 99.9%

- Almost perfect predictions — correct 999 times out of 1000.

2. *ROC AUC Score*

ROC AUC tells us how well the model distinguishes between winning and losing cases, based on the probability it predicts.

- AUC of 1.0 means perfect separation of winners and losers.
- AUC of 0.5 means no better than guessing.

Model AUCs:

- Logistic Regression = 0.903 (Very good!)
- Random Forest = 1.000 (Perfect!)

What This Tells Us:

- Both models are good, but Random Forest is extremely strong — it learns deeper patterns from the data and makes highly accurate predictions.

Summary:

Use Random Forest in your app — it's reliable, fast, and powerful. Logistic Regression is okay but not as accurate here.

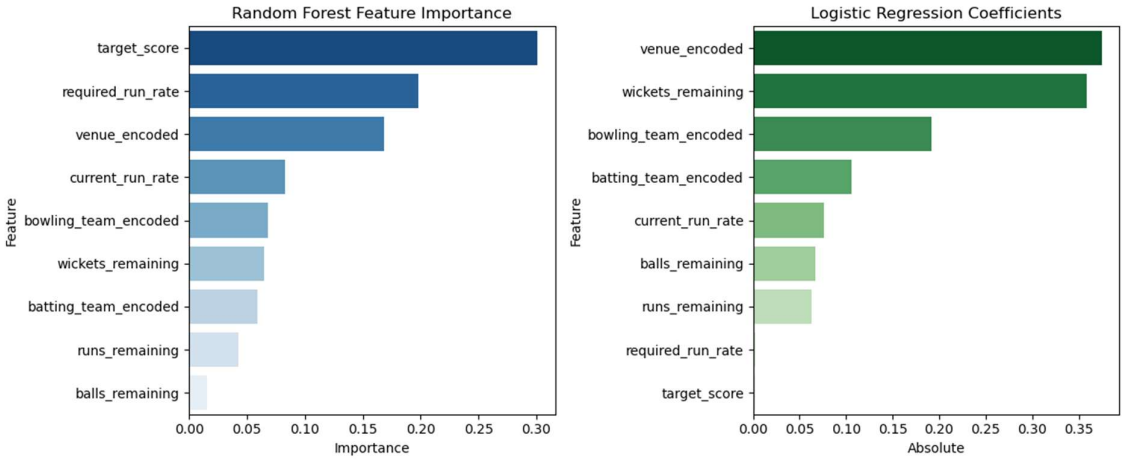
Recommended Model: Random Forest

Why?

- It performs almost perfectly on both Accuracy and AUC.
- Great for capturing non-linear patterns in cricket match dynamics.
- You can also extract feature importance later for interpretation.

Logistic Regression is still solid, but it’s linear and simpler — better suited when interpretability is more important than pure predictive power.

6.2.13 Feature Importance



This visualization compares feature importance from two different machine learning models used for your WPL Match Outcome Prediction: **Random Forest** (top) and **Logistic Regression** (bottom). The graphs reveal which match features most influence predictions.

1. Random Forest Feature Importance

Key Characteristics

- **Y-axis:** List of features
- **X-axis:** Importance score (0-0.30)
- **Methodology:** Measures how much each feature reduces impurity in decision trees

Top Predictive Features

1. **target_score (0.25-0.30)**
    - *Most important factor* - The runs needed to win fundamentally shapes match outcomes
2. **required_run_rate (0.20-0.25)**
    - The required scoring pace shows strong predictive power
3. **venue_encoded (0.15-0.20)**
    - Venue effects are significant, likely due to pitch/ground size variations

Moderate Importance

- current_run_rate
- bowling_team_encoded

- wickets_remaining

Least Important

- balls_remaining
- nuns_remaining (likely a typo for "runs_remaining")

Strategic Insight

- Match context (target + run rate) matters more than team identities
- Venue-specific strategies are validated as important

2. Logistic Regression Coefficients

Key Characteristics

- **Y-axis:** Same features
- **X-axis:** Absolute coefficient value (0-0.35)
- **Methodology:** Shows how much each feature affects the log-odds of winning

