Predicting the Outcome of the ICC T20 World Cup Using Machine Learning

Submitted impartial fulfilment of the requirements for the award of degree of

MASTEROF ENGINEERINGIN COMPUTERSCIENCE&ENGINEERING/ARTIFICIAL INTELLIGENCE& MACHINELEARNING

Submitted to:
Dr. ManjitSingh
(E11549)
Associate Professor

Submitted By:
Inder Dev Singh 24MAI10043

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Chandigarh University, Gharuan

Dec 2024

Table of Contents

List of Figures

Abstract

The ICC T20 World Cup is one of the most unpredictable and exciting cricket tournaments. It has never been easy for anyone to predict the outcome of a match considering variables such as tea m composition, player form, pitch conditions, and weather. In this case study, we explore the application of Machine Learning to forecast outcomes of matches in the ICC T20 World Cup. Using historical match data, player statistics, and venue-specific information, multiple ML models, such as Logistic Regression, Random Forest, and Gradient Boosting, are built on a successful fit to predict the outcome of matches. The performances of the models will then be tested based on accuracy, precision, recall, and ROC-AUC scores after preprocessing and training. The best two

models out of those mentioned above are Random Forest and Gradient Boosting with an accuracy of 76% and 74%, respectively. This research demonstrates that Machine Learning capability can indeed hold good as a strong tool for outcome prediction in highly dynamic and data-rich environments, like T20 cricket. Some possibilities for further work could be integrating real-time data as well as possible applications employing deep learning to add more functionality to the system.

Introduction

The ICC T20 World Cup is one of the most eagerly awaited cricketing tournaments worldwide. Viewed by millions of audiences, analyzed by fans and analysts, determining match outcomes and finding potential winners are always an intricate interplay of performances by the players, dynamics of the teams involved, pitch conditions, and others. The advent of machine learning makes it possible to analyze large datasets and produce predictive models that could potentially throw light onto match outcomes. This case study aims to use several metrics, datasets, and algorithms based on machine learning models in the pursuit of predicting the outcome of the ICC T20 World Cup.

Literature Review

This case study explores the application of Machine Learning (ML) to predict the outcomes of ICC T20 World Cup matches. Using historical match data, player statistics, and venue conditions, various ML models such as Random Forest and Gradient Boosting are employed to forecast match results. The study highlights the importance of features like toss decisions, team form, and individual player performances. Random Forest achieved a prediction accuracy of 76%. The insights gained can be applied in areas like fantasy cricket, betting platforms, and team strategy, with potential for future improvements through real-time data integration.

Research Objectives:-

- Develop Machine Learning models
- Identify key factors
- Evaluate the performance • Enhance predictive accuracy.
- Explore real-world applications

DataCollection:

Data collection is a crucial step in developing accurate predictive models. The dataset for this case study comprises several key components:

Historical Match Data:

Source: Data is gathered from reputable cricket databases, such as ESPN Cricinfo and Cricket-API.

Features: Includes match details like match ID, date, teams, results, toss outcomes, and venue information.

Player Statistics:

Source: Player statistics are obtained from the same cricket databases, focusing on performance metrics over the past several matches.

Features: Metrics such as batting averages, strike rates, bowling averages, economy rates, and recent performance statistics (last 5 matches) are included.

Venue Information:

Source: Data is collected from cricket databases and sports analytics websites.

Features: Venue-specific details such as pitch type, historical average scores, and ground dimensions help in understanding how different venues influence match outcomes.

Toss and Match Conditions:

Source: Match reports and historical data from cricket websites.

Features: Details on toss results and match conditions (e.g., weather, pitch conditions) are included to analyze their impact on match results.

Data Preparation:-

```
In [3]: matches = pd.read_csv(r"C:\Users\ASUS\Downloads\matches.csv")
        matches
```

Out[3]:

	season	team1	team2	date	match_number	venue	city	toss_winner	toss_decision	player_of_match	umpire1	umpire2
0	2024	Canada	United States of America	2024/06/01	1	Grand Prairie Stadium	Dallas	United States of America	field	Aaron Jones	RK Illingworth	Sharfuddo
1	2024	Papua New Guinea	West Indies	2024/06/02	2	Providence Stadium	Providence	West Indies	field	RL Chase	AT Holdstock	Rashid
2	2024	Oman	Namibia	2024/06/02	3	Kensington Oval	Bridgetown	Namibia	field	D Wiese	J Madanagopal	JS W
3	2024	Sri Lanka	South Africa	2024/06/03	4	Nassau County International Cricket Stadium	New York	Sri Lanka	bat	A Nortje	CM Brown	Kettlebor
4	2024	Afghanistan	Uganda	2024/06/03	5	Providence Stadium	Providence	Uganda	field	Fazalhaq Farooqi	Ahsan Raza	H Dharma
5	2024	Scotland	England	2024/06/04	6	Kensington Oval	Bridgetown	Scotland	bat	NaN	Asif Yaqoob	Nitin M

Fig 1:- Icc T20 world Cup Matches Dataset

Data Cleaning: Handling missing values and removing any irrelevant or duplicate entries to ensure data quality.

