

**Department of Electrical and Computer Engineering**

**North South University**

**Senior Design Project-I (CSE499A)**

**Forecasting Predictive Outcomes of the T20 International Cricket and T20 Club Cricket Competitions Matches Using Machine-Learning Django Flutter**

|  |  |
| --- | --- |
| **Md. Taufiq Al Hasib Sadi** | **ID# 1311370042** |
| **Gigawashu Das** | **ID# 1421310042** |
| **Md. Ekramul Haq** | **ID# 1921740642** |
| **Israt Jerin** | **ID# 1631893042** |

**Faculty Advisor:**

**Dr. Mohammad Monirujjaman Khan**

**Associate Professor**

**ECE Department**

**Spring, 2023**

# LETTER OF TRANSMITTAL

17 June, 2023

To

Dr. Rajesh Palit

Chairman,

Department of Electrical and Computer Engineering

North South University, Dhaka

Subject: **Submission of Capstone Project Report on “Forecasting Predictive Outcomes of the T20 International Cricket and T20 Club Cricket Competitions Matches Using Machine-Learning Django Flutter”**

Dear Sir,

With due respect, we would like to submit our **Capstone Project Report** on ***“Forecasting Predictive Outcomes of the T20 International Cricket and T20 Club Cricket Competition Matches Using Machine-Learning Django Flutter”*** as a part of our BSc program. The report deals with machine learning-based phone and web application creation; which is developed by using Django and Flutter for the Men's and Women's T20 International Cricket and T20 Club Cricket Competitions to forecast likely scores and victory likelihood. This research was very valuable to us as it helped us gain experience in practical fields and apply it in real life. We tried to the maximum competence to meet all the dimensions required from this report.

We will be highly obliged if you kindly receive this report and provide your valuable judgment. It would be our immense pleasure if you find this report useful and informative and have an apparent perspective on the issue.

Sincerely Yours,

.........................................................

Md. Taufiq Al Hasib Sadi

ECE Department

North South University, Bangladesh

........................................................

Gigawashu Das

ECE Department

North South University, Bangladesh

.........................................................

Md. Ekramul Haq

ECE Department

North South University, Bangladesh

.........................................................

Israt Jerin

ECE Department

North South University, Bangladesh

# APPROVAL

Md. Taufiq Al Hasib Sadi (ID # 1311370042), Gigawashu Das (ID # 1421310042), Md.Ekramul Haq (ID # 1921740642), and Israt Jerin (ID # 1631893042) from the Electrical and Computer Engineering Department of North South University have worked on the Senior Design Project titled *“****Forecasting Predictive Outcomes of the T20 International Cricket and T20 Club Cricket Competitions Matches Using Machine-Learning Django Flutter****”* under the supervision of Dr. Mohammad Monirujjaman Khan in partial fulfillment of the requirement for the degree of Bachelors of Science in Engineering and has been accepted as satisfactory.

**Supervisor’s Signature**

…………………………………….

**Dr. Mohammad Monirujjaman Khan**

**Associate Professor**

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh.

**Chairman’s Signature**

…………………………………….

**Dr. Rajesh Palit**

**Professor**

Department of Electrical and Computer Engineering

North South University

Dhaka, Bangladesh

# DECLARATION

This is to declare that this project is our original work. No part of this work has been submitted elsewhere partially or fully for the award of any other degree or diploma. All project related information will remain confidential and shall not be disclosed without the formal consent of the project supervisor. Relevant previous works presented in this report have been properly acknowledged and cited. The plagiarism policy, as stated by the supervisor, has been maintained.

Students’ names & Signatures

\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

**1. Md. Taufiq Al Hasib Sadi**

\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

**2. Gigawashu Das**

\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

**3. Md. Ekramul Haq**

\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

**4. Israt Jerin**

# ACKNOWLEDGEMENT

First and foremost, praise to the Almighty Allah for showers of blessings throughout our research work to complete the thesis successfully.

We would like to express our deepest gratitude to our honorable research supervisor, Dr. Mohammad Monirujjaman Khan, Associate Professor, Department of Electrical and Computer Engineering, North South University, for guiding us to accomplish this challenging task. This research would not have been possible without his insightful suggestions, constant guidance, astute criticism, unimaginable care, and incredible patience. His vision and motivation have deeply inspired us. He has taught us the methodology to carry out the work and to present the thesis work as carefully as possible. It was a privilege and honor to have him as our supervisor and mentor.

# ABSTRACT

**Forecasting Predictive Outcomes of the T20 International Cricket and T20 Club Cricket Competitions Matches Using Machine-Learning Django Flutter**

Cricket is a game of uncertainty where two teams compete against each other; utilize machine-learning techniques, gather relevant historical match data and analyze those data to generate match score and outcomes for both T20 International Cricket and T20 Club Cricket competitions is a revolutionary study of modern days. Consequently, in this study, a web application and a phone application are created to perform predictive analysis on live T20 International Cricket and some popular T20 Club Cricket competitions matches in order to forecast the final score and the winner of the game of live match and also before the match begins. To examine match data and train the model for prediction, Xgboost Regression, Random Forest Regression and Logistic Regression has been utilized. These predictions are made using algorithms such as Xgboost Regressor, Random Forest Regressor and Logistic Regression, while the web application is built with Django Web Framework, APIs are built with Django Rest Framework and phone application is built with Flutter Framework; deployed with google play store and web host server. In case of score prediction for T20 International Men's from other tournaments, Xgboost Regressor has best R2 Score and lower Mean Absolute Error which is 0.98 and 2.88; in case of first innings win for T20 International Women's from other tournaments, Logistic Regression has best accuracy score which is 87% and in case of second innings win for T20 International Women's from other tournaments, Logistic Regression has best accuracy score which is 90%.

TABLE OF CONTENTS

[LETTER OF TRANSMITTAL i](#_Toc138467246)

[APPROVAL iii](#_Toc138467247)

[DECLARATION iv](#_Toc138467248)

[ACKNOWLEDGEMENT v](#_Toc138467249)

[ABSTRACT vi](#_Toc138467250)

[LIST OF FIGURES ix](#_Toc138467251)

[LIST OF TABLES xi](#_Toc138467252)

[**Chapter 1 Introduction** 1](#_Toc138467253)

[1.1 Background and Motivation 1](#_Toc138467254)

[1.2 Purpose and Goal of the Project 2](#_Toc138467255)

[1.3 Organization of the Report 2](#_Toc138467256)

[**Chapter 2 Research Literature Review** 4](#_Toc138467257)

[2.1 Existing Research and Limitations 4](#_Toc138467258)

[**Chapter 3 Methodology** 9](#_Toc138467259)

[3.1 System Design 9](#_Toc138467260)

[3.1.1 Proposed Methodology of Complete Research 9](#_Toc138467261)

[3.1.2 Proposed Methodology of Real World Applications 10](#_Toc138467262)

[3.1.3 Database Design and Entity Relationship Diagram 12](#_Toc138467263)

[3.2 Hardware and/or Software Components 12](#_Toc138467264)

[3.2.1 Data Collection and Description of Collected Data 12](#_Toc138467265)

[3.2.2 Data Exploration 13](#_Toc138467272)

[3.2.3 Data Pre-processing and Feature Data Extraction 13](#_Toc138467273)

[3.2.4 Feature Data 15](#_Toc138467274)

[3.2.5 Machine Learning Models Building 30](#_Toc138467275)

[3.2.6 Train Models 32](#_Toc138467276)

[3.2.7 Machine Learning Model Evaluation 33](#_Toc138467277)

[3.2.8 Required Tools 35](#_Toc138467278)

[**Chapter 4 Result, Analysis and Discussion** 37](#_Toc138467279)

[4.1 First Innings Score Prediction Models Result Analysis 37](#_Toc138467280)

[4.2 First Innings Win Prediction Models Result Analysis 48](#_Toc138467281)

[4.3 Second Innings Win Prediction Models Result Analysis 59](#_Toc138467282)

[**Chapter 5 Impacts of the Project** 72](#_Toc138467283)

[5.1 Impact of this project on societal, health, safety, legal and cultural issues 72](#_Toc138467284)

[5.1.1 Societal Impact 72](#_Toc138467285)

[5.1.2 Health Impact 72](#_Toc138467286)

[5.1.3 Safety Impact 72](#_Toc138467287)

[5.1.4 Legal Impact 72](#_Toc138467288)

[5.1.5 Cultural Impact 73](#_Toc138467289)

[5.2 Impact of this project on environmental and sustainability 73](#_Toc138467290)

[**Chapter 6 Project Planning and Budget** 74](#_Toc138467291)

[6.1 Gantt Chart 74](#_Toc138467292)

[6.2 Budget 74](#_Toc138467293)

[6.2.1 Hardware 74](#_Toc138467294)

[6.2.2 Software 74](#_Toc138467304)

[6.2.3 Data 75](#_Toc138467305)

[6.2.4 Hosting and Deployment 75](#_Toc138467306)

[**Chapter 7 Complex Engineering Problems and Activities** 76](#_Toc138467307)

[7.1 Complex Engineering Problems (CEP) 76](#_Toc138467308)

[7.2 Complex Engineering Activities (CEA) 77](#_Toc138467309)

[**Chapter 8 Conclusion** 79](#_Toc138467310)

[8.1 Summary 79](#_Toc138467311)

[8.2 Limitations 79](#_Toc138467312)

[8.3 Future Improvement 81](#_Toc138467313)

[**References** 83](#_Toc138467314)

# LIST OF FIGURES

[Figure 1. Workflow of the research 9](#_Toc138466233)

[Figure 2. Architecture of Real World Application 10](#_Toc138466234)

[Figure 3. Entity Relationship (ER) diagram of database 11](#_Toc138466235)

[Figure 4. A snippet of the BBL score prediction model feature dataset 14](#_Toc138466236)

[Figure 5. A snippet of the BBL first innings win prediction model feature dataset 15](#_Toc138466237)

[Figure 6. A snippet of the BBL second innings win prediction model feature dataset 15](#_Toc138466238)

[Figure 7. A snippet of the BPL first innings score prediction model feature dataset 16](#_Toc138466239)

[Figure 8. A snippet of the BPL first innings win prediction model feature dataset 16](#_Toc138466240)

[Figure 9. A snippet of the BPL second innings win prediction model feature dataset 16](#_Toc138466241)

[Figure 10. A snippet of the CPL first innings score prediction model feature dataset 17](#_Toc138466242)

[Figure 11. A snippet of the CPL first innings win prediction model feature dataset 17](#_Toc138466243)

[Figure 12. A snippet of the CPL second innings win prediction model feature dataset 18](#_Toc138466244)

[Figure 13. A snippet of the CSA T20 first innings score prediction model feature dataset 18](#_Toc138466245)

[Figure 14. A snippet of the CSA T20 first innings win prediction model feature dataset 19](#_Toc138466246)

[Figure 15. A snippet of the CSA T20 second innings win prediction model feature dataset 19](#_Toc138466247)

[Figure 16. A snippet of the IPL first innings score prediction model feature dataset 20](#_Toc138466248)

[Figure 17. A snippet of the IPL first innings win prediction model feature dataset 20](#_Toc138466249)

[Figure 18. A snippet of the IPL second innings win prediction model feature dataset 20](#_Toc138466250)

[Figure 19. A snippet of the LPL first innings score prediction model feature dataset 21](#_Toc138466251)

[Figure 20. A snippet of the LPL first innings win prediction model feature dataset 21](#_Toc138466252)

[Figure 21. A snippet of the LPL second innings win prediction model feature dataset 22](#_Toc138466253)

[Figure 22. A snippet of the PSL first innings score prediction model feature dataset 22](#_Toc138466254)

[Figure 23. A snippet of the PSL first innings win prediction model feature dataset 23](#_Toc138466255)

[Figure 24. A snippet of the PSL second innings win prediction model feature dataset 23](#_Toc138466256)

[Figure 25. A snippet of the T20 Blast first innings score prediction model feature dataset 24](#_Toc138466257)

[Figure 26. A snippet of the T20 Blast first innings win prediction model feature dataset 24](#_Toc138466258)

[Figure 27. A snippet of the T20 Blast second innings win prediction model feature dataset 24](#_Toc138466259)

[Figure 28. A snippet of the Men’s T20i first innings score prediction model feature dataset 25](#_Toc138466260)

[Figure 29. A snippet of the Men’s T20i first innings win prediction model feature dataset 25](#_Toc138466261)

[Figure 30. A snippet of the Men’s T20i second innings win prediction model feature dataset 26](#_Toc138466262)

[Figure 31. A snippet of the Women’s T20i first innings score prediction model feature dataset 26](#_Toc138466263)

[Figure 32. A snippet of the Women’s T20i first innings win prediction model feature dataset 27](#_Toc138466264)

[Figure 33. A snippet of the Women’s T20i second innings win prediction model feature dataset 27](#_Toc138466265)

[Figure 34. A snippet of the WBBL first innings score prediction model feature dataset 28](#_Toc138466266)

[Figure 35. A snippet of the WBBL first innings win prediction model feature dataset 28](#_Toc138466267)

[Figure 36. A snippet of the WBBL second innings win prediction model feature dataset 28](#_Toc138466268)

[Figure 37. Illustration of gradient tree boosting 30](#_Toc138466269)

[Figure 38. Sigmoid function in graphic view 31](#_Toc138466270)

[Figure 39. Confusing matrix structure 33](#_Toc138466271)

[Figure 40. Predicted Values vs Accurate Values Graph of BBL score prediction models 36](#_Toc138466272)

[Figure 41. Predicted Values vs Accurate Values Graph of BPL score prediction models 37](#_Toc138466273)

[Figure 42. Predicted Values vs Accurate Values Graph of CPL score prediction models 37](#_Toc138466274)

[Figure 43. Predicted Values vs Accurate Values Graph of CSA T20 score prediction models 38](#_Toc138466275)