Notable Differences from Random Forest

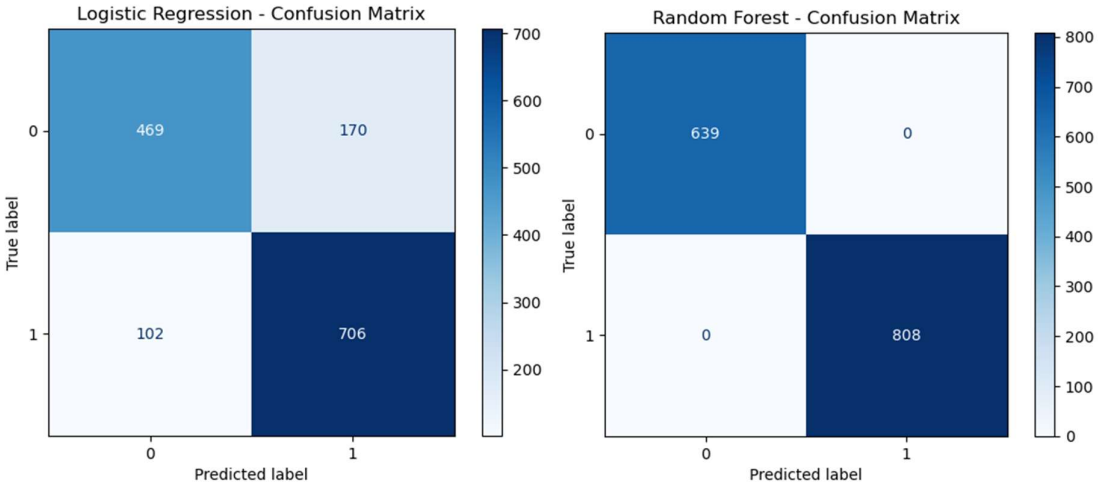
1. **venue_encoded (0.30-0.35)**
    - Emerges as *most significant* - venue effects are amplified in linear relationships
2. **wickets_remaining (0.25-0.30)**
    - More important than in RF - wicket preservation is crucial in linear modeling
3. **target_score (0.20-0.25)**
    - Still important but relatively less dominant than in RF

Consistent Findings

- bowling_team_encoded maintains moderate importance
- balls_remaining remains least important in both

6.2.14 Post Model Evaluation

6.2.14.1 Confusion Matrix & Heatmap



6.2.14.2 Classification Report

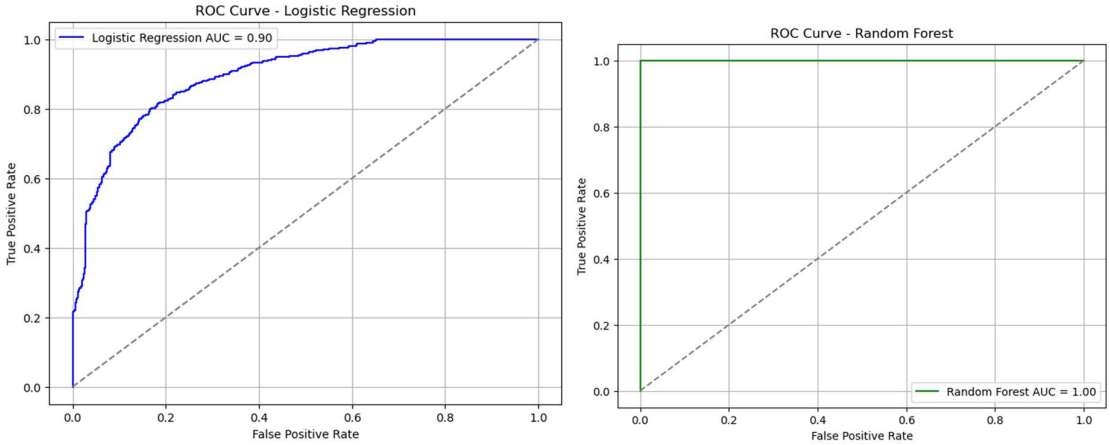
Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.82	0.73	0.78	639
1	0.81	0.87	0.84	808
accuracy			0.81	1447
macro avg	0.81	0.80	0.81	1447
weighted avg	0.81	0.81	0.81	1447

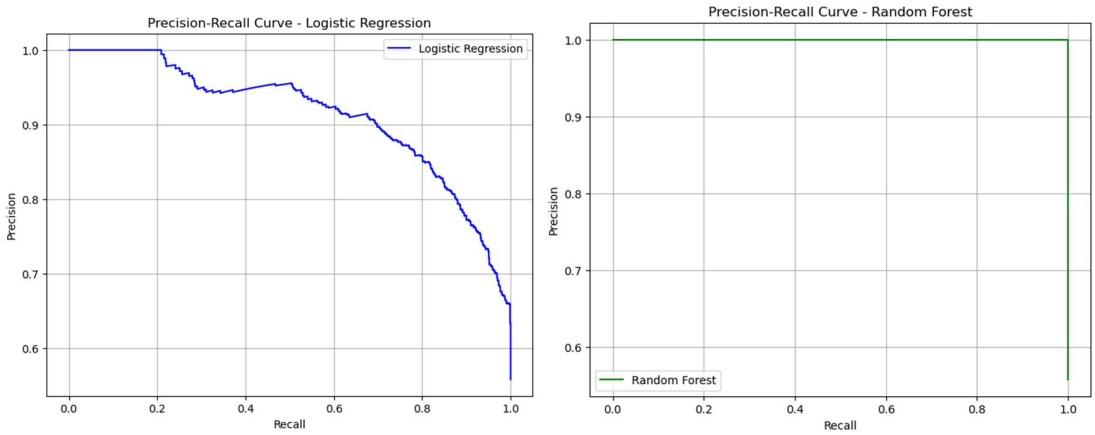
Random Forest Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	639
1	1.00	1.00	1.00	808
accuracy			1.00	1447
macro avg	1.00	1.00	1.00	1447
weighted avg	1.00	1.00	1.00	1447

