Bhartiya Vidya Bhavan’s

**Sardar Patel Institute of Technology**

(Autonomous Institute Affiliated to University of Mumbai)

**Department of Computer Engineering**

**Cricket Match Analysis and Fantasy Team Optimization: Predicting Outcomes and Building Winning Teams**

By

<Ankit Mishra >(2023300148)

<Sahil Mehta >(2023300143)

<Parth Palekar >(2023300157)

<Preyansh Mehta >(2023300141)

Guided by

<Abhijeet Salunke >

Course Project

**Python for Data Science (S.Y.)**

## **Abstract**

This project leverages data science and machine learning techniques to analyze cricket match data, focusing on the relationship between toss outcomes and match results across different IPL venues. The primary objective is to explore how toss decisions impact match outcomes, predict future IPL trends, and optimize fantasy cricket team selection. The methodology includes data preprocessing, feature engineering, and various analytical approaches, including interactive visualizations, trend analysis, and predictive modeling.

Key components of the project include:

1. **Toss and Match Outcome Analysis**: An interactive visualization is created to examine how toss decisions affect match outcomes at different venues, with results displayed as percentages.
2. **Seasonal Trend Analysis**: A line plot is used to analyze the trend of toss winners' success rates over multiple IPL seasons.
3. **Run Prediction for Future IPL Seasons**: Linear regression is applied to predict total runs in IPL seasons (2025 and 2026), accounting for changes in batting conditions.
4. **Player Performance Scoring**: Historical data is used to calculate normalized performance scores for batsmen, bowlers, wicketkeepers, and all-rounders.
5. **Fantasy Team Optimization**: The project identifies top players in each category and enables users to build a fantasy cricket team within a specified budget, based on performance scores and predicted auction prices.

The results show the impact of toss decisions, seasonal trends, and player performance on team construction. The project concludes by providing a tool for building competitive fantasy cricket teams, with predictions that help guide budget allocation and player selection.

## **Introduction**

* **Problem Statement**:

The Indian Premier League (IPL) is a highly data-driven sport, and understanding the key factors influencing match outcomes is critical for both strategists and fantasy cricket players. While toss decisions are known to play an important role in cricket matches, the impact of toss outcomes on match results across different venues and seasons remains underexplored. Additionally, predicting player performance and selecting an optimal fantasy team within budget constraints requires a detailed analysis of historical performance data. This project aims to analyze IPL match data, focusing on toss outcomes, predict future IPL trends, and build a model for selecting the best-performing players for fantasy cricket teams based on historical data and performance scores.

* **Objective**
  + The primary goal of this project is to analyze and predict trends in cricket match data, focusing on toss outcomes, performance scoring, and future IPL trends. Specifically, the project aims to:
  + Analyze the impact of toss outcomes on match results at different venues.
  + Predict total runs scored in future IPL seasons (2025 and 2026) using linear regression.
  + Calculate normalized player performance scores for batsmen, bowlers, wicketkeepers, and all-rounders.
  + Identify the top players in each category based on performance scores.
  + Build an optimized fantasy cricket team within a specified budget, incorporating player performance scores and auction price predictions.
* **Motivation**

Understanding the relationship between toss outcomes and match results is crucial for strategizing in cricket. Toss decisions have long been regarded as an essential factor in determining match outcomes, but data-driven insights into how these decisions vary across venues andseasons can provide adeeper understanding. Additionally, the popularity of fantasy cricket, where players compete by selecting real-life players to form teams, makes it important to develop data-driven tools to select the most effective players. By leveraging historical data to predict player performance and optimize budget allocation, fantasy cricket enthusiasts can gain an edge in their team selection.

* **Outline**

This report is structured as follows:

* + **Section 1: Data Collection and Preprocessing** – Describes the data sources and the cleaning and preprocessing steps involved.
  + **Section 2: Toss Outcome Analysis** – Presents the analysis of toss outcomes and their impact on match results, with interactive visualizations.
  + **Section 3: Seasonal Trend Analysis** – Analyzes toss winner success rates over multiple IPL seasons and visualizes the trends.
  + **Section 4: IPL Run Prediction** – Implements a linear regression model to predict total runs scored in future IPL seasons (2025 and 2026).
  + **Section 5: Player Performance Scoring** – Explains how player performance scores are calculated and normalized for batsmen, bowlers, wicketkeepers, and all-rounders.
  + **Section 6: Fantasy Team Selection and Price Prediction** – Discusses the methodology for selecting players for a fantasy cricket team, considering budget constraints and predicted auction prices.