[Figure 44. Predicted Values vs Accurate Values Graph of IPL score prediction models 39](#_Toc138466276)

[Figure 45. Predicted Values vs Accurate Values Graph of LPL score prediction models 39](#_Toc138466277)

[Figure 46. Predicted Values vs Accurate Values Graph of PSL score prediction models 40](#_Toc138466278)

[Figure 47. Predicted Values vs Accurate Values Graph of T20 Blast score prediction models 41](#_Toc138466279)

[Figure 48. Predicted Values vs Accurate Values Graph of Men’s T20i score prediction models 41](#_Toc138466280)

[Figure 49. Predicted Values vs Accurate Values Graph of Women’s T20i score prediction models 42](#_Toc138466281)

[Figure 50. Predicted Values vs Accurate Values Graph of WBBL score prediction models 43](#_Toc138466282)

[Figure 51. BBL first innings win prediction model confusion matrix 44](#_Toc138466283)

[Figure 52. BPL first innings win prediction model confusion matrix 45](#_Toc138466284)

[Figure 53. CPL first innings win prediction model confusion matrix 46](#_Toc138466285)

[Figure 54. CSA T20 first innings win prediction model confusion matrix 47](#_Toc138466286)

[Figure 55. IPL first innings win prediction model confusion matrix 48](#_Toc138466287)

[Figure 56. LPL first innings win prediction model confusion matrix 49](#_Toc138466288)

[Figure 57. PSL first innings win prediction model confusion matrix 50](#_Toc138466289)

[Figure 58. T20 Blast first innings win prediction model confusion matrix 51](#_Toc138466290)

[Figure 59. Men’s T20i first innings win prediction model confusion matrix 52](#_Toc138466291)

[Figure 60. Women’s T20i first innings win prediction model confusion matrix 53](#_Toc138466292)

[Figure 61. WBBL first innings win prediction model confusion matrix 54](#_Toc138466293)

[Figure 62. BBL second innings win prediction model confusion matrix 55](#_Toc138466294)

[Figure 63. BPL second innings win prediction model confusion matrix 56](#_Toc138466295)

[Figure 64. CPL second innings win prediction model confusion matrix 57](#_Toc138466296)

[Figure 65. CSA T20 second innings win prediction model confusion matrix 58](#_Toc138466297)

[Figure 66. IPL second innings win prediction model confusion matrix 59](#_Toc138466298)

[Figure 67. LPL second innings win prediction model confusion matrix 60](#_Toc138466299)

[Figure 68. PSL second innings win prediction model confusion matrix 61](#_Toc138466300)

[Figure 69. T20 Blast second innings win prediction model confusion matrix 62](#_Toc138466301)

[Figure 70. Men’s T20i second innings win prediction model confusion matrix 63](#_Toc138466302)

[Figure 71. Women’s T20i second innings win prediction model confusion matrix 64](#_Toc138466303)

[Figure 72. WBBL second innings win prediction model confusion matrix 65](#_Toc138466304)

[Figure 73. Gantt Chart. 68](#_Toc138466305)

# LIST OF TABLES

[Table I. List of Software Tools 34](#_Toc138466344)

[Table II. Evaluation table of BBL score prediction models 36](#_Toc138466345)

[Table III. Evaluation table of BBL score prediction models 37](#_Toc138466346)

[Table IV. Evaluation table of CPL score prediction models 38](#_Toc138466347)

[Table V. Evaluation table of CSA T20 score prediction models 38](#_Toc138466348)

[Table VI. Evaluation table of IPL score prediction models 39](#_Toc138466349)

[Table VII. Evaluation table of LPL score prediction models 39](#_Toc138466350)

[Table VIII. Evaluation table of PSL score prediction models 40](#_Toc138466351)

[Table IX. Evaluation table of T20 Blast score prediction models 41](#_Toc138466352)

[Table X. Evaluation table of Men’s T20i score prediction models 42](#_Toc138466353)

[Table XI. Evaluation table of Women’s T20i score prediction models 42](#_Toc138466354)

[Table XII. Evaluation table of WBBL score prediction models 43](#_Toc138466355)

[Table XIII. Accuracy Score of first innings win model of all tournaments 43](#_Toc138466356)

[Table XIV. Accuracy Score of Second innings win model of all tournaments 54](#_Toc138466357)

[Table XV. Complex Engineering Problem Attributes 70](#_Toc138466358)

[Table XVI. Complex Engineering Activities 71](#_Toc138466359)

# **Chapter 1 Introduction**

## Background and Motivation

Cricket is a popular sport that people worldwide enjoy. The first reference to cricket being played as an adult sport was in 1611, and in the same year, a dictionary defined cricket as a boys' game in London and the southeastern countries of England. After that, in 1745, women's cricket was added to the dictionary. This game spreads daily to Asia, North America, South America, and Australia. There are various formats of cricket, like Test cricket, known as a five or four day game; ODI (one-day international) cricket, known as the fifty-over game; T-20 cricket, which is twenty overs [1]; T-10 cricket, which is ten overs; and The Hundreds, which is a hundred balls game. Due to reduced playtime and speedy decisions made, the T-20 variant of this game has become extremely popular. More than seventy-nine nations are currently participating in the T20-International and Non-International games [2]. Due to the popularity of the International T-20 Cricket across cricketing nations, many countries have started their own leagues where crowds of people attend the games and millions more watch on television. Like IPL in India, Big Bash League and Women's Big Bash League in Australia, T20 Blast and Women's Cricket Super League in England, PSL in Pakistan, CPL in the West Indies, BPL in Bangladesh, CSA T20 Challenge in South Africa, Super Smash in New Zealand, and LPL in Sri Lanka. All of the leagues are very popular commercial leagues [3]; those have been considered in this research.

As a result, machine learning is a subfield of artificial intelligence (AI) and data science that uses collected data and algorithms to mimic how people learn, progressively improving its accuracy [4]; in the areas of Twenty20 (T20) cricket, especially IPL and PSL cricket game analytics, machine learning was more applicable and useful due to the availability of historical datasets [7, 8, 9, 10, 11, 12, 13]. The success of a team and its result against another team can be  forecasted before effectively through the machine learning algorithms; in case of score prediction Linear Regression, Rigid Regression, Naive Bayes, Random Forest, Multi-Class Support Vector Machine (SVM), and Decision Trees had presented the result through the algorithms previously [8, 18, 19]. In case win prediction Naive Bayes, Extreme Gradient Boosting, SVM, Logistic Regression, Random Forests and Multiple Linear Regression had presented the result through the algorithms previously [6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 20]. Using machine learning technology in phone and web applications is more revolutionary because it has provided companies with insight into patterns in customer behavior and business strategies. The world's top businesses, including Facebook, Netflix, Google, Amazon, Netflix, Oval Money, Uber, Snapchat Filters, and Tinder, are major examples of this [5].

Get inspired by the massive popularity of T20 cricket and the millions of devoted fans worldwide. Use machine learning to improve decision-making and acquire useful insights in cricket prediction. To create accurate predictions on match outcomes and scores, analyze massive volumes of data, detect trends, and uncover hidden truths. Address existing system constraints and give full coverage. Differentiate yourself by providing user-friendly phone and online applications that ensure a seamless and accessible user experience. Provide cricket fans with a new and better predictive analysis platform to revolutionize the way cricket is understood and enjoyed.

## Purpose and Goal of the Project

This ambitious project aims to revolutionize the world of cricket by harnessing the power of machine learning and cutting-edge technologies. By leveraging the vast amount of historical data, the project seeks to provide accurate and insightful predictions for the scores and victory likelihood of Men's and Women's T20 International Cricket and T20 Club Cricket Competitions. Integrating machine learning algorithms will enable the identification of hidden patterns and trends, empowering teams and cricket enthusiasts to make informed decisions. Developing user-friendly phone and web applications using the Flutter and Django frameworks will also create a seamless and convenient platform for accessing the prediction of first-innings scores, first-innings win, and second-innings win. The goal is to provide cricket enthusiasts with a comprehensive hub that offers accurate forecasts and enhances their overall viewing experience. This project aims to elevate how cricket is understood, enjoyed, and analyzed by delivering valuable insights and fascinating features.

## Organization of the Report

Chapter 1 introduces the report, including background information and the motivation behind the project. It also outlines the purpose and goals of the project and describes how the report is organized. Chapter 2 presents a literature review of existing research on the topic, including relevant articles. It also discusses the limitations of the existing research. Chapter 3 details the methodology used in the project, including the system design and the proposed methodology for the complete research. It also describes the machine learning and software components, the dataset used, the applicable input and output of a real-world application, the user role and access, and the required tools. The chapter may also cover real-world application implementation. Chapter 4 focuses on the investigation or experiment, result analysis, and discussion related to the project. Chapter 5 examines the project's impacts on societal, health, safety, legal, cultural, and environmental aspects and its contribution to sustainability. Chapter 6 deals with project planning and budget considerations. Chapter 7 addresses complex engineering problems (CEP) and complex engineering activities (CEA) encountered during the project. Chapter 8 presents the conclusions of the report, summarizing the findings. It also discusses the limitations of the project and suggests areas for future improvement. The report concludes with a list of references cited throughout the document.