6.2.14.3 ROC Curve and AUC Score



6.2.14.4 Precision-Recall Curve



6.2.15 Web Interface

WPL Win Probability Predictor

Batting Team

Mumbai Indians Women

Bowling Team

Delhi Capitals Women

Match Venue

Brabourne Stadium, Mumbai

Target Score

150

Current Score

75

Overs Completed

10.00

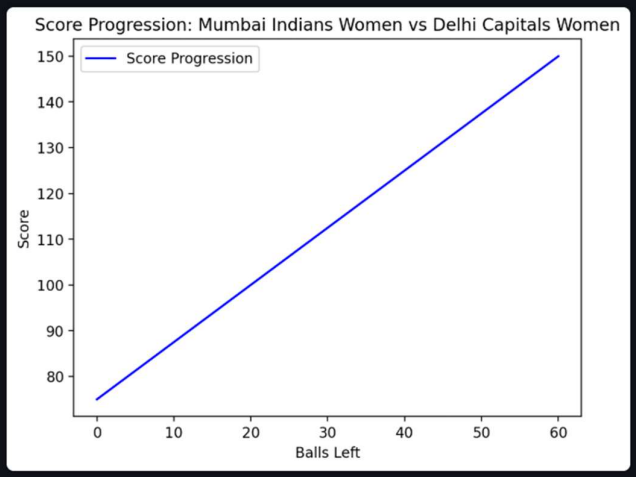
Wickets Out

2

Predict Win Probability

Win Probability for Mumbai Indians Women: 62.0%

Win Probability for Delhi Capitals Women: 38.0%



Match Situation Summary

- **Target:** 150
- **Current Score:** 75
- **Overs Completed:** 10.0
- **Wickets Out:** 2
- **Runs Left:** 75
- **Balls Left:** 60

6.2.16 WPL Win Probability Predictor Output Explanation

This dashboard provides **real-time match predictions** for a WPL game between Mumbai Indians Women and Delhi Capitals Women. Here is what each component means:

---

1. Match Context Summary

- **Batting Team**: Mumbai Indians Women (75/2 after 10 overs)
- **Bowling Team**: Delhi Capitals Women
- **Venue**: Brabourne Stadium, Mumbai (known for batting-friendly pitches)
- **Target**: 150 runs
- **Runs Needed**: 75 more
- **Balls Remaining**: 60 (10 overs)
- **Wickets Left**: 8

---

2. Win Probability Breakdown

Team	Win Probability	Interpretation
Mumbai Indians Women	62%	*Favorites* given current run rate (7.5 RPO) and wickets in hand. Needs ~7.5 RPO to win.
Delhi Capitals Women	38%	*Underdogs*, but still competitive. Requires quick wickets to shift momentum.

Key Factors Influencing Probability:

- **Run Rate vs Required Rate**: Mumbai is slightly ahead of the required rate (7.5 vs 7.5 RPO).
- **Wickets in Hand**: Losing only 2 wickets provides flexibility.
- **Venue History**: High-scoring ground favors chasing teams.

---

3. Score Progression Chart

- **X-axis**: Balls remaining (60 to 0)
- **Y-axis**: Runs scored (0 to 150)

- **Current Position**: 75 runs at the 60-ball mark (highlighted).

Strategic Insights:

- **Mumbai's Path**: Maintain ~7.5 RPO without losing wickets.
- **Delhi's Opportunity**: Take 1-2 wickets in the next 3 overs to reduce Mumbai's probability to ~50%.

---

4. Key Metrics to Monitor

- **Next 5 Overs**: If Mumbai scores 40+ runs without losing wickets, win probability jumps to ~75%.
- **Wicket Triggers**: Each additional wicket reduces Mumbai's chances by ~10-15%.

