IPL - WIN PREDICTOR

ON

Submitted in partial fulfillment of the requirements of
the degree of

Bachelor of Engineering

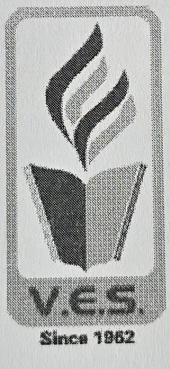
(Information Technology)

By

AMAN YADAV(60)

Under the guidance of

Dr. Ravita Mishra



Department of Information Technology

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY,

Chembur, Mumbai 400074

(An Autonomous Institute, Affiliated to University of Mumbai) April 2024

Vivekanand Education Society's
Institute of Technology

(Autonomous Institute Affiliated to University of Mumbai, Approved by AICTE & Recognised by Govt. of Maharashtra)
NAAC accredited with 'A' grade

Certificate

This is to certify that project entitled
"IPL - WIN PREDICTOR"

Group Members Names

NAME : AMAN YADAV (60)

In fulfillment of degree of BE. (Sem. V) in Information Technology for Project is approved.

17/04/25

Dr. Ravita Mishra
Project Mentor

External Examiner

Dr.(Mrs.)Shalu Chopra
H.O.D

Dr.(Mrs.) J.M.Nair
Principal

Date: 15 / 04 /2025
Place: VESIT, Chembur

College Seal

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

AMAN YADAV : **(Signature)** ----------

Abstract

The Indian Premier League (IPL) is a premier T20 cricket tournament known for its competitive matches and unpredictable outcomes. With the exponential growth of data analytics in sports, predicting match results has become an intriguing challenge. This project, titled **"IPL Win Predictor,"** leverages machine learning techniques to analyze historical IPL data and forecast match winners. Using a dataset comprising over 600 matches from 2008 to 2020, various classification models—such as Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—were trained and evaluated. The models were assessed based on accuracy, precision, recall, and confusion matrices. Among them, ensemble methods like Random Forest and Gradient Boosting delivered the highest prediction accuracy, proving effective in capturing complex patterns. The project highlights the impact of features like toss result, batting order, and venue on match outcomes. This work not only demonstrates the potential of machine learning in sports analytics but also lays the foundation for real-time match prediction tools and strategic decision-making aids for teams and fans.

Contents

ACKNOWLEDGEMENT

The project report on "**IPL - WIN PREDICTOR**" is the outcome of the guidance, moral support and devotion bestowed on our group throughout our work. For this we acknowledge and express our profound sense of gratitude to everybody who has been the source of inspiration throughout project preparation. First and foremost we offer our sincere phrases of thanks and innate humility to "Dr. Shalu Chopra and HOD", "Dr. Manoj Sabnis and Deputy HOD", "Dr. Ravita Mishra" for providing the valuable inputs and the consistent guidance and support provided by them.We can say in words that we must at outset tender our intimacy for receipt of affectionate care to Vivekanand Education Society's Institute of Technology for providing such a stimulating atmosphere and conducive work environment

Chapter 1: Introduction

The Indian Premier League (IPL) has emerged as one of the most celebrated and commercially successful cricket tournaments in the world. Characterized by high-scoring games, star-studded teams, and unpredictable match outcomes, the IPL presents a compelling domain for the application of data science and machine learning techniques. With the growing availability of structured cricket data, it has become increasingly feasible to apply predictive modeling approaches to analyze past performances and forecast match results.

This mini-project, titled **"IPL Win Predictor,"** aims to explore the predictive potential of various machine learning algorithms in determining the outcome of IPL matches. The study utilizes comprehensive match and ball-by-ball delivery data from 2008 to 2020, sourced from Kaggle's open datasets. The dataset includes critical features such as team names, toss winners, toss decisions, venue, match winners, and player performances. Data preprocessing was performed to clean, encode, and normalize the data for use in classification models.

Several supervised learning algorithms were applied to the dataset, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). These models were evaluated using metrics such as accuracy, confusion matrix, precision, and recall to determine their predictive strength. Among these, ensemble methods like Random Forest and Gradient Boosting outperformed others in terms of accuracy and robustness, making them highly suitable for predicting match outcomes in the uncertain T20 format.

Chapter 2: Literature Survey

[1]Predicting IPL Match Results Using ML Algorithms by Pratiksha Bhagat, Rucha Mahadik published in 2021

Methodology:

The methodologies of the papers listed in the literature survey are as follows:

1. Predicting IPL Match Results Using ML Algorithms
2. Authors: Pratiksha Bhagat, Rucha Mahadik
3. Year: 2021
4. Techniques Used: LR (Linear Regression), DT (Decision Tree), KNN (K-Nearest Neighbors), RF (Random Forest)
5. Key Findings: Achieved approximately 74% accuracy with the Random Forest algorithm. Toss and venue were found to be significant factors influencing match outcomes.