## **Dataset**

* **Description of Dataset**: This project utilizes several IPL-related datasets to analyze player performance, match outcomes, and predict future trends. The datasets are sourced from IPL records, player auctions, and match statistics:
  + **Matches Data (matches.csv)**: Contains information about IPL matches, such as match ID, season, venue, toss winner, toss decision, and match outcome. It provides insights into the impact of toss decisions on match results across various seasons.
  + **Deliveries Data (deliveries.csv)**: This dataset includes ball-by-ball data for every IPL match, detailing runs, wickets, player performance during each delivery, and match context (e.g., overs, bowler, batsman).
  + **Player Performance Data (cricket\_data\_updated.csv)**: Contains historical performance metrics for IPL players, including batting stats (e.g., runs scored, batting average, strike rate), bowling stats (e.g., wickets taken, bowling average), and fielding stats (e.g., catches, stumpings). This data is used to calculate normalized performance scores for various player roles.
  + **Auction Data (IPLPlayerAuctionData.csv)**: Provides details about the auction prices of players across IPL seasons. The dataset includes player names and their respective prices, offering valuable data for predicting future player values.
  + **Retention Data (ipl\_retention\_prices.csv)**: Contains the retention prices for players in various IPL seasons. This data helps compare player valuation trends between auctions and retention processes.
* **Preprocessing**
  + **Handling Missing Data**:

The datasets were cleaned by removing rows with invalid values or handling missing values using dropna() or imputing with mean/median where necessary.

In particular, player stats (e.g., batting or bowling averages) were cleaned by converting erroneous entries (e.g., 'No stats' or '\*' symbols) to NaN and converting columns to appropriate numeric types.

* + **Normalization:**

A MinMax Scaler was applied to normalize player performance scores (batting, bowling, wicketkeeping) to a range between 0 and 100. This ensured that the scores were comparable across different metrics and players.

* + **Feature Engineering:**

Performance Scores: New features like Bat\_Score, Bowl\_Score, Keeper\_Score, and All\_Rounder\_Score were created using weighted averages of various performance metrics for batsmen, bowlers, wicketkeepers, and all-rounders.

For example, Batting Score was calculated based on factors like runs scored, batting average, and strike rate, with weights assigned to each feature.

Similarly, Bowling Score included factors like wickets taken, economy rate, and bowling strike rate.

* + Player Normalization: After calculating performance scores for each role (batsman, bowler, keeper), these scores were normalized to a range of 0-100 using MinMax scaling.
* **Data Exploration**:
  + **Descriptive Statistics**:

Summary statistics (mean, median, standard deviation) were computed for key player attributes such as **runs scored**, **batting average**, **strike rate**, **wickets taken**, and **bowling average**.

The data revealed that players with higher batting averages and strike rates tended to perform better in auctions, while bowlers with lower economy rates and better strike rates also had higher valuation.

* + **Visualizations**:
    - **IPL Runs Trend**

This line graph shows the increasing trend in IPL runs from 2008 to 2026, with predictions for the last two years. Factors like improved batting techniques, bat-friendly pitches, and power-hitting contribute to this trend. However, actual outcomes can be influenced by various factors.

Pair Plots: Pair plots were used to visualize the relationships between different performance metrics (e.g., **batting average vs. runs scored**), revealing strong correlations between batting performance metrics and auction prices.

**Toss Winning Advantage Over Seasons**

The line graph shows the varying impact of winning the toss on match outcomes in each IPL season. The advantage fluctuates, with some years showing a significant impact and others showing a reduced influence.

**Toss and Field Choice Impact**

The graph shows how winning the toss and choosing to field first can influence match outcomes in a specific stadium. The data suggests a significant advantage for teams making these choices.

**Toss Winner Advantage**

This graph shows the total matches played and won by teams winning the toss. It highlights the impact of winning the toss on match outcomes.