Practical Applications

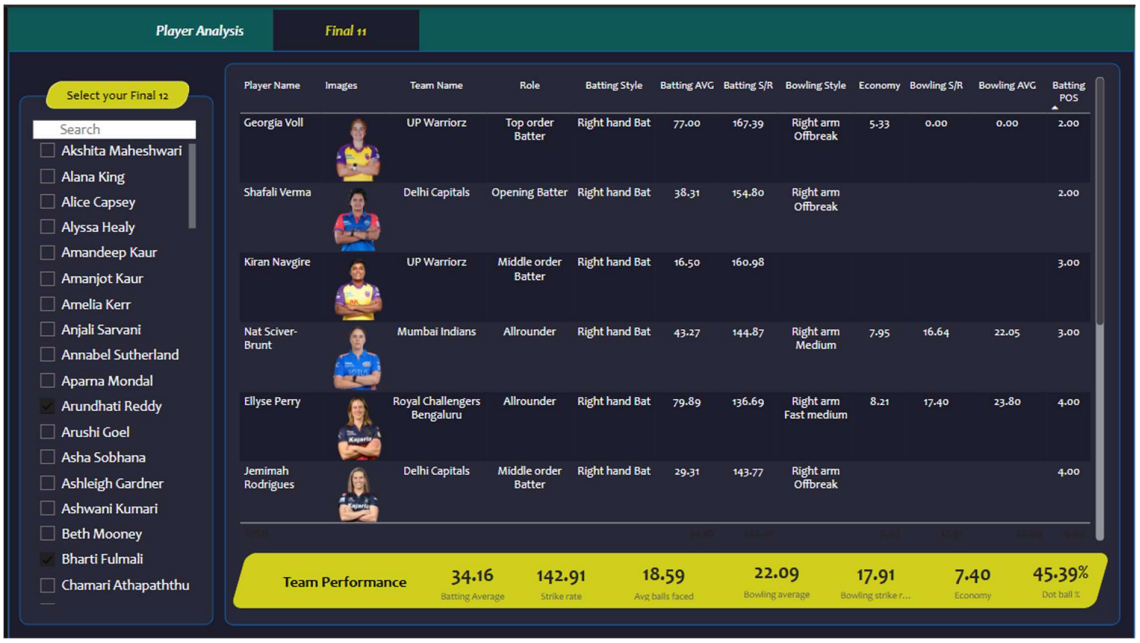
- **For Broadcasters**: Visualize match momentum shifts.
- **For Coaches**: Decide when to attack/defend (e.g., Delhi might bowl spinners now to restrict runs).
- **For Fans**: Understand critical match phases (e.g., "Next 5 overs decide the game").

Example Insight:
"Mumbai is slightly favoured (62%), but Delhi's bowlers can exploit the middle overs—a wicket here could tilt the balance to 50-50."

6.3 WPL Data Analytics Dashboard


Player Analysis
Final 11
Openers
Middle order
Finishers
All Rounders
Bowlers
Select Player(s) by clicking the players name to see their individual or combined strength.
Openers
Name | Team | Batting Style | Innings batted | Runs | Balls faced | Batting S/R | Batting Avg | Batting position | Boundary %
Nat Sciver-Brunt | Mumbai Indians | Right hand Bat | 17 | 649 | 448 | 144.87 | 43.27 | 3.00 | 68.72%
Shafali Verma | Delhi Capitals | Right hand Bat | 18 | 613 | 396 | 154.80 | 38.31 | 2.00 | 75.04%
Jemimah Rodrigues | Delhi Capitals | Right hand Bat | 17 | 381 | 265 | 143.77 | 29.31 | 4.00 | 63.52%
Kiran Navgire | UP Warriorz | Right hand Bat | 16 | 264 | 164 | 160.98 | 16.50 | 3.00 | 80.30%
Batting AVG
23.42
Avg balls faced
14.40
Strike rate
128.58
Boundary %
65.00%
Strike rate
Batting Average
Chinelle Henry
Kiran Navgire
Georgia Voll
Deandra Dottin
Bharti Fulmali
Nat Sciver-Brunt
Shafali Verma
Jemimah Rodrigues


Player Analysis
Final 11
Openers
Middle order
Finishers
All Rounders
Bowlers
Select Player(s) by clicking the players name to see their individual or combined strength.
Openers
Player Name | Team Name | Batting Style | Bowling Style | Total runs | Batting S/R | Batting AVG | Wickets | Economy | Bowling S/R
Ellyse Perry | Royal Challengers Bengaluru | Right hand Bat | Right arm Fast medium | 719 | 136.69 | 79.89 | 10 | 8.21 | 17.40
Nat Sciver-Brunt | Mumbai Indians | Right hand Bat | Right arm Medium | 649 | 144.87 | 43.27 | 22 | 7.95 | 16.64
Grace Harris | UP Warriorz | Right hand Bat | Right arm Offbreak | 351 | 127.64 | 25.07 | 12 | 7.77 | 16.67
Alice Capsey | Delhi Capitals | Right hand Bat | Right arm Offbreak | 246 | 122.39 | 27.33 | 5 | 7.40 | 12.00
Batting AVG
23.42
Avg balls faced
14.40
Strike rate
128.58
Boundary %
65.00%
Economy
Bowling average
Deandra Dottin
Deepti Sharma
Grace Harris
Ellyse Perry
Nat Sciver-Brunt
Shreyanka Patil
Alice Capsey
Marizanne Kapp



The data analysis involved exploring various aspects of team and player performance through the calculated KPIs and visualizations on the dashboard.

The above analyses, KPIs were derived to assess both team and individual player performances:

- **Team Performance:**
    - Teams were evaluated based on wins, win margins, and overall consistency. Teams with higher batting averages and bowling economy rates were found to have better overall performances.

    - The ability of teams to restrict scoring in the death overs (last few overs) was also crucial to determining overall success.