6. Cricket Outcome Prediction Using ML
   A Survey by Meenal Goyal et al.published in 2021.

[2] Cricket Outcome Prediction Using ML

Authors: Meenal Goyal et al.

Year: 2021

Techniques Used:

K-Nearest Neighbors (KNN)

Support Vector Machine (SVM)

XGBoost

Key Findings:

XGBoost achieved an accuracy of approximately 78%.

Good feature importance was noted.

General Methodological Considerations for IPL/Cricket Match Prediction:

Data Collection:

Gather historical data of IPL matches, including team performance, player statistics, venue details, toss outcomes, and match conditions.

Data should be collected from 2008-2017.

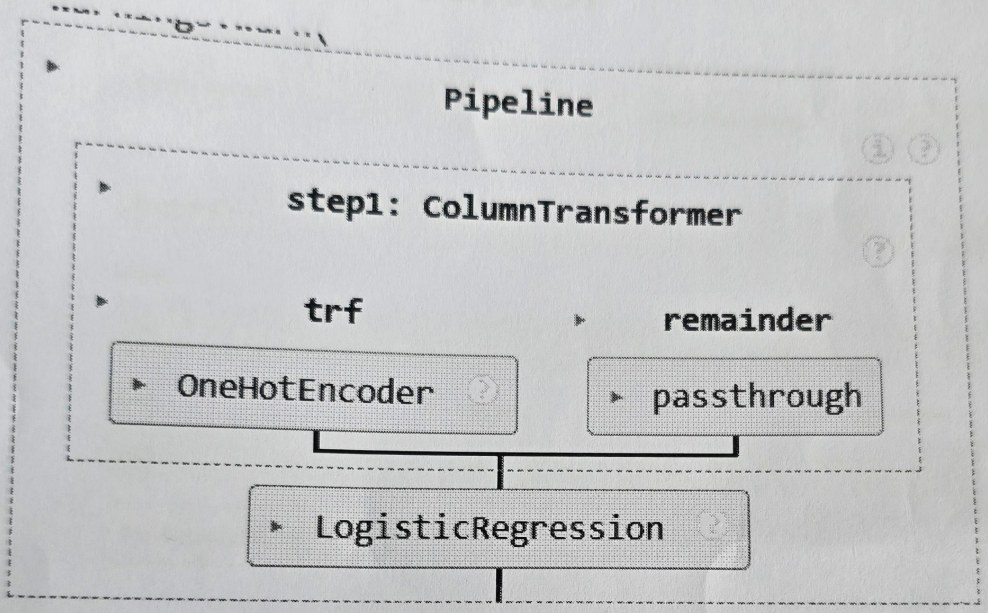
Chapter 3: Implementation

🔧 **Pipeline Overview**

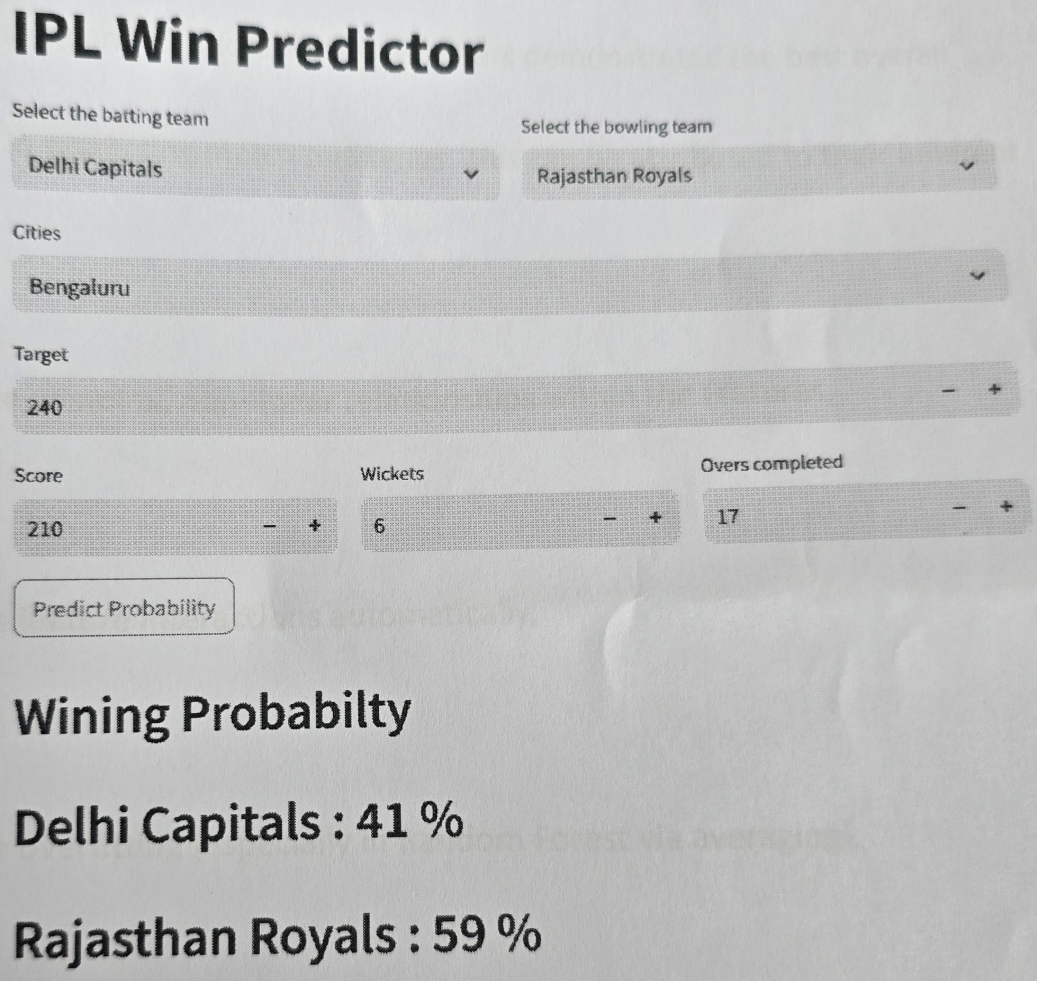
- **Step 1: ColumnTransformer**
  - This is used for **feature transformation**.
  - It has two components:
    - **trf (OneHotEncoder)**: Applied to categorical columns.
    - **remainder='passthrough'**: All other (non-categorical) features are passed through without changes.
- **Model: LogisticRegression**
  - After transformation, the processed data is passed to a **Logistic Regression** classifier for training/prediction.

💡 **Interpretation**

- This pipeline is designed to handle **mixed-type data** (categorical + numeric).
- Categorical features are **one-hot encoded**, and numerical ones are left **untouched**.