# **Chapter 2 Research Literature Review**

## Existing Research and Limitations

As machine learning technology continues to evolve, many related theories and technological applications are gaining more and more attention. At present, it has been widely used in the predictive intelligent system. In the world of sports, particularly cricket, machine learning has forged its way in the prediction of team and individual performance. Lamsal et al. [6] developed some models for predicting the outcome of IPL matches where they trained multiple machine learning models, including Naive Bayes, Extreme Gradient Boosting, SVM, Logistic Regression, Random Forests, and Multiple Linear Regression using data from the IPL's first ten seasons. The accuracy rates of the models are as follows: Naive Bayes (50%), Extreme Gradient Boosting (55%), SVM (63.33%), Logistic Regression (68.33%), Random Forests (68.33%), and Multiple Linear Regression (68.33%). It should be noted that this study has limitations due to the use of only two machine-learning models and five features. H. Barot et al. [7] some developed machine learning models to predict IPL match result. The authors have emphasized analyzing the characteristics of IPL cricket matches. Moving on to the study of the Indian Premier League, this paper has labeled the Batsmen and Bowlers in a unique manner based on their performance. Aside from the traditional attributes like the toss and location of the games, a few critical variables like team form and team power; furthermore, a new analysis of batting and bowling based on the Batting Index and Bowling Index has suggested. Machine Learning methods such as SVM, Logistic Regression, Random Tree, Random Forest, and Naive Bayes have been used to forecast matches. The precision of the Decision Tree and Logistic Regression methods is greater than 87% and 95%, respectively. R. Ahmed et al. [8] applied two machine learning algorithms: Linear Regression and Rigid Regression for forecasting the first innings score of IPL matches. The authors did not mention Accuracy, they chose Linear Regression for model development since it predicts more accurately. After integrating a machine learning model and backend with Flask to display the predicted score in a web application, they connected the front-end (HTML) and the back-end (python). V. Bhatnagar et al. [9] built a machine learning model to forecast the likelihood of a team winning in the IPL by using Logistic Regression (LR) techniques. The accuracy of the suggested LR algorithm is 80%. The website was built with Streamlit and Python and then launched on a Heroku server. The website doesn't exist now. The significance of the study of E. Mundhe et al. [10] have created a web application using Python and Flask to do predictive analysis on a live IPL match in order to forecast the outcome of the game and the winner before it starts. The Multivariate Polynomial Regression and Random Forest Classifier algorithm's accuracy has 67.3% and 66%, or 6 out of 10 times the actual score has been come inside the projected score range, when it has used to predict the runs scored for a live match at the conclusion of the 20 overs match. For score prediction, XGboost has 55.67%, Multivariate Polynomial Regression has 67.3%, SVM has 47.6%, Random Forest Classifier 66% accuracy. For win prediction, the Logistic Regression has 42.10% and Random Forest Classifier has 55% accuracy. S. Agrawal et al. [11] has proposed some machine learning-based approach. Based on collaboration and each player's capacity to contribute to the match's success, this model foresaw outcomes. The Support vector machine, CTree, and Naive Bayes are three alternative machine learning methods that are employed to forecast the winning team based on prior performance. After being transformed into numerical form, the prediction data sorted into three categories for victory, defeat, and tie. These classifiers achieved accuracy scores of 95.96%, 97.98%, and 98.98%, respectively. With a 98.98% accuracy rate, the NaIve Bayes classifier produced the best results. K. R. Reddy et al. [12] investigated the 2008 to 2019 IPL dataset from kaggle and reviewed some pertinent literature. Their study is more noteworthy since they provided a chatbot utilizing IBM Cognos Tools, IBM Watson Assistant, and Flask to display the findings of their analysis. They conducted their research using machine learning and neural network models, and the findings indicate that for IPL games, teams, venues, the winner of the toss, the match's location, and the decision made after winning the toss are significant determinants of victory. Sequential Model, Linear Regression, and Random Forest Regression from these three Machine Learning Models have been studied, and Sequential Model performed the best in terms of predicting the results of an IPL game. The accuracy of the sequential model is 91.3%. M. Awais et al. [13] has examined team strength, the winner of the toss and the outcome of the toss; to predict the outcome of PSL (Pakistan Super League) matches by using Gaussian Naive Bayes, K-Nearest Neighbours (KNN), and Random Forest. KNN, Gaussian Naive Bayes, and Random Forest each have prediction accuracy of 49.99%, 19.86%, and 60.1%, respectively. The study of K. C. Srikantaiah et al. [14] have indicated that they have been made predictions about the outcome of a match between two teams using a variety of variables, including the team's composition, the batting and bowling averages of each player, and the team's performance in prior matches, in addition to more conventional variables like the toss, venue, day/night, and the likelihood of winning by batting first at a given match venue against a given team. Using machine learning algorithms including SVM, Random Forest Classifier (RFC), Logistic Regression, and K-Nearest Neighbour (KNN), they suggested a model for forecasting the results of the IPL matches. The Random Forest method outperformed other algorithms, according to experimental findings, with an accuracy of 88.10%. The accuracy of KNN, Logistic Regression, SVM and Linear Regression is 49.34%, 51.40%, 32.6% and 44.05% respectively. To create a model utilizing machine learning techniques for forecasting the player's performance and identifying the greatest all-rounder, M. Shetty et al. [15] have concentrated on other aspects such as pitch type, weather, venue and opponent. They look at novel machine classification techniques such Decision Tree, Weighted Random Forest, Support Vector Machine, and Naive Bayes. Using the Random Forest Algorithm, the model provided accuracy for performance of batters, bowlers, and all-rounders of 76%, 67%, and 96%, respectively. According to the study of A. Basit et al. [16], they have discussed the probable winner of the seventh T20 World Cup. In order to forecast who would win the T20 cricket world championship, the author has compared numerous machine learning algorithms. Data was gathered from ESPNCricinfo 1. Once the dataset has been gathered, it is checked for integrity and cleanliness using a variety of approaches. The next step is to build predictive models utilizing a number of machine learning techniques from the decision tree algorithm family, such as Random Forest Classifier, Decision Trees, and Extra Trees. They make the observation that Random Forest outperformed the other Classifiers when evaluated using the extracted dataset (s). It has 80.86% accuracy. The accuracy of the Extra Trees, Decision Trees (ID3) and Decision Trees (C4.5) is 79.67%, 74.69% and 79.73%, respectively. Arjun et al. [17] has examined sixteen various training model components. They considered several factors including the average number of fours and sixes a player hits, player average strike rate, the number of times player is staying not out, the number of half-centuries and centuries player has scored, the number of games player has played with current and average run, the number of wide balls and no-balls, the average run granted by the player, and the most recent number of maiden over player has bowled and they collected data of all T-20 league, international, and domestic matches. They use data from five thousands T-20 matches from all data to train Naive Bayes, Randomized Forest, Decision Trees, and Support Vector Machine (SVM) ML algorithms, which are then used to test the simulation findings. When put to the test, the SVM algorithm outperformed the other models with a prediction accuracy of 63.89%. K. Passi et al. [18] took into account the team performance of individual team members based on runs scored by each batter and wickets taken by a bowler. Writers forecasted results based on performance using machine learning with improved accuracy. The author used a variety of classifiers, including Naive Bayes, Random Forest, Multi-Class SVM, and decision trees, to predict a model and classify the problem. The management of the club finds it challenging to choose a roster of eleven players from a total of fifteen players after evaluating each player's performance and match-critical contributions. This domain has an intriguing feature that reliably forecasts results based on individual performance, such as runs scored by a batsman and wickets taken by bowlers, as well as past performance, pitch condition, and weather conditions. There are several classifiers that can forecast a batsman's run total and a bowler's wicket total, but random forest has been found to be the most accurate at 90.74% and 92.25%, respectively. Nevertheless, the results of the SVM were, respectively, 51.45% and 68.78% and also the maximum accuracy of Decision Tree Algorithm were, respectively, 80.46% and 86.5%. They also show how test and train data splitting can impact on accuracy scores. The author studied and described a number of factors that affected how T20 matches turned out. A method called multivariate regression has been suggested to measure the team points. Every individual and team's past performance has been scrutinized, and the likelihood of defeating an opponent has been calculated using that data. Ultimately, other attributes or characteristics that may be utilized to forecast matches are discovered. For this experiment, a number of machine learning classifiers are trained and observed, the Logistic Regression, KNN, SVM, Decision Tree and Random Forest emerged as accuracy rates respectively of 29.84%, 64.78%, 88.79%, 89.15% and 89.15%. R. Ul Mustafa et al. [19] proposed a machine learning algorithm that analyzes user sentiment on social media sites like Twitter and forecasts game results. In this essay, the author used data from social media sites like Twitter to predict the outcome of a cricket match. Everyone wrote remarks in support of their side on social media before the game began since it is a place where one may express one's thoughts without fear. Based on the amount of tweets in support of each team and the predicted scores for each team, three alternative methods are employed to determine the winner. Author used the three methods mentioned above to predict the outcome of the match before it began. Author then used supervised learning techniques to evaluate this outcome and claimed that the Support vector machine (SVM) produced better results than other methods like NaIve Bayes and Logistic Regression. With the use of these strategies, the author was able to achieve an accuracy of around 75%. This model is applicable to all cricket matches, and it can also be used to provide outcomes for other games like hockey, baseball, and basketball. Decision trees and MLP networks have been used by J. Kumar et al. [20] to develop a model predicting the result of ODI matches. The toss, run rate, remaining wickets, strike rate, venue, historical data, team strength, and other parameters were used by the author to estimate the model's outcome. All of these elements have an impact on the outcome of a match, both before and during play. Decision trees and a Multilayer Perceptron Network (MLP) are two separate machine learning approaches used to research and determine the impact of a match. A prediction system has been created using these classifiers. In statistical technique, a mathematical equation is used to establish a relationship between variables, but in this model, assumptions have already been made on the data variables and their main relationships. The accuracy of the Decision trees and MLP networks methods is 55.1% and 57.4%, respectively.

The following observations have been made after a detailed examination of the literature reviews − (i) most of the articles emphasized on IPL (Indian Premier League), some of articles on T20 International or World Cup Cricket and only one article on PSL (Pakistan Super League), (ii) in general, those articles did not care about female T20 cricket, (iii) also, those articles did not care about other T20 Leagues for example BBL (Big Bash League), WBL (Women Big Bash), BPL (Bangladesh Premier League) and etc., (iv) most importantly, in terms of determining victory chance of a T20 match, all proposed system of those articles are out of date because of used data is older, (v) furthermore, in terms of forecasting the likely result, the proposed system of those articles' accuracy is poor, and they are also out of date, and (vi) most of the article did not emphasize to develop Real World Application.

These investigations have motivated us to study the International T20 games and most popular T20 Franchises leagues with latest datasets and apply three machine learning approaches (xgboost regression, random forest regression and logistic regression).

# **Chapter 3 Methodology**

## System Design

### Proposed Methodology of Complete Research

According to the objective of the research, some tournaments of T20 Cricket games have been considered. At the starting of the discussion of proposed methodology of the research, the considered tournaments or cricket games that have been considered for the research are pointed out here − (i) Men’s T20 International, (ii) Women’s T20 International, (iii) Bangladesh Premier League (BPL), (iv) Indian Premier League (IPL), (v) Big Bash League (BBL), (vi) Women’s Big Bash League (WBBL), (vii) T20 Blast, (viii) Pakistan Super League (PSL), (ix) Lankan Premier League (LPL), (x) Caribbean Premier League (CPL), (xi) Men’s Super Smash, (xii) Women’s Super Smash, and (xiii) Women's Cricket Super League (WCSL). This list of the tournaments has been considered in the data collection process of the research.

There are three modules proposed in this study namely first innings score prediction, first innings win prediction and second innings win prediction. First Innings score prediction module will predict the probable final score of batting team, first innings win prediction module will predict the probable winning chance of batting team by that final score and second innings win prediction module will predict the probable winning chance of batting team; when second innings will be running. Most importantly, xgboost algorithm, random forest regression algorithm and logistic regression algorithm are considered for building a machine learning model in this study. For the first inning score prediction module, there has been a comparison with xgboost algorithm and random forest regression algorithm to choose the best model for Real World Applications. For the win prediction module, the logistic regression algorithm has been highly considered.

Fig. 1 is presenting the workflow of the research. Firstly, research study started from collecting past information of related research objectives. Secondly, data acquisition from the collected past information. Thirdly, data is pre-processed to filter only required information to extract feature data for machine learning models. Fourthly, machine learning models are developed by using xgboost algorithm, random forest regression algorithm and logistic regression algorithm then evaluate those models by training and testing those models by using extracted feature data. Subsequently, according to the evaluation of models, a deep learning model extracted for declared modules of Real World Applications. Then, real world web and phone applications will be developed by the name of structured modules and will be tested and bug fixed.

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Diagram\main_work flow_transparent.png |

Figure 1. Workflow of the research

### Proposed Methodology of Real World Applications

Fig. 2 is presenting the architecture of Real World Application of the research. There two parts of the Real World Application one is Web Application and the other is Phone Application. To develop web application and Application Programming Interfaces (API) Django Framework will be considered; to develop phone application Flutter framework will be considered.

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Diagram\Full System Architecture_transparent.png |

Figure 2. Architecture of Real World Application

### Database Design and Entity Relationship Diagram

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Diagram\Database ER diagram_transparent.png |

Figure 3. Entity Relationship (ER) diagram of database

Fig. 3 is presenting the ER diagram of the Real World Applications database of the research. PostgreSQL has been used to build the database.

## Hardware and/or Software Components

### Data Collection and Description of Collected Data

The information has been acquired from cricsheet.org [19]. Data is accessible on this page in YAML format, which is a human-readable format. There is a separate file for each match that contains information about all matches (ODI, TEST, T20, and Club Competitions), the season year, details of ball by ball innings data, match winner, player of the match, match number, location, stadium name, umpires' particulars, participating teams, and margin of victory. The collected data contains 8229 T20 International and Club Competitions contests since 2005 to 2023. The description of collected data has been pointed here –

**ID** – The attribute needs to contain the information about the unique id for a match.

**SEASON** – The attribute needs to contain the information about the year when the match has been conducted.

**VENUE** – The attribute needs to contain the information about the venue where the match took place.

**TEAM** – The attribute needs to contain the information about the batting team and bowling team.

**TOSS** – The attribute needs to contain the information about the winner of the toss.

**INNINGS** – The attribute needs to contain the information about which innings is running, which team is doing bat and bowl.

**WINNER OF THE MATCH** – The attribute needs to contain the information about the winner of the match.

**BALL BY BALL INFORMATION** – The attribute needs to contain the information about the runs in every ball, wicket information.



### Data Exploration

Data exploration is the process of data acquisition from data files. The given data was in YAML format with a separate file for each match, necessitating thorough preparation. The yaml and os Python package has been used to first read data into Python and normalize it. It was afterwards transformed into a data frame using a different program called pandas. After then, all of these files were joined together to make a single data frame that contained the information from each match. In order to find any abnormalities, exploratory data analysis was done after that. Usually, datasets have errors that need to be fixed before using the methods. There might be a number of null values, or no result game values, that affect the result. When using machine learning algorithms, data pre-processing creates a more understandable structure.

### Data Pre-processing and Feature Data Extraction

Collected data contained unnecessary contents and missing values for different columns due to error(s) in recording or parsing. To avoid improper features data there has been done some processing techniques:

* **Checking Tournament Type:** Tournament type has been checked from data and those statistics have been deleted since those are not the respective T20 International and T20 tournaments data.
* **Checking Null Values:** Those statistics have been removed, those components have null values, and those games have no outcomes.
* **First and Second Innings Data Separation:** To reach the goal of the research this part is most important. A manual python function has been written for this process. After exploration of data, there was a column named "innings". The "innings" attribute had contained "ball by ball" first and second innings data with team names, winner of the match, ball number with runs of that ball with whether it was out to any player in that ball, and venue name. All of these data has been extracted by the manual python function. Note that, main feature data has been extracted from these data.
* **Batting and Bowling Team Name Extraction:** Batting team and bowling team name of every row has been extracted from the "team names". In the "teams" attribute, there were two team names, first one was batting team and second one was bowling team.
* **City Name:** City name has been extracted from the "venue" attribute by removing unnecessary alphabets and by taking only the city name of the venue. City name has been considered by replacing venue name for reducing data clashes and complexities.
* **Current Score:** Current score of every match has been counted by doing cumulative sum of "runs" attribute according to ball.
* **Wicket Left:** Wicket left of every ball has been extracted and counted from the "player dismissed" attribute. First, if there has been found a player name that has been replaced by integer value 1. Then, cumulative sum has been counted by group of match Id. Last, the result sum value of every row has been subtracted with 10.
* **Delivered Balls:** Balls bowled mean how many balls already done. Balls bowled data have been extracted from "balls" attribute.
* **Ball Left:** "Ball left" means how many balls have been left to play. This feature has been extracted by subtraction of how many balls already done with 120.
* **Total Score:** The total score is the total sum of runs of every match.
* **Target Score:** The target score has been counted by adding one with the total score of every match.
* **Run Left:** Runs left means how many runs left to reach the target. Runs left has been counted by subtracting target runs with current runs.
* **Current Run Rate:** Current run rate is notated by crr. According to the rules of a T20 Cricket match, there are six balls per over. T20 Cricket match current run rate calculated by multiplying the current runs with the number of balls per over then divided by the number of balls already played.

|  |  |
| --- | --- |
|  | (1) |

* **Required Run Rate:** Required run rate is notated by rrr. According to the rules of a T20 Cricket match, there are six balls per over and a total of hundred & twenty balls per match. T20 Cricket match required run rate calculated by multiplying the required runs with the number of balls per over then divided by number of balls left from total balls.

|  |  |
| --- | --- |
|  | (2) |

In this process some python technical functions are used that are allowed by numpy and pandas libraries

### Feature Data

Extracted Feature Dataset of every tournament which is the part of the research has been shown here:

* **BBL Feature Data:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\bbl\bbl_score_feature_data.PNG |

Figure 4. A snippet of the BBL score prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\bbl\bbl_first_innings_win_feature_data.PNG |

Figure 5. A snippet of the BBL first innings win prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\bbl\bbl_second_innings_win_feature_data.PNG |

Figure 6. A snippet of the BBL second innings win prediction model feature dataset

Fig. 4 presented a snippet of the BBL score prediction model feature dataset; Fig. 5 presented a snippet of the BBL first innings win prediction model feature dataset and Fig. 6 presented a snippet of the BBL second innings win prediction model feature dataset.