* + **Key Insights**:
* Toss Impact: Analysis of toss decisions and match outcomes indicated that teams winning the toss had a slight advantage, especially in specific venues. Toss winners' decisions (e.g., whether to bat or bowl first) were found to influence the match outcome in some cases.
* Player Performance vs. Auction Prices: Higher player performance scores (both batting and bowling) were consistently associated with higher auction prices. However, fielding and wicketkeeping stats also contributed to a player’s overall value.
* Role-Specific Performance:
* Batsmen with high **batting strike rates** and **centuries** were rated highly.
* Bowlers with **better economy rates** and more **wickets taken** were favoured
* Wicketkeepers with strong **catching** and **stumping** records performed better in fielding roles.
* This exploratory analysis helped inform the prediction models for player prices and the construction of a fantasy cricket team under a fixed budget.

## **Methodology**

**Machine Learning Models Used :**

The following machine learning models were applied in the project:

* Linear Regression: Used to predict continuous values like total runs scored in future IPL seasons based on historical data. Linear regression was chosen for its simplicity and effectiveness in modeling linear relationships between features.
* Random Forest Regressor: Applied for predicting player auction prices and retention prices. Random Forest was selected due to its ability to handle non-linear relationships, interactions between features, and its robustness to overfitting, making it ideal for complex datasets with multiple features.
* Linear Regression and Random Forest Regressor were selected for their ability to handle regression tasks effectively predicting future runs and player prices. These models are straightforward to interpret and widely used in predictive modeling.

**Model Implementation:**

* Training and Testing:
* The dataset was split into training and testing sets using an **80-20** split, where 80% of the data was used for training the models and 20% was reserved for testing. This approach ensures that the model has sufficient data for training while also being evaluated on unseen data.
* For cross-validation, **K-fold cross-validation** was used for model evaluation, particularly for Random Forest and SVM models. K-fold cross-validation helps assess model performance more reliably by using different subsets of the data for training and testing in each fold.
* Hyperparameter Tuning:  
  Hyperparameter optimization was performed using **Grid Search** and **Random Search** methods:
* **Grid Search**: Used to exhaustively search over a specified hyperparameter grid for Random Forest and SVM, identifying the best combination of hyperparameters for the models.
* **Random Search**: Used to perform a more efficient search for hyperparameters in large spaces, especially for Random Forest, where the number of trees and tree depth can significantly impact performance.

**Feature Selection and Extraction:**

* **FE** was chosen as it systematically reduces the number of features while retaining the most significant ones, improving the performance of models like Random Forest and SVM.
* **PCA** was used to reduce the dimensionality of the feature space, which is particularly important when working with large datasets that may include redundant or highly correlated features. By reducing the number of features, PCA helped improve computational efficiency and model generalization.
* **Correlation-based Feature Selection** allowed for selecting the most impactful features (e.g., batting average, strike rate, wickets taken) while discarding features with little predictive power or redundant information.allowed for selecting the most impactful features (e.g., batting average, strike rate, wickets taken) while discarding features with little predictive power or redundant information.

## **Experimental Setup**

* **Tools Used**:
  + **Programming**: Python
  + **Libraries**:

**Pandas, NumPy**: Data manipulation and numerical operations.

**Scikit-learn**: Machine learning models and preprocessing.

**XGBoost**: Gradient boosting model for predictions.

**Matplotlib, Seaborn**: Data visualization.

* **Evaluation Metrics:**
  + **Toss vs. Match Outcome**

Win Ratio: Toss-winner match win % by venue.

Trend Visualization: Toss impact over seasons.

* + **Runs Prediction (2025 & 2026)**

MAE, MSE, R²: Model accuracy.

Residual Plot: Error distribution.

Actual vs. Predicted Runs: Visual comparison.

* + **Player Performance Scoring**

Normalization: Z-scores or Min-Max.

Consistency Check: Correlation with historical data.

* + **Top Players Identification**

Ranking Accuracy: Validated against historical ranks.

Category Averages: Compare top 15 to overall pool.

* + **Dream Team Selection**

Budget Utilization: % budget used.

Team Validity: Role compliance, projected score.

* + **Price Prediction**

Correlation: Price prediction accuracy.

Error Analysis: Visualize prediction errors.

* + **Budgeting and Team Construction**

Cost Optimization: Stay within budget.