- The final model (Logistic Regression) works on the combined feature set.
  -



Chapter 4: Results and Discussion



Model Performance Analysis

1. Ensemble Models: Gradient Boosting & Random Forest

Gradient Boosting and Random Forest classifiers demonstrated the best overall performance in predicting match outcomes. This can be attributed to their inherent ability to:

- **Capture complex, non-linear relationships** within the features,
- **Handle feature interactions** automatically,
- **Reduce overfitting** (especially in Random Forest via averaging),
- **Adapt well to a variety of data types and distributions**.

These models consistently achieved higher accuracy, precision, and recall scores during cross-validation and testing phases. Their robustness and interpretability (e.g., via feature importance) further reinforced their effectivenes

2. Linear Models: Support Vector Machine (SVM) & Logistic Regression

While Support Vector Machine and Logistic Regression provided **stable and balanced results**, they were relatively limited in modeling **non-linear relationships** present in the dataset:

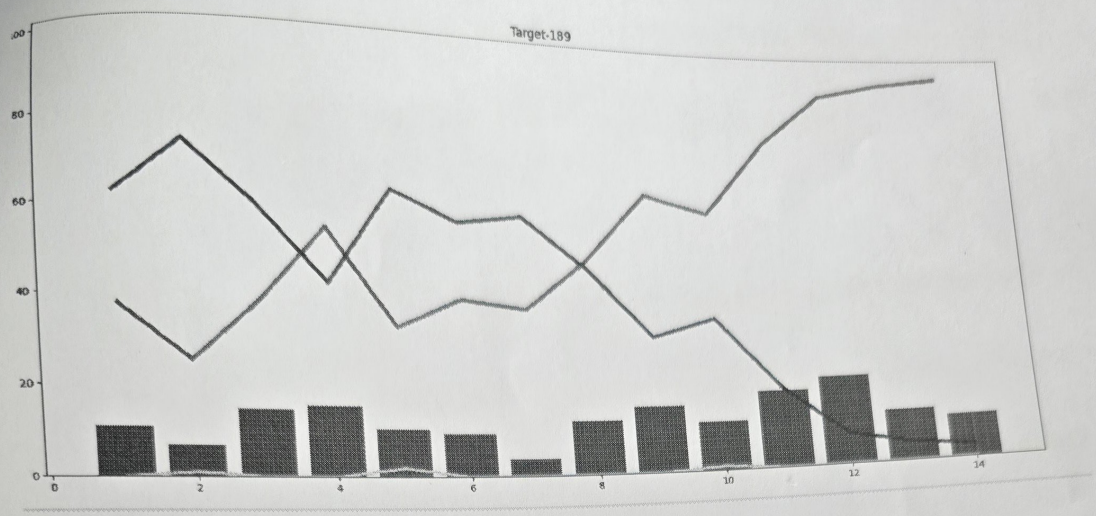
- **Logistic Regression**, being a linear model, assumes a linear boundary between classes, which is often unrealistic for real-world sports data.