* **BPL Feature Data:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\bpl\score_prediction_feature_data.jpg |

Figure 7. A snippet of the BPL first innings score prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\bpl\bpl_1st_innings_win_feature_data.jpg |

Figure 8. A snippet of the BPL first innings win prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\bpl\bpl_2nd_innings_win_feature_data.jpg |

Figure 9. A snippet of the BPL second innings win prediction model feature dataset

Fig. 7 presented a snippet of the BPL score prediction model feature dataset; Fig. 8 presented a snippet of the BPL first innings win prediction model feature dataset and Fig. 9 presented a snippet of the BPL second innings win prediction model feature dataset.

* **CPL Feature Data:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\cpl\cpl_score_feature_data.PNG |

Figure 10. A snippet of the CPL first innings score prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\cpl\cpl_first_innings_win_feature_data.PNG |

Figure 11. A snippet of the CPL first innings win prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\cpl\cpl_second_innings_win_feature_data.PNG |

Figure 12. A snippet of the CPL second innings win prediction model feature dataset

Fig. 10 presented a snippet of the CPL score prediction model feature dataset; Fig. 11 presented a snippet of the CPL first innings win prediction model feature dataset and Fig. 12 presented a snippet of the CPL second innings win prediction model feature dataset.

* **CSA T20 Feature Data:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\csa t20 challenge\csa_t20_score_prediction_feature_data.PNG |

Figure 13. A snippet of the CSA T20 first innings score prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\csa t20 challenge\csa_t20_first_innings_win_feature_data.PNG |

Figure 14. A snippet of the CSA T20 first innings win prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\csa t20 challenge\csa_t20_second_innings_win_feature_data.PNG |

Figure 15. A snippet of the CSA T20 second innings win prediction model feature dataset

Fig. 13 presented a snippet of the CSA T20 score prediction model feature dataset; Fig. 14 presented a snippet of the CSA T20 first innings win prediction model feature dataset and Fig. 15 presented a snippet of the CSA T20 second innings win prediction model feature dataset.

* **IPL Feature Data:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\ipl\ipl_score_feature_data.PNG |

Figure 16. A snippet of the IPL first innings score prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\ipl\ipl_first_innings_win_feature_data.PNG |

Figure 17. A snippet of the IPL first innings win prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\ipl\ipl_second_innings_win_feature_data.PNG |

Figure 18. A snippet of the IPL second innings win prediction model feature dataset

Fig. 16 presented a snippet of the IPL score prediction model feature dataset; Fig. 17 presented a snippet of the IPL first innings win prediction model feature dataset and Fig. 18 presented a snippet of the IPL second innings win prediction model feature dataset.

* **LPL Feature Data:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\lpl\lpl_score_feature_data.PNG |

Figure 19. A snippet of the LPL first innings score prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\lpl\lpl_first_innings_win_feature_data.PNG |

Figure 20. A snippet of the LPL first innings win prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\lpl\lpl_second_innings_win_feature_data.PNG |

Figure 21. A snippet of the LPL second innings win prediction model feature dataset

Fig. 19 presented a snippet of the LPL score prediction model feature dataset; Fig. 20 presented a snippet of the LPL first innings win prediction model feature dataset and Fig. 21 presented a snippet of the LPL second innings win prediction model feature dataset.

* **PSL Feature Data:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\psl\score_prediction_feature_data.jpg |

Figure 22. A snippet of the PSL first innings score prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\psl\psl_1st_innings_win_feature_data.jpg |

Figure 23. A snippet of the PSL first innings win prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\psl\psl_2nd_innings_win_feature_data.jpg |

Figure 24. A snippet of the PSL second innings win prediction model feature dataset

Fig. 22 presented a snippet of the PSL score prediction model feature dataset; Fig. 23 presented a snippet of the PSL first innings win prediction model feature dataset and Fig. 24 presented a snippet of the PSL second innings win prediction model feature dataset.

* **T20 Blast Feature Data:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\t20 blast\t20_blast_score_feature_data.PNG |

Figure 25. A snippet of the T20 Blast first innings score prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\t20 blast\t20_blast_first_innings_win_feature_data.PNG |

Figure 26. A snippet of the T20 Blast first innings win prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\t20 blast\t20_blast_second_innings_win_feature_data.PNG |

Figure 27. A snippet of the T20 Blast second innings win prediction model feature dataset

Fig. 25 presented a snippet of the T20 Blastscore prediction model feature dataset; Fig. 26 presented a snippet of the T20 Blast first innings win prediction model feature dataset and Fig. 27 presented a snippet of the T20 Blast second innings win prediction model feature dataset.

* **Men’s T20i Feature Data:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\t20 men's\t20_men_first_innings_score_feature_data.PNG |

Figure 28. A snippet of the Men’s T20i first innings score prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\t20 men's\t20_men_first_innings_win_feature_data.PNG |

Figure 29. A snippet of the Men’s T20i first innings win prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\t20 men's\t20_men_second_innings_win_feature_data.PNG |

Figure 30. A snippet of the Men’s T20i second innings win prediction model feature dataset

Fig. 28 presented a snippet of the Men’s T20 International score prediction model feature dataset; Fig. 29 presented a snippet of the Men’s T20 International first innings win prediction model feature dataset and Fig. 30 presented a snippet of the Men’s T20 International second innings win prediction model feature dataset.

* **Women’s T20i Feature Data:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\t20 women's\t20_women_score_feature_data.PNG |

Figure 31. A snippet of the Women’s T20i first innings score prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\t20 women's\t20_women_first_innings_win_feature_data.PNG |

Figure 32. A snippet of the Women’s T20i first innings win prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\t20 women's\t20_women_second_innings_win_feature_data.PNG |

Figure 33. A snippet of the Women’s T20i second innings win prediction model feature dataset

Fig. 31 presented a snippet of the Women’s T20 International score prediction model feature dataset; Fig. 32 presented a snippet of the Women’s T20 International first innings win prediction model feature dataset and Fig. 33 presented a snippet of the Women’s T20 International second innings win prediction model feature dataset.

* **WBBL Feature Data:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\wbbl\wbl_score_feature_data.PNG |

Figure 34. A snippet of the WBBL first innings score prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\wbbl\wbl_first_innings_win_feature_data.PNG |

Figure 35. A snippet of the WBBL first innings win prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\wbbl\wbl_second_innings_win_feature_data.PNG |

Figure 36. A snippet of the WBBL second innings win prediction model feature dataset

Fig. 34 presented a snippet of the WBBL score prediction model feature dataset; Fig. 35 presented a snippet of the WBBL first innings win prediction model feature dataset and Fig. 36 presented a snippet of the WBBL second innings win prediction model feature dataset.

* **Men’s Super Smash:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\super smash men's\ssm_score_feature_data.PNG |

Figure 37. A snippet of the Men’s Super Smash first innings score prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\super smash men's\ssm_1st_innings_win_feature_data.PNG |

Figure 38. A snippet of the Men’s Super Smash first innings win prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\super smash men's\ssm_2nd_innings_win_feature_data.PNG |

Figure 39. A snippet of the Men’s Super Smash second innings win prediction model feature dataset

Fig. 37 presented a snippet of the Men’s Super Smash score prediction model feature dataset; Fig. 38 presented a snippet of the Men’s Super Smash first innings win prediction model feature dataset and Fig. 39 presented a snippet of the Men’s Super Smash second innings win prediction model feature dataset.

* **Women’s Super Smash:**

|  |
| --- |
| **E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\super smash women's\ssf_score_feature_data.PNG** |

Figure 40. A snippet of the Women’s Super Smash first innings score prediction model feature dataset

|  |
| --- |
| **E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\super smash women's\ssf_1st_innings_feature_data.PNG** |

Figure 41. A snippet of the Women’s Super Smash first innings win prediction model feature dataset

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\super smash women's\ssf_2nd_innings_feature_data.PNG |

Figure 42. A snippet of the Women’s Super Smash second innings win prediction model feature dataset

Fig. 40 presented a snippet of the Women’s Super Smash score prediction model feature dataset; Fig. 41 presented a snippet of the Women’s Super Smash first innings win prediction model feature dataset and Fig. 42 presented a snippet of the Women’s Super Smash second innings win prediction model feature dataset.

### Machine Learning Models Building

XGboost (eXtreme Gradient Boosting) algorithm and Random Forest Algorithm have been used for Regression to forecast probable score and Logistic Regression algorithm for classification to win probability prediction. Scikit-learn library has provided various tools and libraries to develop and evaluate perfect Machine Learning (ML) Model. Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python.

* **Random Forest Regression:** Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting [22]. Equation 3 represents the Random Forest Regressor model.

|  |  |
| --- | --- |
| *RandomForestRegressor()* | (3) |

**Xgboost Regression:** XGBoost (Extreme Gradient Boosting) is a popular supervised-learning algorithm used for regression and classification on large datasets. It uses sequentially-built shallow decision trees to provide accurate results and a highly-scalable training method that avoids over fitting. So, XGBoost is a distributed gradient-boosted decision tree (GBDT) machine learning library [23]. When gradient boosting is used for regression, the weak learners are regression trees, and each regression tree translates an input data point to one of its leafs, each of which holds a continuous score. XGBoost minimizes a regularized (L1 and L2) objective function composed of a convex loss function (based on the difference between the expected and goal outputs) and a model complexity penalty term. (in other words, the regression tree functions). Iteratively, new trees are added to forecast the residuals or mistakes of earlier trees, which are then merged with previous trees to make the final projection. Gradient boosting gets its name from the fact that it employs a gradient descent method to reduce loss when adding new models. [24]. Fig. 43 is a brief illustration on how gradient tree boosting works. Equation 4 represents the XGBoost Regressor model.

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Diagram\xgboost_illustration.png |

Figure 43. Illustration of gradient tree boosting

|  |  |
| --- | --- |
| *XGBRegressor(n\_estimator=1000, learning\_rate=0.2, max\_depth=20)* | (4) |

n\_estimators — the number of runs XGBoost will try to learn. learning\_rate — learning speed.

* **Logistic Regression:** Logistic regression is an example of supervised learning. This type of statistical model (also known as logit model) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn’t vote, based on a given dataset of independent variables [25]. The sigmoid or Logistic function can be expressed as and are presented in Fig. 44. Equation 6 represents the Logistic Regression model. Equation 5 represents the internal function which has been used by logistic regression when regression has been performed.

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Diagram\logistic-regression-in-machine-learning.png |

Figure 44. Sigmoid function in graphic view

|  |  |  |
| --- | --- | --- |
|  | (5) | |
| *LogisticRegression()* | | (6) |

### Train Models

The model train process has started through data splitting into train and test sets using the Train Test Split method which is provided by the Scikit-learn library. 90% of data from every feature dataset has been splitted to train machine learning models. 10% data has been saved as test datasets to use at models evaluation process.

At the beginning, three steps prominent Pipeline method has been used to train machine learning models which has been presented at equation 7. First step as step 1, Two methods have been used namely  Columntransformer and Onehotencoder have been used  to transform every column that contains objects (string) values to numeric values as integers. Second step as step2, StandardScaler method has been used to remove the mean and scales each feature/variable to unit variance. Third step as step3, the machine learning model has been used. Finally, the train dataset has been fitted into the pipeline to train the machine learning models.

|  |  |
| --- | --- |
| *Pipeline(steps=[*  *(‘step1’, <method>),*  *(‘step2’, <method>),*  *(‘step3’, <machine learning model>)*  *])* | (7) |

### Machine Learning Model Evaluation

To evaluate all of machine learning model performance effectively, the performance metric must be defined first. The efficiency and quality of any machine learning models are judged by the numerous prescribed performance metrics, which are themselves a single index.

#### **R-square Score**

The R2 (R-Square) score [24] is a very important metric that is used to evaluate the performance of a regression-based machine learning model. It is pronounced as R squared and is also known as the coefficient of determination. It works by measuring the amount of variance in the predictions explained by the dataset. R2 Score scale is from 0 to 1. If R2 Score is better than 0.9 that means the ML model is good and score is closer to 1 means the model is perfect.

The r2 score was calculated using a Python program,

|  |  |
| --- | --- |
| *r2\_score ( y\_true, y\_predict )* | (8) |

Here, y\_true = actual target values, y\_predict = estimated target values

#### **Confusion Matrix**

A confusion matrix is a table that is used to define a classification algorithm's performance. A confusion matrix visualizes and summarizes a classification algorithm's performance. Fig. 45 is presenting the structure of the confusion matrix.

|  |
| --- |
| E:\CSE499 CSE498R\The Hundrad DL research\diagram\confusion matrix.jpg |

Figure 45. Confusing matrix structure

True Negative (TN) indicates that prediction is negative and it’s true. True Positive (TP) indicates that prediction is positive and it’s true. False Negative (FN) indicates that prediction is negative but it’s false. False Positive (FP) indicates that prediction is positive but it’s false.

The confusion matrix was calculated using a Python program,

|  |  |
| --- | --- |
| *confusion\_matrix ( y\_true, y\_predict )* | (9) |

Here, y\_true = actual target values, y\_predict = estimated target values

#### **Accuracy Score**

This compares the actual results to the classifier's projected results for a particular input dataset. The collection of real true labels in the testing dataset must match the matching set of anticipated labels for the best accuracy score. Best value: 1 and Worst value: 0. The accuracy score was calculated using a Python program,

|  |  |
| --- | --- |
| *accuracy\_score ( y\_true, y\_predict )* | (10) |

Here, y\_true = actual target values, y\_predict = estimated target values

#### **F1 Score**

The F1 score is a machine learning assessment statistic that gauges the accuracy of a model. It combines a model's accuracy and recall scores. The accuracy statistic calculates how many times a model predicted correctly over the full dataset. The f1 score was calculated using a Python program,

|  |  |
| --- | --- |
| *f1\_score ( y\_true, y\_predict )* | (11) |

Here, y\_true = actual target values, y\_predict = estimated target values

#### **Mean Absolute Error**

The average variance between the significant values in the dataset and the predicted values in the same dataset is defined as the mean absolute error (MAE).