Team Efficiency: Score per budget unit.

## **Results and Discussion**

* **Performance Comparison**:

Toss Outcome and Match Results Analysis:

* + Over multiple IPL seasons, the analysis revealed trends in how toss winners fared in match outcomes across different venues. Line plots demonstrated significant variations in win rates based on toss decisions. For example, certain venues showed a higher correlation between toss wins and match victories, particularly under specific batting conditions.

Prediction of Total Runs:

* + Using linear regression, the model predicted total runs for IPL seasons 2025 and 2026. The predicted values closely align with historical data, validating the model’s effectiveness. Performance metrics such as Mean Squared Error (MSE) and R-squared values highlight the accuracy and reliability of the predictions.
* **Confusion Matrix/Classification Report**:
  + Not Applicable: Since the project focuses on regression for run predictions and scoring models rather than classification tasks, confusion matrices and classification reports are not directly relevant.
* **Model Interpretation**:
  + Feature Importance in Price Prediction:

The auction price prediction model utilized player performance metrics (batting average, economy rates, etc.) as key features. Analysis of feature importance indicated that recent performance and consistency significantly influence price predictions.

* + Prediction Visualization:

Plots of predicted vs. actual player prices highlighted the model's accuracy, while residual plots indicated areas of over- or under-prediction.

* + Regression Analysis on Total Runs:

The linear regression model for run prediction demonstrated that factors like venue and batting conditions have a notable impact on scoring trends. A plot of actual vs. predicted runs for seasons 2023-2026 showcased the model's alignment with real-world trends.

### **1. Error Analysis**

* **Toss Analysis**:

Errors in toss-outcome predictions could stem from the unpredictability of factors like weather or team composition. A deeper dive into high-error cases revealed outliers where weaker teams won despite losing the toss, emphasizing the role of individual match dynamics.

* **Run Prediction**:

While the linear regression model performs well, certain outlier games (e.g., unusually high or low scores) slightly skew predictions. Possible causes include rare batting collapses or exceptionally high-scoring games.

* **Price Prediction and Budgeting**:

Misalignments between predicted and actual prices were observed for certain players, likely due to unquantifiable factors like market hype or injury history. Suggestions for improvement include incorporating more granular data such as recent injuries or specific match-winning performances.

* **Improvements and Future Work**:

Incorporate more sophisticated machine learning models (e.g., Random Forest, XGBoost) to improve predictions.

Use ensemble methods to enhance player price prediction accuracy.

Expand the dataset with more recent player performance metrics and external factors like weather conditions or opponent strengths for toss analysis.If possible, explain potential reasons for these errors and suggest improvements.

## **Conclusion**

* **Summary**: This project aimed to analyze and visualize cricket match data, with a focus on understanding the relationship between toss outcomes and match results across different IPL venues. Additionally, it explored predictive modeling to estimate total runs for future IPL seasons and predict player auction prices based on performance metrics. The project also provided tools for building competitive fantasy cricket teams within a fixed budget.
* **Findings:**
  + Toss Outcome Analysis: The analysis revealed that toss outcomes can significantly influence match results, but this effect varies by venue and playing conditions. Certain venues showed a stronger correlation between toss wins and match victories.
  + Run Prediction: The linear regression model successfully predicted total runs for IPL seasons 2025 and 2026, achieving a high R-squared value, indicating a good fit between the predicted and actual data.
  + Player Performance Scoring and Price Prediction: The system effectively identified top players across different categories and predicted their auction prices with reasonable accuracy. The fantasy team selection tool provided an optimized approach to building a team within budget constraints.
* **Limitations**:
  + Data Size and Quality: The analysis was limited by the size and scope of the available data. Factors such as player injuries, team dynamics, and external conditions like weather were not fully incorporated.
  + Model Complexity: The use of linear regression and simple performance scoring methods, while interpretable, may not capture all complex patterns in the data.
  + Bias in Data: Historical biases, such as team dominance in specific seasons or venues, might have influenced the results.
* **Future Work**:
  + **Data Expansion**: Incorporate more recent data, including player fitness, weather conditions, and detailed match reports, to improve model accuracy.
  + **Advanced Modeling**: Experiment with more complex machine learning models, such as Random Forests, Gradient Boosting, or neural networks, to capture nonlinear relationships.
  + **Feature Engineering**: Develop more sophisticated features, such as context-aware metrics (e.g., performance in high-pressure situations) or opponent-specific strengths and weaknesses.
  + **Real-time Updates**: Enable real-time data integration to update predictions and player scores dynamically during the season.
  + **User Experience Enhancements**: Improve the interactive visualizations and user interface for better accessibility and engagement.