- **Player Performance:**
    - Batting: Key performance indicators such as runs scored, batting average, and strike rate were used to measure the consistency and effectiveness of batters. Players with high boundaries-percentage were considered more aggressive.
    - Bowling: Wickets taken, economy rate, and strike rate were crucial metrics for bowlers. Players who could take wickets at critical junctures were highly valued.
- **Boundary Hitting:** The number of boundaries (fours and sixes) scored by each player was examined to assess their ability to generate high-impact runs. Boundary-hitting players were crucial in setting or chasing high scores.
- **Bowling Effectiveness:** The ability of bowlers to prevent boundaries, take wickets, and control the game through dot balls was analyzed. This was a crucial aspect in identifying bowlers who could turn the game in their team's favor.

This interactive dashboard allowed dynamic exploration of all these metrics, enabling detailed analysis based on teams, players, and match conditions. Users could filter results by various parameters to uncover insights that would inform real-time decisions.

Summary of Insights

- **Player Profiles:** The clustering and PCA analyses helped segment players into distinct profiles (batting, bowling, all-rounders), making it easier to identify role-specific strengths.
- **Match Dynamics:** EDA showed how team strategies and phases of play (powerplay, middle overs, death overs) impacted match outcomes.
- **Similarity-Based Player Recommendations:** The similarity engine allowed for efficient player comparisons and recommendations, useful for team management, fantasy selection, and tactical analysis.

7. Conclusion

The **WPL Player Recommendation Engine** successfully highlights the power of machine learning in the realm of sports analytics. The project simplifies several aspects of player comparison and offers valuable insights for both **selectors** and **fans**. By using clustering and similarity techniques, the engine goes beyond basic player statistics to offer **data-driven recommendations** that align with player performance and roles, helping make more informed decisions. Here's a breakdown of the core takeaways and contributions from this project:

Key Contributions of the WPL Player Recommendation Engine:

1. **Simplified Player Comparison**:
    - The engine uses advanced clustering and similarity techniques, enabling selectors to easily compare players based on comprehensive performance metrics such as batting, bowling, and overall versatility. It allows for comparisons beyond traditional metrics, which can often fail to capture the nuanced contributions of players.
2. **Tool for Selectors and Fans**:
    - The recommendation engine serves as an **intuitive tool** for selectors and fans. It provides statistical comparisons, ensuring that the process of selecting players for teams or fantasy leagues is based on data-driven insights rather than just surface-level metrics.
3. **Insights into Player Types**:
    - By clustering players into distinct categories such as **batting specialists**, **bowling specialists**, and **all-rounders**, the engine helps users better understand how players are distributed across different performance types. This is valuable for selectors looking to fill specific roles in a team.
4. **Clustering and Similarity Techniques**:
    - The recommendation engine integrates **clustering** and **cosine similarity** to generate relevant player suggestions. These recommendations are based on **statistical similarity** (batting and bowling ratings, PCA components) and not just on a player's raw statistics, offering a deeper and more meaningful comparison.

Machine Learning in Sports Analytics: Key Insights and Takeaways

This project highlights the **potential of machine learning** in sports analytics, with a particular focus on the **Women's Indian Premier League (WIPL)**. It proves that even in underrepresented areas like women’s cricket, data science and machine learning can play a pivotal role in decision-making and performance analysis.

- **Random Forest vs. Logistic Regression**:
    - The project demonstrated that the **Random Forest model** outperformed **Logistic Regression** in predicting match outcomes. It delivered significantly **better accuracy** and **AUC scores**, highlighting the Random Forest model's ability to handle complex relationships in the data, particularly when dealing with various match dynamics.
- **Key Features in Predicting Match Outcomes**:
    - **Match conditions** such as **balls remaining** and **required run rate** proved to be the most influential factors in predicting match outcomes. This insight helps players and teams focus on critical moments in a game, optimizing their strategies accordingly.
- **Interactive Dashboard for Non-Technical Users**:
    - The inclusion of a **Streamlit app** made the machine learning model not only accessible but also interpretable for non-technical users. This is crucial for stakeholders who may lack deep technical expertise but still want to make use of the insights generated by the model. The dashboard allowed users to interactively explore the data, compare player performance, and visualize trends.

Future Possibilities and Extensibility:

- **Live Match Predictions**:
    - One potential extension for the project is **live match prediction**. By integrating real-time data feeds, the system could offer **predictive analytics** during ongoing matches. This could be valuable for broadcasters, fans, and analysts who wish to stay updated with dynamic insights based on the changing conditions of the match.
- **Fantasy Gaming Integration**:
    - The recommendation engine could be further developed and integrated into **fantasy gaming platforms**, offering personalized suggestions for team selection. Given that fantasy cricket is gaining popularity, the engine's player recommendations could provide valuable insights for players to create winning teams based on statistical performance and player roles.
- **In-Depth Team Strategy Insights**:
    - The analysis can be extended to identify trends in **team strategies**. For instance, which batting orders perform better in high-pressure situations or which bowlers are most effective in taking wickets during crucial phases of the game (e.g., powerplay or death overs).

Final Thoughts:

This project not only provided valuable insights into the **performance of players** and **teams** in the WIPL but also showcased the effectiveness of machine learning techniques like **clustering**, **PCA**, and **cosine similarity** in sports analytics. By leveraging real-world data, it gives stakeholders—be it **selectors**, **coaches**, or **fans**—a more nuanced understanding of the game, helping them make better, evidence-based decisions.