The mean absolute error was calculated using a Python program,

|  |  |
| --- | --- |
| *mean\_absolute\_error ( y\_true, y\_predict )* | (12) |

Here, y\_true = actual target values, y\_predict = estimated target values

Results of evaluations of machine learning models have been discussed at chapter 4**.**

### Required Tools

This section has presented all the software tools which have been used throughout the entire research and development process and also presented similar software tools of those tools.

Table I. List of Software Tools

|  |  |  |  |
| --- | --- | --- | --- |
| **Tool** | **Functions** | **Other similar Tools (if any)** | **Why selected this tool** |
| Android Studio | Develop and compile phone applications and build APK | Visual Studio Code  Xcode | It provides a unified environment to build apps for Android phones, tablets, Android Wear, Android TV, and Android Auto. |
| Visual Studio Code | Develop and compile phone applications, and web applications and build APK | Notepad++  Sublime Text  Atom  Brackets | It is a streamlined code editor supporting development operations like debugging, task running, and version control. |
| PyCharm | To develop and compile the main system backend and web application | Visual Studio Code  Wing Python IDE  Selenium IDE  Aptana Studio | It is a dedicated Python Integrated Development Environment (IDE) providing a wide range of essential tools for Python developers, tightly integrated to |
| Anaconda IDE | Anaconda IDE will provide us all essential libraries to develop machine learning models | MATLAB  Azure Machine Learning  IBM Watson Studio | To distribute of required tools and packages for Python and R programming languages for scientific computing. |
| Jupyter Notebook | It allows programmers to compile all aspects of a data sciences project in one place. | Google Colab | It makes easy to show the entire process and data visualizations of a project. |
| Flutter | To build high-quality natively compiled apps for iOS and Android quickly without having to write the code for the two apps separately. | React Native  Swift  Kotlin  Xamarin  Android Native (Java) | For developing cross-platform mobile application. |
| Git | To track source code changes, enabling to work together on non-linear development. | AWS Code-Commit  Gitlab  Bitbucket | It is a free and open-source version control system that efficiently handles small to very large projects. |
| Lucidchart | Lucidchart offers a wide range of templates and shapes that allows customize easily. It saves time and effort when creating diagrams. | Draw.io  Visio  Gliffy  Smartdraw | Lucidchart allows create diagrams that help to visualize complex ideas and processes, making communicating with others and identify potential issues easier. |

# **Chapter 4 Result, Analysis and Discussion**

The analytical discussion on the evaluation of every model has been presented in this section according to tournaments.

## First Innings Score Prediction Models Result Analysis

To evaluate every tournament's score prediction models, predicted values versus accurate values graph has been plotted and r2 score and mean absolute error (MAE) has been calculated.

Deciding factors of best model for score prediction modules: (i) Predicted Values versus Accurate Values graph which graph has more values are nearest to slope, (ii) which models has R2 Score better than another, and (iii) which models has lower mean absolute error (MAE) value.

* **BBL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\bbl\bbl_score_graph.PNG |

Figure 46. Predicted Values vs Accurate Values Graph of BBL score prediction models

Table II. Evaluation table of BBL score prediction models

|  |  |  |  |
| --- | --- | --- | --- |
| **Xgboost Regressor** | | **Random Forest Regressor** | |
| R2 Score | MAE | R2 Score | MAE |
| 0.96 | 5.26 | 0.95 | 6.46 |

The provided graphs illustrate the relationship between the predicted values and the accurate values for score prediction; Fig. 46. One graph represents the predictions made by the XGBoost Regressor, while the other graph represents the predictions made by the Random Forest Regressor. Both graphs show that the predicted values align closely with the accurate values, indicating the models' ability to capture the underlying patterns in the data. Furthermore, an evaluation table is provided, which includes the R2 score and mean absolute error (MAE) for both algorithms. Upon analysis, the XGBoost Regressor achieved an R2 score of 0.96 and an MAE of 5.26, indicating strong correlation and low prediction errors. The Random Forest Regressor obtained an R2 score of 0.96 and an MAE of 6.46, suggesting slightly larger errors. Based on these results, the XGBoost Regressor is chosen as the preferred algorithm for score prediction due to its higher accuracy and lower MAE than the Random Forest Regressor. Opting for the Random Forest Regressor would result in slightly lower accuracy and higher prediction errors.

* **BPL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\bpl\bpl_score_prediction_graph.PNG |

Figure 47. Predicted Values vs Accurate Values Graph of BPL score prediction models

Table III. Evaluation table of BBL score prediction models

|  |  |  |  |
| --- | --- | --- | --- |
| **Xgboost Regressor** | | **Random Forest Regressor** | |
| R2 Score | MAE | R2 Score | MAE |
| 0.97 | 4.42 | 0.96 | 5.59 |

The provided graphs illustrate the relationship between the predicted values and the accurate values for score prediction; Fig. 47. One graph represents the predictions made by the XGBoost Regressor, while the other represents the predictions made by the Random Forest Regressor. Both graphs show a close alignment between the predicted and accurate values, indicating the models' ability to capture the underlying patterns in the data. Regarding evaluation metrics, the XGBoost Regressor achieved an R2 score of 0.97 and a mean absolute error (MAE) of 4.42, indicating strong accuracy and low prediction errors. On the other hand, the Random Forest Regressor achieved an R2 score of 0.96 and an MAE of 5.59, demonstrating reasonable performance but with slightly lower accuracy and more significant prediction errors compared to the XGBoost Regressor. Based on these results, the XGBoost Regressor is selected as the preferred algorithm for score prediction in this project due to its superior accuracy and more precise predictions. Choosing the Random Forest Regressor would result in slightly lower accuracy and more significant prediction errors, potentially impacting the reliability of the score predictions.

* **CPL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\cpl\cpl_score_graph.PNG |

Figure 48. Predicted Values vs Accurate Values Graph of CPL score prediction models

Table IV. Evaluation table of CPL score prediction models

|  |  |  |  |
| --- | --- | --- | --- |
| **Xgboost Regressor** | | **Random Forest Regressor** | |
| R2 Score | MAE | R2 Score | MAE |
| 0.96 | 3.92 | 0.96 | 4.47 |

The provided graphs showcase the predicted values versus the accurate values for a score prediction, with one diagram representing the predictions made by the XGBoost Regressor and the other graph illustrating the predictions made by the Random Forest Regressor; Fig. 48. Both graphs align closely with the predicted and accurate values, indicating the models' proficiency in capturing the underlying data patterns. Evaluating the metrics, the XGBoost Regressor and the Random Forest Regressor achieve the same R2 score of 0.96, indicating a strong correlation between the predicted and accurate values. However, when considering the mean absolute error (MAE), the XGBoost Regressor has a lower MAE of 3.92, while the Random Forest Regressor has a slightly higher MAE of 4.47. If the Random Forest Regressor is selected instead, there would be a trade-off between accuracy and prediction errors, as it demonstrates a slightly higher MAE than the XGBoost Regressor. This may result in less precise predictions and potentially impact the reliability of the score predictions. Therefore, based on the lower MAE and higher precision, the XGBoost Regressor is selected for this preferred algorithm for score prediction.

* **CSA T20:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\csa t20 challenge\csa_t20_score_prediction_graph.PNG |

Figure 49. Predicted Values vs Accurate Values Graph of CSA T20 score prediction models

Table V. Evaluation table of CSA T20 score prediction models

|  |  |  |  |
| --- | --- | --- | --- |
| **Xgboost Regressor** | | **Random Forest Regressor** | |
| R2 Score | MAE | R2 Score | MAE |
| 0.97 | 3.37 | 0.97 | 4.50 |

The provided graphs showcase the predicted values versus the accurate values for a score prediction, with one diagram representing the predictions made by the XGBoost Regressor and the other graph illustrating the predictions made by the Random Forest Regressor; Fig. 49. Both graphs align closely with the predicted and accurate values, indicating the models' proficiency in capturing the underlying data patterns. Evaluating the metrics, the XGBoost Regressor and the Random Forest Regressor achieve the same R2 score of 0.97, indicating a strong correlation between the predicted and accurate values. However, when considering the mean absolute error (MAE), the XGBoost Regressor has a lower MAE of 3.37, while the Random Forest Regressor has a slightly higher MAE of 4.50. If the Random Forest Regressor is selected instead, there would be a trade-off between accuracy and prediction errors, as it demonstrates a slightly higher MAE than the XGBoost Regressor. This may result in less precise predictions and potentially impact the reliability of the score predictions. Therefore, based on the lower MAE and higher precision, the XGBoost Regressor is selected for this preferred algorithm for score prediction.

* **IPL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\ipl\ipl_score_graph.PNG |

Figure 50. Predicted Values vs Accurate Values Graph of IPL score prediction models

Table VI. Evaluation table of IPL score prediction models

|  |  |  |  |
| --- | --- | --- | --- |
| **Xgboost Regressor** | | **Random Forest Regressor** | |
| R2 Score | MAE | R2 Score | MAE |
| 0.96 | 4.77 | 0.95 | 6.19 |

The provided graphs illustrate the relationship between the predicted values and the accurate values for score prediction; Fig. 50. One graph represents the predictions made by the XGBoost Regressor, while the other graph represents the predictions made by the Random Forest Regressor. Both graphs show that the predicted values align closely with the accurate values, indicating the models' ability to capture the underlying patterns in the data. Furthermore, an evaluation table is provided, which includes the R2 score and mean absolute error (MAE) for both algorithms. Upon analysis, the XGBoost Regressor achieved an R2 score of 0.96 and an MAE of 4.77, indicating strong correlation and low prediction errors. The Random Forest Regressor obtained an R2 score of 0.95 and an MAE of 6.19, suggesting slightly larger errors. Based on these results, the XGBoost Regressor is chosen as the preferred algorithm for score prediction due to its higher accuracy and lower MAE than the Random Forest Regressor. Opting for the Random Forest Regressor would result in slightly lower accuracy and higher prediction errors.

* **LPL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\lpl\lpl_score_graph.PNG |

Figure 51. Predicted Values vs Accurate Values Graph of LPL score prediction models

Table VII. Evaluation table of LPL score prediction models

|  |  |  |  |
| --- | --- | --- | --- |
| **Xgboost Regressor** | | **Random Forest Regressor** | |
| R2 Score | MAE | R2 Score | MAE |
| 0.98 | 2.85 | 0.98 | 3.11 |

The provided graphs showcase the predicted values versus the accurate values for a score prediction, with one diagram representing the predictions made by the XGBoost Regressor and the other graph illustrating the predictions made by the Random Forest Regressor; Fig. 51. Both graphs align closely with the predicted and accurate values, indicating the models' proficiency in capturing the underlying data patterns. Evaluating the metrics, the XGBoost Regressor and the Random Forest Regressor achieve the same R2 score of 0.98, indicating a strong correlation between the predicted and accurate values. However, when considering the mean absolute error (MAE), the XGBoost Regressor has a lower MAE of 2.85, while the Random Forest Regressor has a slightly higher MAE of 3.11. If the Random Forest Regressor is selected instead, there would be a trade-off between accuracy and prediction errors, as it demonstrates a slightly higher MAE than the XGBoost Regressor. This may result in less precise predictions and potentially impact the reliability of the score predictions. Therefore, based on the lower MAE and higher precision, the XGBoost Regressor is selected for this preferred algorithm for score prediction.

* **PSL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\psl\psl_first_innings_score_graph.PNG |

Figure 52. Predicted Values vs Accurate Values Graph of PSL score prediction models

Table VIII. Evaluation table of PSL score prediction models

|  |  |  |  |
| --- | --- | --- | --- |
| **Xgboost Regressor** | | **Random Forest Regressor** | |
| R2 Score | MAE | R2 Score | MAE |
| 0.98 | 3.69 | 0.97 | 4.12 |

The provided graphs illustrate the relationship between the predicted values and the accurate values for score prediction; Fig. 52. One graph represents the predictions made by the XGBoost Regressor, while the other graph represents the predictions made by the Random Forest Regressor. Both graphs show that the predicted values align closely with the accurate values, indicating the models' ability to capture the underlying patterns in the data. Furthermore, an evaluation table is provided, which includes the R2 score and mean absolute error (MAE) for both algorithms. Upon analysis, the XGBoost Regressor achieved an R2 score of 0.98 and an MAE of 3.69, indicating strong correlation and low prediction errors. The Random Forest Regressor obtained an R2 score of 0.97 and an MAE of 4.12, suggesting slightly larger errors. Based on these results, the XGBoost Regressor is chosen as the preferred algorithm for score prediction due to its higher accuracy and lower MAE than the Random Forest Regressor. Opting for the Random Forest Regressor would result in slightly lower accuracy and higher prediction errors.