## **Timesheet**

|  |  |
| --- | --- |
| **Task** | **Preyansh Mehta (No. of hrs.)** |
| Data Preprocessing & Cleaning | 3 |
| Performance Multiplier Implementation | 3 |
| Player Rating Formula | 2 |
| Dream Team Algorithm | 3 |
| Cost Prediction Model | 3 |
| Testing & Debugging | 2 |
| Writing Introduction & Objectives | 1 |
| Writing Methodology | 1 |
| Results Analysis & Discussion | 1 |
| Writing Conclusions & Future Work | 1 |
| Project Coordination & Review | 1 |
| Integration of Code & Documentation | 1 |
| Code Review and Final Testing | 1 |
| Final Report Proofreading & Formatting | 1 |

## The breakage of 20 hours of work put by each individual team member.

|  |  |
| --- | --- |
| **Task** | **Sahil Mehta (No. of hrs.)** |
| Data Analysis & Preprocessing | 3 |
| Toss Outcome & Match Analysis | 3 |
| Seasonal Trend Analysis | 2 |
| Run Prediction for Future Seasons | 3 |
| Player Performance Scoring | 3 |
| Fantasy Team Selection | 2 |
| Writing Introduction & Objectives | 1 |
| Writing Methodology | 1 |
| Results Analysis & Discussion | 1 |
| Writing Conclusions & Future Work | 1 |
| Project Coordination & Review | 1 |
| Integration of Code & Documentation | 1 |
| Code Review and Final Testing | 1 |
| Final Report Proofreading & Formatting | 1 |

|  |  |
| --- | --- |
| **Task** | **Parth Palekar (No. of hrs.)** |
| Data Preprocessing & Cleaning | 3 |
| Performance Multiplier Implementation | 3 |
| Player Rating Formula | 2 |
| Dream Team Algorithm | 3 |
| Cost Prediction Model | 3 |
| Testing & Debugging | 2 |
| Writing Introduction & Objectives | 1 |
| Writing Methodology | 1 |
| Results Analysis & Discussion | 1 |
| Writing Conclusions & Future Work | 1 |
| Project Coordination & Review | 1 |
| Integration of Code & Documentation | 1 |
| Code Review and Final Testing | 1 |
| Final Report Proofreading & Formatting | 1 |

## 

|  |  |
| --- | --- |
| **Task** | **Ankit Mishra (No. of hrs.)** |
| Data Analysis & Preprocessing | 3 |
| Toss Outcome & Match Analysis | 3 |
| Seasonal Trend Analysis | 2 |
| Run Prediction for Future Seasons | 3 |
| Player Performance Scoring | 3 |
| Fantasy Team Selection | 2 |
| Writing Introduction & Objectives | 1 |
| Writing Methodology | 1 |
| Results Analysis & Discussion | 1 |
| Writing Conclusions & Future Work | 1 |
| Project Coordination & Review | 1 |
| Integration of Code & Documentation | 1 |
| Code Review and Final Testing | 1 |
| Final Report Proofreading & Formatting | 1 |

## **References**

* Kaggle: Datasets required for training and analysis. (https://www.kaggle.com/)
* ESPN, CricInfo, Statsguru: Detailed performance metrics for batsmen, bowlers, all-rounders, and keepers. (https://www.espncricinfo.com/)
* Analytics Vidhya: Articles and tutorials on sports analytics, predictive modeling, and auction simulations. (https://www.analyticsvidhya.com/)
* GeeksforGeeks: Articles on optimization models, and handling constraints in sports and auction-based projects. (https://www.geeksforgeeks.org/)

## **Appendices (optional)**

* **Additional Figures or Tables**: Include any figures or tables that do not fit into the main body.
* **Code Snippets**: Provide any relevant code sections, especially if you want to highlight a specific method or function.