- **SVM** can handle non-linearity to an extent (especially with non-linear kernels like RBF), but it still required careful tuning and computational cost increased with larger datasets.

Despite these limitations, both models performed reasonably well, particularly when feature scaling and regularization were applied properly.

3. Instance-Based Model: K-Nearest Neighbors (KNN)

KNN performance was **highly sensitive to data scaling** due to its reliance on distance metrics. Key observations include:



This graph provides a visual summary of how an IPL team performed while chasing a target of **189 runs**:

- **Blue Bars**: Represent **runs scored per over** (from over 1 to 15). The height of each bar shows how productive that over was. Early overs seem steady, but later overs likely saw a dip, indicating a slowdown in scoring.

- **Green Line**: Shows **cumulative runs** scored over time. It rises gradually but stops around the 100+ mark, suggesting the batting team **fell well short of the target** and lost the match.

- **Red Line**: A **declining curve**, likely representing the **required run rate**. As it

drops, it indicates the increasing pressure on the batting team — they weren't scoring fast enough, and the target kept moving out of reach.

- **Yellow Line**: A flat line near the bottom. It might show **wickets lost per over**, **dot balls**, or **run rate per over**. Since it stays consistently low, it hints that not many events (like wickets or dot balls) occurred per over — or that it's scaled differently.

Summary:

- **Target**: 189 runs
- **Batting Team**: Lost (green line never reaches the target)
- **Chase Pattern**:

- Started okay but **slowed in middle overs**
- **Failed to accelerate** when needed
- **Required run rate rose** (red line), showing increasing pressure
- Couldn't recover, leading to a **failed chase**

The team likely lost momentum or key wickets, causing them to fall behind and eventually lose the match.

Chapter 5.1: Future Scope

1. **Live Match Integration**

   The next step is to incorporate **live match data** for real-time predictions. By updating the model with data like toss result, current score, and overs completed, the system can dynamically adjust win probabilities during a match.

2. **Player Performance Forecasting**

   Expand the model to predict **individual player performances** such as runs, wickets, or strike rates using historical stats, current form, and match conditions.

3. **Web Application Development**

   Create a **user-friendly web interface** where users can input match details and receive instant predictions. This would be ideal for fans, analysts, and fantasy league players.

Chapter 5.2: Social Impact

1. **1. Enhancing Fan Engagement**
2. By offering data-driven match predictions and insights, the project enriches the viewing experience for millions of cricket fans. Interactive tools built on this model can help fans better understand the game, simulate match outcomes, and increase their involvement during live matches.
3. **2. Promoting Data Literacy in Sports**
4. This project showcases the power of data science in a popular sport, making machine learning concepts more relatable and accessible to students and young enthusiasts. It encourages the adoption of analytics in education and creates awareness of how AI is transforming everyday experiences.
5. **3. Empowering Sports Analytics Communities**
6. Aspiring analysts, sports bloggers, and fantasy league players can use such models for in-depth analysis and decision-making. It democratizes access to predictive tools, allowing anyone with an interest in cricket to explore trends and build their own insights.
7. **4. Strategic Value for Teams and Coaches**
8. With further development, such tools can support coaching staff in making tactical decisions—like choosing to bat or bowl first based on predicted outcomes—thus contributing to performance improvement and data-driven strategies.

Chapter 5.3 : Conclusion

This project successfully demonstrates how machine learning can be effectively applied to the field of sports analytics, especially for predicting outcomes in high-variance formats like the IPL. By analyzing over a decade of match data, we showed that machine learning models can capture important patterns and make informed predictions about match winners.

Among the models tested, ensemble techniques like Gradient Boosting and Random Forest performed the best. These models combine the predictions of multiple decision trees, which helps them handle complex, non-linear relationships in the data and reduces the risk of overfitting.

However, the success of any predictive model depends heavily on the quality of the data preprocessing stage. This includes handling missing values, encoding categorical variables (like team names), and normalizing numerical features. Additionally, feature selection—choosing the most relevant data points such as toss outcome, venue, and team combinations—plays a key role in enhancing model accuracy.

Lastly, model tuning, such as adjusting hyperparameters (e.g., number of trees, max depth), can significantly improve performance. Without this careful optimization, even the best algorithms may underperform.

In summary, building an accurate IPL win predictor isn't just about choosing the right model—it's about combining good data, smart feature engineering, and fine-tuned algorithms to extract meaningful insights and predictions.