* **T20 Blast:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\t20 blast\t20_blast_score_prediction_graph.PNG |

Figure 53. Predicted Values vs Accurate Values Graph of T20 Blast score prediction models

Table IX. Evaluation table of T20 Blast score prediction models

|  |  |  |  |
| --- | --- | --- | --- |
| **Xgboost Regressor** | | **Random Forest Regressor** | |
| R2 Score | MAE | R2 Score | MAE |
| 0.97 | 4.49 | 0.96 | 6.18 |

The provided graphs illustrate the relationship between the predicted values and the accurate values for score prediction; Fig. 53. One graph represents the predictions made by the XGBoost Regressor, while the other graph represents the predictions made by the Random Forest Regressor. Both graphs show that the predicted values align closely with the accurate values, indicating the models' ability to capture the underlying patterns in the data. Furthermore, an evaluation table is provided, which includes the R2 score and mean absolute error (MAE) for both algorithms. Upon analysis, the XGBoost Regressor achieved an R2 score of 0.97 and an MAE of 4.49, indicating strong correlation and low prediction errors. The Random Forest Regressor obtained an R2 score of 0.96 and an MAE of 6.18, suggesting slightly larger errors. Based on these results, the XGBoost Regressor is chosen as the preferred algorithm for score prediction due to its higher accuracy and lower MAE than the Random Forest Regressor. Opting for the Random Forest Regressor would result in slightly lower accuracy and higher prediction errors.

* **Men’s T20i:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\t20 men's\t20_men_score_prediction_comparison_graph.PNG |

Figure 54. Predicted Values vs Accurate Values Graph of Men’s T20i score prediction models

Table X. Evaluation table of Men’s T20i score prediction models

|  |  |  |  |
| --- | --- | --- | --- |
| **Xgboost Regressor** | | **Random Forest Regressor** | |
| R2 Score | MAE | R2 Score | MAE |
| 0.98 | 2.88 | 0.97 | 4.20 |

The provided graphs illustrate the relationship between the predicted values and the accurate values for score prediction; Fig. 54. One graph represents the predictions made by the XGBoost Regressor, while the other graph represents the predictions made by the Random Forest Regressor. Both graphs show that the predicted values align closely with the accurate values, indicating the models' ability to capture the underlying patterns in the data. Furthermore, an evaluation table is provided, which includes the R2 score and mean absolute error (MAE) for both algorithms. Upon analysis, the XGBoost Regressor achieved an R2 score of 0.98 and an MAE of 2.88, indicating strong correlation and low prediction errors. The Random Forest Regressor obtained an R2 score of 0.97 and an MAE of 4.20, suggesting slightly larger errors. Based on these results, the XGBoost Regressor is chosen as the preferred algorithm for score prediction due to its higher accuracy and lower MAE than the Random Forest Regressor. Opting for the Random Forest Regressor would result in slightly lower accuracy and higher prediction errors.

* **Women’s T20i:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\t20 women's\t20_women_score_prediction_graphical_view.PNG |

Figure 55. Predicted Values vs Accurate Values Graph of Women’s T20i score prediction models

Table XI. Evaluation table of Women’s T20i score prediction models

|  |  |  |  |
| --- | --- | --- | --- |
| **Xgboost Regressor** | | **Random Forest Regressor** | |
| R2 Score | MAE | R2 Score | MAE |
| 0.98 | 2.12 | 0.98 | 2.79 |

The provided graphs showcase the predicted values versus the accurate values for a score prediction, with one diagram representing the predictions made by the XGBoost Regressor and the other graph illustrating the predictions made by the Random Forest Regressor; Fig. 55. Both graphs align closely with the predicted and accurate values, indicating the models' proficiency in capturing the underlying data patterns. Evaluating the metrics, the XGBoost Regressor and the Random Forest Regressor achieve the same R2 score of 0.98, indicating a strong correlation between the predicted and accurate values. However, when considering the mean absolute error (MAE), the XGBoost Regressor has a lower MAE of 2.12, while the Random Forest Regressor has a slightly higher MAE of 2.79. If the Random Forest Regressor is selected instead, there would be a trade-off between accuracy and prediction errors, as it demonstrates a slightly higher MAE than the XGBoost Regressor. This may result in less precise predictions and potentially impact the reliability of the score predictions. Therefore, based on the lower MAE and higher precision, the XGBoost Regressor is selected for this preferred algorithm for score prediction.

* **WBBL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\wbbl\wbl_score_graph.PNG |

Figure 56. Predicted Values vs Accurate Values Graph of WBBL score prediction models

Table XII. Evaluation table of WBBL score prediction models

|  |  |  |  |
| --- | --- | --- | --- |
| **Xgboost Regressor** | | **Random Forest Regressor** | |
| R2 Score | MAE | R2 Score | MAE |
| 0.97 | 3.42 | 0.96 | 4.36 |

The provided graphs illustrate the relationship between the predicted values and the accurate values for score prediction; Fig. 56. One graph represents the predictions made by the XGBoost Regressor, while the other graph represents the predictions made by the Random Forest Regressor. Both graphs show that the predicted values align closely with the accurate values, indicating the models' ability to capture the underlying patterns in the data. Furthermore, an evaluation table is provided, which includes the R2 score and mean absolute error (MAE) for both algorithms. Upon analysis, the XGBoost Regressor achieved an R2 score of 0.97 and an MAE of 3.42, indicating strong correlation and low prediction errors. The Random Forest Regressor obtained an R2 score of 0.96 and an MAE of 4.36, suggesting slightly larger errors. Based on these results, the XGBoost Regressor is chosen as the preferred algorithm for score prediction due to its higher accuracy and lower MAE than the Random Forest Regressor. Opting for the Random Forest Regressor would result in slightly lower accuracy and higher prediction errors.

* **Men’s Super Smash:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\super smash men's\ssm_score_prediction_graph.PNG |

Figure 57. Predicted Values vs Accurate Values Graph of Men’s Super Smash score prediction models

Table XIII. Evaluation table of Men’s Super Smash score prediction models

|  |  |  |  |
| --- | --- | --- | --- |
| **Xgboost Regressor** | | **Random Forest Regressor** | |
| R2 Score | MAE | R2 Score | MAE |
| 0.98 | 3.75 | 0.97 | 4.55 |

The provided graphs illustrate the relationship between the predicted values and the accurate values for score prediction; Fig. 57. One graph represents the predictions made by the XGBoost Regressor, while the other graph represents the predictions made by the Random Forest Regressor. Both graphs show that the predicted values align closely with the accurate values, indicating the models' ability to capture the underlying patterns in the data. Furthermore, an evaluation table is provided, which includes the R2 score and mean absolute error (MAE) for both algorithms. Upon analysis, the XGBoost Regressor achieved an R2 score of 0.98 and an MAE of 3.75, indicating strong correlation and low prediction errors. The Random Forest Regressor obtained an R2 score of 0.97 and an MAE of 4.55, suggesting slightly larger errors. Based on these results, the XGBoost Regressor is chosen as the preferred algorithm for score prediction due to its higher accuracy and lower MAE than the Random Forest Regressor. Opting for the Random Forest Regressor would result in slightly lower accuracy and higher prediction errors.

* **Women’s Super Smash:**

|  |
| --- |
| **E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\super smash women's\ssf_score_prediction_graph.PNG** |

Figure 58. Predicted Values vs Accurate Values Graph of Women’s Super Smash score prediction models

Table XIV. Evaluation table of Women’s Super Smash score prediction models

|  |  |  |  |
| --- | --- | --- | --- |
| **Xgboost Regressor** | | **Random Forest Regressor** | |
| R2 Score | MAE | R2 Score | MAE |
| 0.98 | 2.14 | 0.98 | 2.38 |

The provided graphs showcase the predicted values versus the accurate values for a score prediction, with one diagram representing the predictions made by the XGBoost Regressor and the other graph illustrating the predictions made by the Random Forest Regressor; Fig. 58. Both graphs align closely with the predicted and accurate values, indicating the models' proficiency in capturing the underlying data patterns. Evaluating the metrics, the XGBoost Regressor and the Random Forest Regressor achieve the same R2 score of 0.98, indicating a strong correlation between the predicted and accurate values. However, when considering the mean absolute error (MAE), the XGBoost Regressor has a lower MAE of 2.14, while the Random Forest Regressor has a slightly higher MAE of 2.38. If the Random Forest Regressor is selected instead, there would be a trade-off between accuracy and prediction errors, as it demonstrates a slightly higher MAE than the XGBoost Regressor. This may result in less precise predictions and potentially impact the reliability of the score predictions. Therefore, based on the lower MAE and higher precision, the XGBoost Regressor is selected for this preferred algorithm for score prediction.

## First Innings Win Prediction Models Result Analysis

To evaluate every tournament first innings win prediction models, actual values vs predicted values graph of confusion matrix has been plotted and accuracy score has been calculated.

Table XV. Accuracy Score of first innings win model of all tournaments

|  |  |
| --- | --- |
| Tournament Name | Accuracy Score |
| BBL | 0.76 |
| BPL | 0.85 |
| CPL | 0.79 |
| CSA T20 | 0.81 |
| IPL | 0.77 |
| LPL | 0.84 |
| PSL | 0.76 |
| T20 Blast | 0.74 |
| Men’s T20i | 0.79 |
| Women’s T20i | 0.87 |
| WBBL | 0.74 |
| Men’s Super Smash | 0.78 |
| Women’s Super Smash | 0.87 |

The table provides the accuracy scores of the first innings win model for various cricket tournaments. Each row represents a specific tournament, and the corresponding accuracy score is listed. The scores range from 0.74 to 0.87, indicating the model's performance in predicting the first innings winner. The Women's T20i and Women’s Super Smash tournament has the highest accuracy score of 0.87, demonstrating the model's predictive solid capabilities in women's T20 international matches. The BPL and LPL tournaments also exhibit notable accuracy scores of 0.85 and 0.84, respectively. Conversely, the T20 Blast and PSL display slightly lower accuracy scores of 0.74 and 0.76, respectively. These accuracy scores provide valuable insights for cricket enthusiasts and teams, aiding them in making informed decisions based on the model's predictions regarding the first innings winner in different cricket tournaments.

* **BBL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\bbl\bbl_first_innings_win_graph.PNG |

Figure 59. BBL first innings win prediction model confusion matrix

The model achieved an accuracy score of 0.76, correctly classifying 76% of the instances. It accurately predicted "Win" in 35.89% of cases and "Lose" in 39.61% of cases. However, it struggled in certain instances, misclassifying 13.51% of "Lose" instances as "Win" and 10.99% of "Win" instances as "Lose".

* **BPL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\bpl\bpl_1st_innings_win_graph.PNG |

Figure 60. BPL first innings win prediction model confusion matrix

The model achieved an accuracy score of 0.85, correctly classifying 85% of the instances. It accurately predicted "Win" in 10.31% of cases and "Lose" in 74.77% of cases. However, it struggled in certain instances, misclassifying 9.80% of "Lose" instances as "Win" and 5.12% of "Win" instances as "Lose".

* **CPL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\cpl\cpl_first_innings_win_graph.PNG |

Figure 61. CPL first innings win prediction model confusion matrix

The model achieved an accuracy score of 0.79, correctly classifying 79% of the instances. It accurately predicted "Win" in 18.90% of cases and "Lose" in 59.86% of cases. However, it struggled in certain instances, misclassifying 12.94% of "Lose" instances as "Win" and 8.30% of "Win" instances as "Lose".

* **CSA T20:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\csa t20 challenge\csa_t20_first_innings_win_graph.PNG |

Figure 62. CSA T20 first innings win prediction model confusion matrix

The model achieved an accuracy score of 0.81, correctly classifying 81% of the instances. It accurately predicted "Win" in 27.20% of cases and "Lose" in 53.61% of cases. However, it struggled in certain instances, misclassifying 11.63% of "Lose" instances as "Win" and 7.56% of "Win" instances as "Lose".

* **IPL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\ipl\ipl_first_innings_win_graph.PNG |

Figure 63. IPL first innings win prediction model confusion matrix

The model achieved an accuracy score of 0.77, correctly classifying 77% of the instances. It accurately predicted "Win" in 23.25% of cases and "Lose" in 53.88% of cases. However, it struggled in certain instances, misclassifying 13.19% of "Lose" instances as "Win" and 9.67% of "Win" instances as "Lose".

* **LPL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\lpl\lpl_first_innings_win_graph.PNG |

Figure 64. LPL first innings win prediction model confusion matrix

The model achieved an accuracy score of 0.84, correctly classifying 84% of the instances. It accurately predicted "Win" in 26.81% of cases and "Lose" in 56.95% of cases. However, it struggled in certain instances, misclassifying 10.21% of "Lose" instances as "Win" and 6.03% of "Win" instances as "Lose".

* **PSL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\psl\psl_1st_innings_win_graph.PNG |

Figure 65. PSL first innings win prediction model confusion matrix

The model achieved an accuracy score of 0.76, correctly classifying 76% of the instances. It accurately predicted "Win" in 33.76% of cases and "Lose" in 41.88% of cases. However, it struggled in certain instances, misclassifying 12.01% of "Lose" instances as "Win" and 12.35% of "Win" instances as "Lose".

* **T20 Blast:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\t20 blast\t20_blast_first_innings_win_graph.PNG |

Figure 66. T20 Blast first innings win prediction model confusion matrix

The model achieved an accuracy score of 0.74, correctly classifying 74% of the instances. It accurately predicted "Win" in 40.53% of cases and "Lose" in 33.16% of cases. However, it struggled in certain instances, misclassifying 12.56% of "Lose" instances as "Win" and 13.75% of "Win" instances as "Lose".

* **Men’s T20i:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\t20 men's\t20_men_first_innings_win_logistic_graph.PNG |

Figure 67. Men’s T20i first innings win prediction model confusion matrix

The model achieved an accuracy score of 0.79, correctly classifying 79% of the instances. It accurately predicted "Win" in 41.85% of cases and "Lose" in 37.52% of cases. However, it struggled in certain instances, misclassifying 9.85% of "Lose" instances as "Win" and 10.77% of "Win" instances as "Lose".

* **Women’s T20i:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\t20 women's\t20_women_first_innings_win_graph.PNG |

Figure 68. Women’s T20i first innings win prediction model confusion matrix

The model achieved an accuracy score of 0.87, correctly classifying 87% of the instances. It accurately predicted "Win" in 49.75% of cases and "Lose" in 36.87% of cases. However, it struggled in certain instances, misclassifying 6.08% of "Lose" instances as "Win" and 7.30% of "Win" instances as "Lose".