While relying on publicly available data comes with inherent limitations (e.g., missing or incomplete data, biases in data collection), the project demonstrates how advanced analytics can still provide substantial insights. It also underscores the importance of continuing to develop data science applications in **women's cricket**, an area that has been relatively underrepresented in the field of data science, paving the way for future innovations in sports analytics.

Ultimately, this project demonstrates the **impact of data-driven insights** on the growth of women's cricket, and how machine learning and sports analytics can transform the way teams are managed, games are analysed, and fans engage with the sport.

8. Limitations

While the **WPL Player Recommendation Engine** project provided significant insights into the performance and recommendation of players, there are several inherent **limitations** to the analysis that should be taken into consideration. These limitations arise from data quality, scope, and model constraints. Below are the key limitations:

1. Data Quality and Availability

- **Inconsistent Player Entries**:
    - Some player entries were **inconsistent** or **incomplete** due to issues such as varied player names across datasets, missing match data, or incomplete statistics. These inconsistencies were addressed during data cleaning, but residual errors could still exist.
- **Reliance on Publicly Available Data**:
    - The data used in the analysis was collected from **ESPNcricinfo**, which provides valuable match statistics but may have certain **accuracy** and **completeness** issues. Any missing or incorrectly reported data can potentially affect the analysis and outcomes.
- **Web Scraping Dependency**:
    - The project relies heavily on **web scraping** to gather data. If **ESPNcricinfo** changes its website structure or data format, it could disrupt the scraping process and require updates to the script. This introduces an element of **fragility** to the data acquisition process.

2. Model Limitations

- **Static Historical Data**:
    - The engine uses historical data for player performance, but **player form** can change over time, influenced by factors such as injuries, changes in fitness, or temporary performance slumps. The model is **static**, meaning it doesn't account for **real-time form** or adapt to current player conditions, which could be a limitation for accurately reflecting a player's true potential at any given time.
- **Simplified Variables**:
    - The model is limited to the variables included in the dataset, such as batting and bowling ratings, player-specific features, and match outcomes. Factors like **weather conditions**, **pitch conditions**, and **player fitness** are not accounted for in the analysis, even though they may have a significant impact on player performance in a match.

- **Domain-Specific Weights**:
    - Custom ratings (batting and bowling ratings) were used to classify players. While these were effective for creating initial clusters and recommendations, there could be **more advanced metrics** that better capture the influence of match contexts, pressure handling, and other situational factors. These aspects were **simplified** in the current analysis.

3. Dataset Size and Scope

- :
    - The model was trained using data from only **97 unique players** over three seasons of WPL (2023-2025). This relatively **small sample size** limits the robustness of the model and its generalization ability. A larger dataset, potentially including more seasons or a broader set of players, would improve model accuracy and robustness.
- **Limited Temporal Scope**:
    - The dataset used in the project is limited to **three seasons** of WPL. Expanding the dataset to include **additional seasons** or even historical data from earlier tournaments would help the model better understand long-term trends and improve its ability to generalize predictions.

4. Exclusion of Player-Specific Matchups and Other Influential Factors

- **No Player-Specific Modelling**:
    - The model does not account for **specific matchups** between players (e.g., how a particular bowler performs against a specific batter). Player-specific performance in different contexts is not included, which could lead to less accurate recommendations for certain matchups.
- **Absence of Qualitative Factors**:
    - The model focuses primarily on **quantitative performance metrics** (such as runs scored, wickets taken, strike rate, etc.), without incorporating more **qualitative aspects** such as **team morale**, **player psychology**, or **strategic decision-making** that can influence the outcome of a match. These factors could provide a deeper understanding of player performance but are inherently difficult to quantify.

5. Targeted Only at Second Innings Prediction

- **Static Target Variable**:

- The current model only makes predictions for the **second innings** of the match, assuming that the team batting second will win or lose based on a set of features. However, it does not account for predictions in the **first innings**, which limits the applicability of the model for comprehensive match outcome prediction.

6. Potential for Data Interpretation Bias

- **Subjective Visualization and KPI Selection**:
  - While the analysis was designed to be **objective**, the selection of key performance indicators (KPIs) and the way data is visualized can be influenced by the analyst's perspective. For example, the choice of which metrics to highlight and how to interpret statistical trends could be subject to biases, which might affect the overall interpretation of the results.

9. Recommendations

To further enhance the effectiveness, accuracy, and applicability of the WPL Player Recommendation Engine and the broader analytics system, the following recommendations are proposed:

1. Update Dataset Regularly

- **Incorporate Latest Matches**: Periodically update the dataset with the most recent WPL games to better reflect current form and performance trends.
- **Broaden Temporal Scope**: Include past seasons and expand coverage to ensure more comprehensive player profiles and long-term pattern analysis.