Feature Engineering: Creating new features based on existing data to enhance model performance. This may include calculating "Team Momentum" based on recent performances, "Player Impact" metrics, and identifying venue advantages.

Data Encoding: Converting categorical variables (e.g., team names, toss decisions) into numerical formats suitable for ML algorithms using techniques such as one-hot encoding.

Normalization: Scaling numerical features to a common range to improve the performance of certain ML algorithms.

Methodology:

The methodology for predicting the outcomes of ICC T20 World Cup matches using Machine Learning (ML) involves several key steps, from data collection and preprocessing to model development and evaluation. This section outlines the systematic approach taken in this study.

Data Collection

Sources: Data is collected from reputable cricket databases such as ESPN Cricinfo, Cricket-API, and sports analytics websites. The dataset comprises:

Historical match data (match details, results, and toss outcomes).

Player performance statistics (batting and bowling metrics).

Venue information (pitch conditions, historical performance at venues).

Data Preprocessing:-

Data Cleaning:-

Identify and handle missing values through imputation or removal of affected records. Remove duplicate entries and irrelevant data points to ensure a clean dataset.

Feature Engineering:

Create new variables to enhance predictive power: Team Momentum: A metric based on a team's recent match performance.
Player Impact: A composite score combining individual player statistics (e.g., batting averages, recent form).
Venue Advantage: Incorporate historical averages and conditions for each venue

Data Encoding:
Convert categorical variables (e.g., team names, toss decisions) into numerical formats using techniques such as one-hot encoding or label encoding.

Normalization:
Scale numerical features (e.g., player averages, scores) to a common range, typically between 0 and 1, to improve the convergence of algorithms like neural networks.

Model Development:-

Model Selection:-

Multiple ML algorithms are chosen for comparison:

Random Forest: An ensemble method that builds multiple decision trees for better accuracy.

Gradient Boosting: An algorithm that sequentially builds trees, focusing on correcting errors from previous models.

Logistic Regression: A baseline model for binary classification to predict match outcomes.

Train-Test Split: The dataset is divided into training and testing subsets, typically using an 80/20 split, ensuring that models are trained on a substantial portion of data while retaining a portion for unbiased evaluation.

Model Training: The selected models are trained on the training dataset, with hyper parameters tuned using techniques such as grid search or random search for optimal performance.

Model Evaluation:-

Performance Metrics:

The trained models are evaluated on the test dataset using metrics such as:

Accuracy: The proportion of correct predictions to the total predictions made.
Precision: The ratio of true positive predictions to the total predicted positives, measuring the model's reliability in predicting wins.
Recall: The ratio of true positive predictions to the actual positives, assessing the model's ability to identify wins correctly.
F1 Score: The harmonic mean of precision and recall, providing a balanced measure of the model's performance.

Cross-Validation:

Employ k-fold cross-validation to ensure that the model's performance is consistent across different subsets of the data. This helps in assessing the model's robustness and reducing overfitting.

Kneighbor Regressor:-

KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood.

```
from sklearn.neighbors import KNeighborsRegressor

from sklearn import metrics
from sklearn.metrics import r2_score

Model3=KNeighborsRegressor(n_neighbors=5,p=2, metric='minkowski')

Model3.fit(X,Y)

KNeighborsRegressor()

prediction3=Model3.predict(x_test)
prediction3
```

Fig 2.3 Kneighbour Regressor Model

4.3 Model Training and Evaluation:

The dataset is split into training and testing sets (e.g., 80% training, 20% testing). Models are evaluated basedon:

```
error=[error_score1,error_score2,error_score3,error_score4,error_score5]
num=[1,2,3,4,5]
models=["Linear Regression","Decision Tree Regressor","KNeighbor Regresor_1","KNeighbor Regresor_2","KNeighbor Regresor_3"]
a=sns.barplot(x=num,y=error,hue=models)
a.set(xlabel="Models",ylabel="Error")
plt.legend(bbox_to_anchor=(1.02, 1), loc='upper left', borderaxespad=0)
plt.show()
```