* **WBBL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\wbbl\wbl_first_innings_win_graph.PNG |

Figure 69. WBBL first innings win prediction model confusion matrix

The model achieved an accuracy score of 0.74, correctly classifying 74% of the instances. It accurately predicted "Win" in 31.71% of cases and "Lose" in 42.31% of cases. However, it struggled in certain instances, misclassifying 13.44% of "Lose" instances as "Win" and 12.53% of "Win" instances as "Lose".

* **Men’s Super Smash:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\super smash men's\ssm_1st_innings_win_graph.PNG |

Figure 70. Men’s Super Smash first innings win prediction model confusion matrix

The model achieved an accuracy score of 0.78, correctly classifying 78% of the instances. It accurately predicted "Win" in 32.44% of cases and "Lose" in 45.61% of cases. However, it struggled in certain instances, misclassifying 10.63% of "Lose" instances as "Win" and 11.32% of "Win" instances as "Lose".

* **Women’s Super Smash:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\super smash women's\ssf_1st_innings_win_graph.PNG |

Figure 71. Women’s Super Smash first innings win prediction model confusion matrix

The model achieved an accuracy score of 0.87, correctly classifying 87% of the instances. It accurately predicted "Win" in 48.28% of cases and "Lose" in 38.32% of cases. However, it struggled in certain instances, misclassifying 4.81% of "Lose" instances as "Win" and 8.59% of "Win" instances as "Lose".

## Second Innings Win Prediction Models Result Analysis

To evaluate every tournament second innings win prediction models, actual values vs predicted values graph of confusion matrix has been plotted and accuracy score has been calculated.

Table XVI. Accuracy Score of Second innings win model of all tournaments

|  |  |
| --- | --- |
| Tournament Name | Accuracy Score |
| BBL | 0.84 |
| BPL | 0.84 |
| CPL | 0.80 |
| CSA T20 | 0.82 |
| IPL | 0.81 |
| LPL | 0.87 |
| PSL | 0.83 |
| T20 Blast | 0.83 |
| Men’s T20i | 0.84 |
| Women’s T20i | 0.90 |
| WBBL | 0.84 |
| Men’s Super Smash | 0.84 |
| Women’s Super Smash | 0.89 |

The table presents the accuracy scores of the second innings win model for various cricket tournaments. Each row corresponds to a specific tournament and provides the corresponding accuracy score. The accuracy scores range from 0.80 to 0.90, indicating the model's performance in predicting the winner of the second innings. The Women's T20i tournament has the highest accuracy score of 0.90, indicating the model's predictive solid capabilities in women's T20 international matches. The Women’s Super Smash and LPL also demonstrate a notable accuracy score of 0.89 and 0.87 respectively, while the BBL, BPL, Men’s Super Smash and Men's T20i exhibit consistent accuracy scores 0.84. The remaining tournaments, including the CPL, CSA T20, IPL, PSL, T20 Blast and WBBL, showcase accuracy scores ranging from 0.80 to 0.83. These accuracy scores provide valuable insights for cricket enthusiasts and teams, enabling them to make informed decisions based on the model's predictions regarding the winner of the second innings in different cricket tournaments.

* **BBL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\bbl\bbl_second_innings_win_graph.PNG |

Figure 72. BBL second innings win prediction model confusion matrix

The model achieved an accuracy score of 0.84, correctly classifying 84% of the instances. It accurately predicted "Win" in 40.15% of cases and "Lose" in 43.57% of cases. However, it struggled in certain instances, misclassifying 8.57% of "Lose" instances as "Win" and 7.72% of "Win" instances as "Lose".

* **BPL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\bpl\bpl_2nd_innings_win_graph.PNG |

Figure 73. BPL second innings win prediction model confusion matrix

The model achieved an accuracy score of 0.84, correctly classifying 84% of the instances. It accurately predicted "Win" in 15.11% of cases and "Lose" in 68.40% of cases. However, it struggled in certain instances, misclassifying 11.26% of "Lose" instances as "Win" and 5.23% of "Win" instances as "Lose."

* **CPL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\cpl\cpl_second_innings_win_graph.PNG |

Figure 74. CPL second innings win prediction model confusion matrix

The model achieved an accuracy score of 0.80, correctly classifying 80% of the instances. It accurately predicted "Win" in 31.30% of cases and "Lose" in 48.68% of cases. However, it struggled in certain instances, misclassifying 9.71% of "Lose" instances as "Win" and 10.31% of "Win" instances as "Lose".

* **CSA T20:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\csa t20 challenge\csa_t20_second_innings_win_graph.PNG |

Figure 75. CSA T20 second innings win prediction model confusion matrix

The model achieved an accuracy score of 0.82, correctly classifying 82% of the instances. It accurately predicted "Win" in 34.00% of cases and "Lose" in 48.42% of cases. However, it struggled in certain instances, misclassifying 9.58% of "Lose" instances as "Win" and 8.01% of "Win" instances as "Lose".

* **IPL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\ipl\ipl_second_innings_win_graph.PNG |

Figure 76. IPL second innings win prediction model confusion matrix

The model achieved an accuracy score of 0.81, correctly classifying 81% of the instances. It accurately predicted "Win" in 31.15% of cases and "Lose" in 49.76% of cases. However, it struggled in certain instances, misclassifying 9.78% of "Lose" instances as "Win" and 9.31% of "Win" instances as "Lose".

* **LPL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\lpl\lpl_second_innings_win_graph.PNG |

Figure 77. LPL second innings win prediction model confusion matrix

The model achieved an accuracy score of 0.87, correctly classifying 87% of the instances. It accurately predicted "Win" in 21.74% of cases and "Lose" in 65.22% of cases. However, it struggled in certain instances, misclassifying 6.93% of "Lose" instances as "Win" and 6.11% of "Win" instances as "Lose".

* **PSL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\psl\psl_2nd_innings_win_graph.PNG |

Figure 78. PSL second innings win prediction model confusion matrix

The model achieved an accuracy score of 0.83, correctly classifying 83% of the instances. It accurately predicted "Win" in 46.68% of cases and "Lose" in 35.12% of cases. However, it struggled in certain instances, misclassifying 8.82% of "Lose" instances as "Win" and 9.37% of "Win" instances as "Lose".

* **T20 Blast:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\t20 blast\t20_blast_second_innings_win_graph.PNG |

Figure 79. T20 Blast second innings win prediction model confusion matrix

The model achieved an accuracy score of 0.83, correctly classifying 83% of the instances. It accurately predicted "Win" in 37.09% of cases and "Lose" in 46.03% of cases. However, it struggled in certain instances, misclassifying 8.45% of "Lose" instances as "Win" and 8.43% of "Win" instances as "Lose".

* **Men’s T20i:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\t20 men's\t20_men_second_innings_win_graph.PNG |

Figure 80. Men’s T20i second innings win prediction model confusion matrix

The model achieved an accuracy score of 0.84, correctly classifying 84% of the instances. It accurately predicted "Win" in 41.85% of cases and "Lose" in 37.52% of cases. However, it struggled in certain instances, misclassifying 9.85% of "Lose" instances as "Win" and 10.77% of "Win" instances as "Lose".

* **Women’s T20i:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\t20 women's\t20_women_second_innings_win_graph.PNG |

Figure 81. Women’s T20i second innings win prediction model confusion matrix

The model achieved an accuracy score of 0.90, correctly classifying 90% of the instances. It accurately predicted "Win" in 34.76% of cases and "Lose" in 54.87% of cases. However, it struggled in certain instances, misclassifying 5.27% of "Lose" instances as "Win" and 5.09% of "Win" instances as "Lose".

* **WBBL:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\wbbl\wbl_second_innings_win_graph.PNG |

Figure 82. WBBL second innings win prediction model confusion matrix

The model achieved an accuracy score of 0.84, correctly classifying 84% of the instances. It accurately predicted "Win" in 44.93% of cases and "Lose" in 39.15% of cases. However, it struggled in certain instances, misclassifying 7.70% of "Lose" instances as "Win" and 8.22% of "Win" instances as "Lose".

* **Men’s Super Smash:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\super smash men's\ssm_2nd_innings_win_graph.PNG |

Figure 83. Men’s Super Smash second innings win prediction model confusion matrix

The model achieved an accuracy score of 0.84, correctly classifying 84% of the instances. It accurately predicted "Win" in 48.47% of cases and "Lose" in 35.75% of cases. However, it struggled in certain instances, misclassifying 8.43% of "Lose" instances as "Win" and 7.35% of "Win" instances as "Lose".

* **Women’s Super Smash:**

|  |
| --- |
| E:\CSE499 CSE498R\T20 ML Project CSE499\Report\Work Data Methodology\super smash women's\ssf_2nd_innings_win_graph.PNG |

Figure 84. Women’s Super Smash second innings win prediction model confusion matrix

The model achieved an accuracy score of 0.89, correctly classifying 89% of the instances. It accurately predicted "Win" in 36.46% of cases and "Lose" in 52.76% of cases. However, it struggled in certain instances, misclassifying 5.53% of "Lose" instances as "Win" and 5.25% of "Win" instances as "Lose".

# **Chapter 5 Impacts of the Project**

## Impact of this project on societal, health, safety, legal and cultural issues

### Societal Impact

The initiative can increase fan involvement by offering precise score projections and victory likelihoods. This can improve T20 cricket enthusiasm and interest, resulting in a more passionate and involved fan base. Using machine learning to estimate scores and the likelihood of winning can make cricket more accessible to a wider audience. Fans unfamiliar with the game's complexities may still enjoy the matches with a greater grasp of the potential results. The initiative can promote data-driven debates between fans, analysts, and specialists, creating better game knowledge and encouraging informed debate.

### Health Impact

Accurate forecasts can improve fans' emotional well-being by increasing their delight and excitement while watching T20 cricket matches. By offering insights into expected scores and winning chances, the initiative can lower stress levels for fans who are involved in the result of the game, allowing them to have reasonable expectations.

### Safety Impact

It is critical to guarantee that the project forecasts do not jeopardize the integrity and fairness of T20 cricket matches. To protect the spirit of fair play, precautions must be taken to avoid any efforts at match-fixing or manipulation based on predictions. Cyber security should be prioritized in the project to secure the integrity of the data used for predictions and to prevent any unwanted access or manipulation, ensuring the system's and users' information's safety.

### Legal Impact

The project must adhere to all applicable laws and rules governing data protection, privacy, and intellectual property rights. It should ensure that suitable consent and licensing agreements exist for data usage and copyrighted works.

### Cultural Impact

The project's forecasts contribute to the game's emerging culture by adding a new degree of analysis and comprehension. It fosters a culture of data-driven insights and dialogues among fans and cricket lovers, encouraging greater participation and appreciation. The research demonstrates the sport's capacity to adapt and absorb contemporary technology by embracing machine learning for cricket forecasts. This illustrates cricket's progressive spirit, emphasizing its eagerness to adopt developments for the benefit of the cricketing community.

## Impact of this project on environmental and sustainability

The initiative has the potential to eliminate the need for needless travel by properly forecasting match results and scores. Fans who rely on in-person attendance may be more likely to make educated selections about which games to attend, lowering their carbon footprint associated with travel.

Accurate forecasting can aid in the optimization of resource allocation in cricket stadiums. Organizers can better plan crowd control, seating arrangements, and other operational factors by forecasting projected scores and victory probabilities. This can result in more efficient resource utilization, such as electricity, water, and trash management.

The availability of accurate forecasts might encourage people to watch matches online instead of attending in person. This transition has the potential to minimize carbon emissions connected with stadium travel while also encouraging the use of digital platforms, which frequently have a reduced environmental impact.

The project's use of machine learning and data analysis may encourage other companies to take similar approaches to sustainability programs. Cricket companies may lower their environmental footprint by employing data-driven insights, such as improving stadium operations, deploying renewable energy solutions, and encouraging recycling and waste management initiatives.

By incorporating technology and data analysis into cricket, the initiative could promote environmental consciousness among cricket fans and the larger community. This increasing understanding can lead to greater attention being paid to sustainability in other parts of cricket, such as stadium design, player transportation, and event administration.

# **Chapter 6 Project Planning and Budget**

## Gantt Chart

|  |
| --- |
| C:\Users\SADI\Downloads\355862457_999922677684275_1784405733702921215_n.png |

Figure 85. Gantt Chart.

## Budget

### Hardware

* High-performance computer or devices: $1,500 - $3,000
* Additional storage: $100 - $200



### Software

* Python (free)
* Django (free)
* Flutter (free)
* Machine learning libraries (free)
* Database (free)
* Integrated development environment (free)

### Data

* T20 Cricket match datasets: Free

### Hosting and Deployment

* Web hosting server: $5 - $50 per month
* Domain name registration: $10 - $20 per year

# **Chapter 7 Complex Engineering Problems and Activities**

## Complex Engineering Problems (CEP)

Table XV demonstrates complex engineering problem attribute.