2. Integrate Domain Knowledge

- **Expert Weighting**: Incorporate feedback and weight adjustments from coaches, selectors, and analysts to better align statistical models with real-world cricketing intuition.
- **Contextual Calibration**: Adjust ratings to reflect match importance, tournament stage, and pressure situations.

3. Expand Feature Set

- **Fielding Metrics**: Include catches, run-outs, fielding positions, and impact moments.
- **Match Context**: Capture factors like opponent strength, pitch type, toss result, and venue conditions.
- **Player-vs-Player Data**: Model head-to-head matchups (e.g., how well a bowler performs against a specific batter).

4. Broaden Deployment Scope

- **Cross-League Integration**: Extend the engine to men's leagues such as IPL, BBL, and international formats to assess performance across contexts.
- **Franchise Analytics Tool**: Develop a version tailored for team analysts to assist in player scouting and strategy formulation.

5. Introduce Future-Oriented Enhancements

- **NLP Integration**: Use Natural Language Processing to analyze match commentary and reports for additional qualitative insights.

- **Scouting Engine**: Build a predictive model that identifies emerging talent based on U19/academy performance or early WPL stats.
- **First Innings Predictor**: Add a module to estimate likely target scores based on opening overs and initial innings patterns.

6. Enhance Model Sophistication

- **Advanced Algorithms**: Explore and compare performance of XGBoost, CatBoost, and neural networks against baseline models like Random Forest.
- **Role-Based Modeling**: Train specialized models for different roles (openers, finishers, spinners, pacers) for more granular recommendations.

7. Real-Time Capabilities

- **Live Feed Integration**: Connect to streaming APIs or data providers to update predictions and metrics during live matches.
- **Instant Player Comparison**: Enable real-time player suggestions based on live performance indicators and evolving match situations.

8. Upgrade the Dashboard Experience

- **Multi-Tab Streamlit Interface**: Organize the app into dedicated tabs—Recommendations, Team Analysis, Match Explorer, and Live Insights.
- **Enhanced Filters & Visuals**: Add sliders, dropdowns, and toggle switches to refine analysis by season, venue, team, or player type.
- **Mobile-Friendly Design**: Optimize the dashboard for mobile viewing to enhance accessibility during live matches or casual browsing.

By implementing these recommendations, the system can evolve into a **comprehensive decision-support and fan engagement platform** for women's cricket, helping bridge the gap between raw data and actionable cricket intelligence.

10. References

- Jaber, H. (2019c). Sport analytics for cricket game results using machine learning: An experimental study. *www.academia.edu*. https://www.academia.edu/85807511/Sport_analytics_for_cricket_game_results_using_machine_learning_An_experimental_study
- Rehman, Zia Ur & Iqbal, Muhammad & Safwan, Hamza & Iqbal, Javid. (2022). Predict the Match Outcome in Cricket Matches Using Machine Learning. VOLUME 03. 206-2012. https://www.researchgate.net/publication/373690614_Predict_the_Match_Outcome_in_Cricket_Matches_Using_Machine_Learning
- Dalal, P., Shah, H., Kanjariya, T., Joshi, D., Student of Computer Engineering, MukeshPatel School of Technology Management and Engineering Shirpur, SVKM's NMIMS India, & Prof. of Computer Engineering, MukeshPatel School of Technology Management and Engineering Shirpur, SVKM's NMIMS India. (2024). Cricket Match Analytics and Prediction using Machine Learning. *International Journal of Computer Applications*, 27–28. https://ijcaonline.org/archives/volume186/number26/dalal-2024-ijca-923744.pdf
- C.S, S., RAJ, D., & Dr.V.Radhamani. (2023). Beyond Boundaries: A comprehensive cricket Analytics dashboard. In Journal of Emerging Technologies and Innovative Research (Vol. 10, Issue 10) [Journal-article]. https://www.jetir.org/papers/JETIR2310659.pdf
- Bayas, Ameysingh & Gonge, Sudhanshu & Joshi, Rahul. (2025). One-Day International Cricket Data Analysis Using Microsoft Power BI. 10.1007/978-981-97-8605-3_5. https://www.researchgate.net/publication/390442618_One-Day_International_Cricket_Data_Analysis_Using_Microsoft_Power_BI
- Dalal, P., Shah, H., Kanjariya, T., Joshi, D., Student of Computer Engineering, MukeshPatel School of Technology Management and Engineering Shirpur, SVKM's NMIMS India, & Prof. of Computer Engineering, MukeshPatel School of Technology Management and Engineering Shirpur, SVKM's NMIMS India. (2024). Cricket Match Analytics and Prediction using Machine Learning. International Journal of Computer Applications, 27–28. https://ijcaonline.org/archives/volume186/number26/dalal-2024-ijca-923744.pdf
- A.M. Mutawa, Korupalli V. Rajesh Kumar, Hemachandran K, M. Murugappan, Using artificial intelligence to predict the next deceptive movement based on video sequence analysis: A case study on a professional cricket player's movements, Journal of Engineering Research, 2025, ISSN 2307-1877, https://doi.org/10.1016/j.jer.2025.01.007. (https://www.sciencedirect.com/science/article/pii/S2307187725000070)
- Biswas, Milon & Akhund, Tajim Md. Niamat Ullah & Mahbub, Md & Islam, Sikder Md Saiful & Sorna, Sadia & Kaiser, M. Shamim. (2021). A Survey on Predicting Player's Performance and Team Recommendation in Game of Cricket Using Machine Learning. 10.1007/978-981-16-0739-4_22. https://www.researchgate.net/publication/353477936_A_Survey_on_Predicting_Playe

r's_Performance_and_Team_Recommendation_in_Game_of_Cricket_Using_Machine_Learning