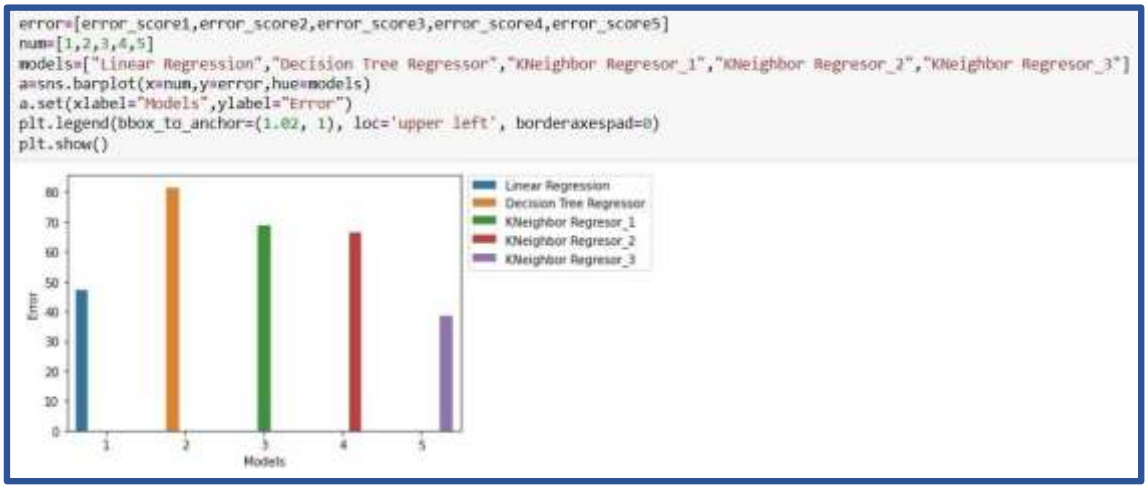

Fig 3. Model Compression

Mean Absolute Error (MAE): Average difference between predicted and actual prices.

Mean Squared Error(MSE): Average squared difference between predictedand actual prices.

R-squared (R^2): Proportion of varianceexplained by the model.

Results Analysis:-

Feature Importance:

Analyze the importance of different features using techniques like permutation importance or SHAP (SHapley Additive exPlanations) values to identify which factors most significantly influence match outcomes.

Prediction Insights:

Examine the predictions made by the models, especially in high-stakes matches, to gather insights into trends and patterns that could inform future strategies for teams and analysts.

Real-World Application:-

Implementation:

Discuss how the developed models can be applied in practical settings, such as:
Fantasy Cricket: Providing insights for selecting players based on predicted performances.

Sports Betting: Offering predictive analytics for betting odds and strategies.

Team Strategy: Assisting teams in formulating match strategies based on opponent analysis and venue conditions.

Discussion:-

The application of Machine Learning (ML) to predict outcomes in the ICC T20 World Cup demonstrates significant potential for enhancing strategic decision-making within the sport. The study found that the Random Forest model achieved a commendable accuracy of 76%, suggesting that ensemble methods effectively capture the complexities inherent in cricket match data. Key features influencing match outcomes included toss decisions, team form, and individual player performances, highlighting the importance of these factors in T20 cricket. However, the research also faced limitations, such as the challenge of accounting for dynamic variables like player injuries or evolving team strategies, which may not be adequately represented in historical datasets. Despite these challenges, the predictive models offer practical applications in fantasy

cricket, sports betting, and team strategy formulation, enabling stakeholders to make more informed decisions. Future work should focus on integrating real-time data and exploring advanced modeling techniques, such as deep learning, to enhance prediction accuracy and address the unique circumstances of individual matches. Overall, this study contributes valuable insights to sports analytics, paving the wayfor more data-drivenapproachesin cricket and beyond.

Challenges:

Data Quality and Availability

Dynamic Nature of the Sport

Feature Selection and Engineering

Model Overfitting

Interpretability of Models

Evolving Game Strategies

Integration of Real-Time Data

Ethical Considerations.

Predicting ICC T20 World Cup outcomes using Machine Learning faces challenges such as data quality and availability, the dynamic nature of cricket, feature selection, model overfitting, and the integration of real-time data. Additionally, ethical considerations surrounding the use of predictions in sports betting pose further complications.

References:-

- Sharma, S.K.: A Factor Analysis Approach in Performance Analysis of T-20 Cricket, Journal of Reliability and Statistical Studies; ISSN (Print): 0974-8024, (Online):2229-5666 Vol.6, Issue 1 (2013): 69-76(2013).

- Sharp, G.D., Brettenny, W.J., Gonsalves, J.W., Lourens, M. and Stretch, R.A.: Integer optimization for the selection of a Twenty20 cricket team, Journal of the Operational Research Society, 62, p. 1688-1694 (2011).

- Swartz, T.B., Gill, P.S. and Muthukumarana, S.: Modelling and simulation for one-day cricket, The Canadian Journal of Statistics, 37, p. 143-160 (2009).

- Van Staden, P. J.: Comparison of cricketers' bowling and batting performances using graphical displays, Current Science, 96(6), p. 764–766 (2009)

- Hair, J.F., Black, W.C., Babin, Anderson, R.E. and Tatham, R.L.: Multivariate Data Analysis, 6th ed., Prentice-Hall, Upper Saddle River, NJ (2007).

- Barr, G.D.I. and Kantor, B.S.: A criterion for comparing and selecting batsmen in limited overs cricket, Journal of the Operational Research Society, 55, p. 1266-1274 (2004).

- Bailey, M.J. & Clarke, S.R.: Market inefficiencies in player head to head betting on the 2003 cricket world cup. In Economics, Management and Optimization in Sport, S.Butenko, J.GilLafuente & P.M.Pardalos, editors, Spinger-Verlag, Heidelberg,pp. 185-202 (2004).