Table XVII. Complex Engineering Problem Attributes

|  |  |  |
| --- | --- | --- |
| **Attributes** | | **Addressing the complex engineering problems (P) in the project** |
| P1 | Depth of knowledge required (K5-K8) | The project requires knowledge of Cricket Rules (K5), Cricket Formats (K5), Data Collection and Pre-processing (K7), Feature Engineering, Ensemble Methods, Model Interpretability (K6), Machine Learning Algorithms, Transfer Learning, and Big Data Processing (K7), Statistical Methods, Predictive Modeling, Research, and Innovation (K8). |
| P2 | Range of conflicting requirements | Prediction accuracy must be balanced against the complexity of the machine learning models used. Data availability is critical for accurate forecasts, but ensuring data quality and reliability can be difficult. Another conflicting goal is to achieve generalization in the models while avoiding over-fitting to specific datasets. Real-time updates to improve forecasts must be balanced against the time necessary for model training. Scalability must be balanced with resource limits when dealing with large amounts of data and user demand. |
| P3 | Depth of analysis required | Depth of analysis needed to select appropriate machine learning algorithms, train and evaluate models, integrate real-time data updates, ensure model interpretability, and optimize performance. |
| P4 | Familiarity of issues | Learn the rules and formats of T20 International Cricket and T20 Club Cricket matches; the process of cricket data collection; fundamental machine learning concepts; the Django and Flutter frameworks; and how to evaluate performance metrics. |
| P5 | Extent of applicable codes | To implement the project, relevant codes include Python for machine learning tasks, Django for web application development, Flutter for mobile application interfaces, machine learning libraries, API integration for real-time data updates, and database management. |
| P6 | Extent of stakeholder involvement | The project team, clients or organizations, end users, cricket experts, data providers, business or marketing teams, and regulatory authorities. |
| P7 | Interdependence | The project's components are interdependent, including the availability of correct data, effective machine learning algorithms, team communication, user feedback, and smooth integration and deployment. |

## Complex Engineering Activities (CEA)

Table XVI demonstrates complex engineering activities.

Table XVIII. Complex Engineering Activities

|  |  |  |
| --- | --- | --- |
| **Attributes** | | **Addressing the complex engineering activities (A) in the project** |
| A1 | Range of resources | The project requires human resources, data resources, computing resources, and software resources for its successful execution. |
| A2 | Level of interactions | Interactions between the internal team, clients or project owners, stakeholders, cross-functional teams, and external entities are required to guarantee successful collaboration, alignment, feedback collection, and compliance. |
| A3 | Innovation | The initiative supports creativity by embracing innovative approaches and technology such as machine learning algorithms, Django, and Flutter frameworks. These enhancements simplify operations, increase efficiency, and enhance the user experience, propelling cricket prediction to new heights. |
| A4 | Consequences for society  / Environment | The project has far-reaching implications for society and the environment. It makes accurate forecasts for cricket matches, boosting the spectator experience. The project's ingenuity and efficient engineering procedures also help to conserve resources and ensure environmental sustainability. |
| A5 | Familiarity | This project requires familiarity with T20 cricket rules, data collection, machine learning techniques, and the Django and Flutter frameworks. Understanding match forms, gathering pertinent data, and employing machine learning approaches. |

# **Chapter 8 Conclusion**

## Summary

The project aims to develop a machine-learning-based system that can forecast the predictive outcomes of T20 International Cricket matches and T20 Club Cricket matches. The system utilizes a combination of Machine Learning, Django, and Flutter technologies to gather and analyze historical cricket match data, including player performance, team statistics, and match conditions. By training the machine learning model on this data, the system can make predictions about future matches, taking into account various factors that may influence the outcome, such as player form, team composition, pitch conditions, and weather. The Django framework is used for building the backend, which handles data processing, model training, and prediction generation. On the other hand, the Flutter framework is employed to create a user-friendly front-end interface where users can input match details and receive the predicted outcomes. This project ultimately provides a tool that can assist cricket enthusiasts, fans, and analysts in making informed predictions about T20 cricket matches.

## Limitations

The project "Forecasting predictive outcomes of the T20 International Cricket and T20 Club Cricket Competitions Matches Using Machine-Learning Django Flutter" may have several limitations, including:

* **Model Training:** Developing accurate machine learning models for sports predictions requires extensive training on relevant data. The success of the project will depend on the availability of a comprehensive and diverse dataset specific to T20 cricket matches. If the training data is limited or biased, it may lead to inaccurate predictions.
* **Dynamic Nature of Cricket:** Cricket is a complex and dynamic sport with various factors influencing match outcomes, including player form, weather’s conditions, pitch conditions, team strategies, and unforeseen events. Incorporating all these variables into a predictive model can be challenging, and the model may struggle to account for such factors accurately.
* **Human Factors:** Cricket is a game played by individuals, and the performance of players can vary significantly from match to match. Injuries, changes in form, and other unpredictable human factors can impact match outcomes. While machine learning algorithms can analyze historical data, they may struggle to account for these human variables effectively.
* **Model Transparency and Explainability:** Machine learning models can be complex and difficult to interpret, especially when using machine learning techniques. The lack of transparency and explainability may make it challenging to understand how the model arrives at its predictions. This can be a limitation when trying to gain insights or validate the model's accuracy.
* **Overfitting and Bias:** Overfitting occurs when a model is too closely tailored to the training data and fails to generalize well to new data. It is essential to ensure the model doesn't overfit the training data, as it can lead to misleading predictions. Additionally, bias in the data or model can result in unfair predictions or reinforce existing biases in cricket analysis.
* **Ethical Considerations:** Predictive models in sports can have significant implications, such as affecting betting markets, team strategies, and fan sentiments. It is essential to consider the ethical implications of the project, such as responsible use of predictions, privacy concerns, and potential impacts on the integrity of the game.
* **Unforeseen Events:** Cricket matches can have unexpected occurrences, such as rain interruptions, player injuries, or disciplinary actions. These events can significantly alter the course and outcome of a match, making it challenging for any predictive model to accurately account for them.
* **User Acceptance:** The success of a predictive model also depends on user acceptance and adoption. Cricket enthusiasts, coaches, and players may have differing opinions about the reliability of machine learning predictions, and their acceptance and willingness to use the predictions may vary.

It's important to consider these limitations and potential challenges when developing and implementing the project to manage expectations and ensure accurate and responsible use of predictive models in cricket forecasting.

## Future Improvement

The project "Forecasting predictive outcomes of the T20 International Cricket and T20 Club Cricket Competitions Matches Using Machine-Learning Django Flutter" has the potential for several future improvements. Here are some possible areas where enhancements can be made:

* **Enhanced Prediction Models:** The project can explore more advanced machine learning algorithms and techniques to improve the accuracy of match outcome predictions. This could involve using ensemble methods, deep learning models, or incorporating additional data sources such as player performance statistics, weather conditions, or pitch characteristics.
* **Real-Time Data Integration:** To make the predictions more accurate, the project can consider integrating real-time data feeds during the matches. This can include live ball-by-ball updates, player performance metrics, and other relevant information. By incorporating real-time data, the models can adapt and adjust their predictions dynamically, taking into account the current match situation.
* **User Feedback and Refinement:** Gathering user feedback and incorporating it into the prediction models can help in continuously improving their performance. The project can implement mechanisms to collect feedback from users about the accuracy of predictions and incorporate this feedback into the learning process to refine the models over time.
* **User-Friendly Interfaces:** The project can focus on developing user-friendly interfaces for different platforms, such as web and mobile applications, using technologies like Django and Flutter. These interfaces should provide intuitive features for users to explore match predictions, view historical performance, and interact with the system effectively.
* **Integration with Betting Platforms:** If legally permissible and ethically responsible, the project can explore integration with betting platforms to provide predictions for users interested in betting on cricket matches. This would require complying with legal regulations and ensuring responsible gambling practices.
* **Expand to Other Cricket Formats:** While the current project focuses on T20 cricket, future improvements could involve expanding the predictive models to cover other formats such as One-Day Internationals (ODIs) or Test matches. This would require incorporating different sets of features and adjusting the models accordingly.
* **Collaboration with Cricket Experts:** Collaborating with cricket experts, coaches, or players can provide valuable insights into the game and help refine the prediction models. Expert opinions can be used to validate the predictions, incorporate domain knowledge, and improve the overall accuracy and reliability of the system.
* **Performance Monitoring and Optimization:** Continuous monitoring and optimization of the prediction models' performance can lead to better outcomes. Techniques such as model retraining, hyperparameter tuning, and performance benchmarking against other prediction systems can be employed to ensure the models remain effective and up to date.
* **Privacy and Security Considerations:** As the project collects and processes user data, it is important to prioritize privacy and security. Future improvements should focus on implementing robust data protection measures, complying with relevant data privacy regulations, and ensuring secure storage and transmission of sensitive user information.
* **Collaboration and Research:** The project can actively engage with the cricket community and academic researchers to foster collaboration and explore novel approaches to cricket match outcome prediction. This can include participating in research conferences, sharing data sets, and contributing to the wider research community

These improvements can collectively enhance the accuracy, usability, and reliability of the project, providing cricket enthusiasts with valuable insights and predictions for T20 cricket matches.

# **References**

1. Twenty20. Accessed on: May. 2, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Twenty20.
2. Official International Cricket Council Website. Accessed on: March. 2, 2023. [Online]. Available: <https://www.icc-cricket.com/>.
3. List of T20 cricket competitions. Accessed on: May. 23, 2023. [Online]. Available: https://en.wikipedia.org/wiki/List\_of\_T20\_cricket\_competitions.
4. Machine Learning. Accessed on: March. 4, 2023. [Online]. Available: <https://www.ibm.com/topics/machine-learning>
5. Tech Accelerator. Ed Burns. Machine Learning [Online]. Available from: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML
6. R. Lamsal and A. Choudhary, "Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning," arXiv preprint, arXiv: 1809.09813 [stat.AP], 2018. Available from: https://doi.org/10.48550/arXiv.1809.09813.
7. H. Barot, A. Kothari, P. Bide, B. Ahir and R. Kankaria, "Analysis and Prediction for the Indian Premier League," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-7, doi: 10.1109/INCET49848.2020.9153972.
8. R. Ahmed, P. Sareen, V. Kumar, R. Jain, P. Nagrath, A. Gupta, S. K. Chawla, “First inning score prediction of an IPL match using machine learning” INNOVATIONS IN COMPUTATIONAL AND COMPUTER TECHNIQUES: ICACCT-2021, doi: 10.1063/5.0108928.
9. V. Bhatnagar, D. Batham, “Estimating the Chances of Winning IPL Match using Machine Learning,“ May 2022, DOI: 10.37591/JoSETTT. Available: <https://www.researchgate.net/publication/360826700_Estimating_the_Chances_of_Winning_IPL_Match_using_Machine_Learning>.
10. E. Mundhe, I. Jain and S. Shah, "Live Cricket Score Prediction Web Application using Machine Learning," 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Pune, India, 2021, pp. 1-6, doi: 10.1109/SMARTGENCON51891.2021.9645855.
11. S. Agrawal, S. P. Singh and J. K. Sharma, "Predicting Results of Indian Premier League T-20 Matches using Machine Learning," 2018 8th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2018, pp. 67-71, doi: 10.1109/CSNT.2018.8820235.
12. K. R. Reddy, P. Sai A, Chandrakanth V, S. S. Reddy, “Super Predictor of Indian Premier League (IPL) using Various ML techniques with help of IBM Cloud,” International Journal for Research in Applied Science & Engineering Technology (IJRASET), June 2022, DOI: 10.22214/ijraset.2022.43654.
13. M. Awais, K. Das, “A Predictive Analysis of PSL Match Winner Using Machine Learning Techniques,“ April 2022, DOI: 10.32350/umtair.11.05. Available: <https://journals.umt.edu.pk/index.php/UMT-AIR/article/view/2198>.
14. K. C. Srikantaiah, A. Khetan, B. Kumar, D. Tolani, and H. Patel, “Prediction of IPL Match Outcome Using Machine Learning Techniques,” Proc. 3rd Int. Conf. Integr. Intell. Comput. Commun. Secur. (ICIIC 2021), 2021, doi:10.2991/ahis.k.210913.049.
15. M. Shetty, S. Rane, C. Pandita and S. Salvi, "Machine learning-based Selection of Optimal sports Team based on the Players Performance," 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2020, pp. 1267-1272, doi: 10.1109/ICCES48766.2020.9137891.
16. A. Basit, M. B. Alvi, F. H. Jaskani, M. Alvi, K. H. Memon and R. A. Shah, "ICC T20 Cricket World Cup 2020 Winner Prediction Using Machine Learning Techniques," 2020 IEEE 23rd International Multi topic Conference (INMIC), Bahawalpur, Pakistan, 2020, pp. 1-6, doi: 10.1109/INMIC50486.2020.9318077.
17. Arjun Singhvi, Ashish Shenoy, Shruthi Racha and Srinivas Tunuguntla. “Prediction of the outcome of a Twenty-20 Cricket Match,” 13 September, 2022, DOI:  10.48550/arXiv.2209.06346. Available: <https://arxiv.org/abs/2209.06346>.
18. K. Passi and N. Pandey, "Increased Prediction Accuracy in the Game of Cricket Using Machine Learning," Int. J. Data Min. Knowl. Manag. Process, vol. 8, no. 2, pp. 19-36, 2018, doi:  10.5121/ijdkp.2018.8203.
19. R. Ul Mustafa, M. S. Nawaz, M. I. Ullah Lali, T. Zia, and W. Mehmood, "Predicting The Cricket Match Outcome Using Crowd Opinions On Social Networks: A Comparative Study Of Machine Learning Methods," Malaysian Journal of Computer Science, 2017, DOI: 10.22452/mjcs.vol30no1.5. Available from: <https://doi.org/10.22452/mjcs.vol30no1.5>.
20. J. Kumar, R. Kumar and P. Kumar, "Outcome Prediction of ODI Cricket Matches using Decision Trees and MLP Networks," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018, pp. 343-347, doi: 10.1109/ICSCCC.2018.8703301.
21. Cricsheet. [Online] Available: <https://cricsheet.org/>.
22. Random Forest Algorithm [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2.
23. Xgboost Algorithm [Online]. Available: <https://www.nvidia.com/en-us/glossary/data-science/xgboost>.
24. Xgboost Regression [Online]. Available: <https://machinelearningmastery.com/xgboost-for-regression>.
25. Logistic Regression [Online]. Available: <https://www.ibm.com/topics/logistic-regression>.
26. R-Square Score [Online]. Available: https://www.geeksforgeeks.org/ml-r-squared-in-regression-analysis/.