- B.Tech, M.E & S, Thameem & J, Spencer & K, Vignesh. (2023). AN INSIGHTS ON CRICKET DATA ANALYTICS. IJARCCE. 12. 10.17148/IJARCCE.2023.125247. https://www.researchgate.net/publication/371051540_AN_INSIGHTS_ON_CRICKET_DATA_ANALYTICS
- Param Dalal, Hirak Shah, Tej Kanjariya, Dhananjay Joshi . Cricket Match Analytics and Prediction using Machine Learning. International Journal of Computer Applications. 186, 26 (Jul 2024), 27-33. DOI=10.5120/ijca2024923744 https://www.ijcaonline.org/archives/volume186/number26/dalal-2024-ijca-923744.pdf
- Cricket Data Analysis Using Power BI. (n.d.). ijcrt.org. https://www.ijcrt.org/papers/IJCRT24A4413.pdf
- PREDICTING OPTIMAL CRICKET TEAM USING MACHINE LEARNING. (2024b). International Research Journal of Modernization in Engineering Technology and Science. https://doi.org/10.56726/irjmets54114
- Chakraborty, S., Mondal, A., Bhattacharjee, A., Mallick, A., Santra, R., Maity, S., & Dey, L. (2023). Cricket data analytics: Forecasting T20 match winners through machine learning. International Journal of Knowledge-based and Intelligent Engineering Systems, 28(1), 73–92. https://doi.org/10.3233/kes-230060
- Jain, V. (2024). A Study on the impact of Data Analytics on Cricket in India. International Journal of Novel Research and Development, 9(1), 190–191. https://www.ijnrd.org/papers/IJNRDTH00099.pdf
- Indika Wickramasinghe, Applications of Machine Learning in cricket: A systematic review, Machine Learning with Applications, Volume 10, 2022, 100435, ISSN 2666-8270, https://doi.org/10.1016/j.mlwa.2022.100435. (https://www.sciencedirect.com/science/article/pii/S2666827022001104)
- Prediction of the outcome of a Twenty-20 Cricket Match. (n.d.). arxiv.or. https://arxiv.org/pdf/2209.06346
- A STUDY ON MACHINE LEARNING APPROACHES FOR PLAYER PERFORMANCE AND MATCH RESULTS PREDICTION. (n.d.). arxiv.org. https://arxiv.org/pdf/2108.10125
- Goel, Rajesh & Davis, Jerryl & Bhatia, Amit & Malhotra, Pulkit & Bhardwaj, Harsh & Hooda, Vikas & Goel, Ankit. (2021). Dynamic cricket match outcome prediction. Journal of Sports Analytics. 7. 1-12. 10.3233/JSA-200510. https://www.researchgate.net/publication/353390999_Dynamic_cricket_match_outcome_prediction
- Cricket Data Analytics. (n.d.). ijirset.com. https://www.ijirset.com/upload/2024/may/334_Cricket.pdf
- Cricket Match Analytics and Prediction using Machine Learning. (n.d.). ijcaonline.org. https://www.ijcaonline.org/archives/volume186/number26/dalal-2024-ijca-923744.pdf
- Jaber, H. (2019d). Sport analytics for cricket game results using machine learning: An experimental study. www.academia.edu. https://www.academia.edu/85807511/Sport_analytics_for_cricket_game_results_using_machine_learning_An_experimental_study
- HK, P., R, P., Kumar, P., Kumar, N., & UG-Students, Department of Computer Science Engineering, Dayananda Sagar Academy of Technology and Management, Bangalore, India. (2023). Cricket score prediction using machine learning.

International Journal of Innovative Research in Technology, 9(8), 109–110. https://ijirt.org/publishedpaper/IJIRT157821_PAPER.pdf

- Kumar, N. S. P., Sai, N. P. M., Naidu, N. R. S. M. K., Rao, N. P. U. M., & Raju, N. K. R. (2023). Cricket Player Analytics using DAX. International Journal of Advanced Research in Science Communication and Technology, 338–342. https://doi.org/10.48175/ijarsct-9140
- Rajeshwarip@skasc.ac.in, Abhinavmc3@gmail.com, & Manikandanelumalai468@gmail.com. (2025). CRICKET ANALYTICS: Enhancing team performance, strategy development and player selection. In International Journal of Research Publication and Reviews (Vol. 6, Issue 4, pp. 2520–2527). https://ijrpr.com/uploads/V6ISSUE4/IJRPR41694.pdf