Player Performance Prediction and Starting Eleven Selection Enhancer Using Machine Learning Algorithm For Cricket

Bhojani, Mirav Hirji
Student ID: 136446
BBIT 4.2C

An Information System Project II documentation submitted to the School of Computing and Engineering Sciences in partial fulfilment of the requirements for the award of the Bachelor's Degree in Business Information Technology of Strathmore University

School of Computing and Engineering Sciences
Strathmore University
Nairobi, Kenya.

December 2023

Declaration

I declare that this work has not been previously submitted and approved for the award of a Bachelor's degree by this or any other University. To the best of my knowledge and belief, the documentation contains no material previously published or written by another person except where due reference is made in the documentation itself.

Student's signature:

…….…….. [*Signature*]

…….…….. [*Date*]

Approval

The Information System Project 2 documentation of Mirav Hirji Bhojani was reviewed and approved *(for examination)* by: Mr. Kelvin Kibunja.

Supervisor's signature:

…….…….. [*Signature*]

…….…….. [*Date*]

Abstract

Kenya in the last two decades has seen a major downfall in cricket. The main reasons for this downfall are, corruption, poor management and outdated technology. Thus, a web application is proposed to minimize the effect from these factors. The proposed system will be used by Cricket Kenya to score all matches being played locally so as to maintain centralized data. The scores will then generate a leaderboards report that will update after every match and will be imported in the machine learning model where the prediction will be made based on the data in the leaderboards. Essentially it there will be two models that is the batting model and the bowling model. The batting model will predict the top 6 batsmen (players) while the bowling model will predict the top 5 bowlers (players). The system development followed the OOAD approach and using the waterfall with feedback methodology and with this the appropriate design diagram were drawn to facilitate development of the system. The major languages used to develop the system are php and python. Tools like VSCode and JupyterNotebook were used as well to fasten the development process. A lot of help was also taken from the libraries for python especially from SciKit to make the prediction model as accurate as possible. Once the systems were developed that is both the CodeIgniter project and the Prediction model then flask was used to deploy the machine learning model on the internet so the front-end user interface can be accessed and to make it easier use the system. Once all the system was ready and deployed then a few tests were carried out mostly black box tests to check if the system was doing as expected to. Finally, a user manual was written to make it easier for a new user to use the system.

Keywords: Cricket Kenya, Leaderboard report, OOAD, waterfall with feedback methodology, SciKit, CodeIgniter and Black box test.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

API – Application Programming Interface

ERD – Entity Relationship Diagram

ICC – International Cricket Council

IDE – Integrated Development Environment

NPCA – Nairobi Provincial Cricket Association

ODI – One Day Internationals

OOAD – Object Oriented Analysis and Design

SVR - Support Vector Regression

T20 – TwentyTwenty Format

T20I- TwentyTwenty International

Acknowledgements

As a student, I am overwhelmed with the help and support that I received in the process of creating this documentation, especially from my supervisor Mr. Kelvin Kibunja as well as the unit lecturer Mr. Titus Kipkosgei who upon approach at any given time have helped us with appropriate response that enabled me to finish my documentation document well before the deadline.

Moreover, I would like to thank Strathmore University and the School of Computing and Engineering Sciences which allowed me to express the skills and knowledge to come up with a system that could benefit the society.

Lastly, there was no viable way of completing the documentation document without the grace of God who gave me very supportive parents and guided me wherever needed despite running under busy schedules and gave me ideas to make my project unique. Finally, wanted to sincerely thank any other supporter that has unintentionally been left out in the acknowledgment.

Chapter 1: Introduction

1.1 Background

Kenya, a cricket playing country that had all the eyes of the best cricket playing countries. In the 1980's Kenya had seen growth in the game, such that the ICC (International Cricket Council) had so much faith in the country that they chose Kenya to host a World cup trophy in 1996 where Kenya managed to turn out as runners up(Roberts, 2019). Also, a World Cup knockout tournament in 2000's which is a big thing in cricket as the ICC wanted Kenya to gain for attraction and funds to develop the game in the country(Sen, 2018).

Moreover, in 2003 which was Kenya's dream year till date where Kenya managed to qualify for the final events(Matongera, 2021). In this process Kenya managed to beat the strongest teams of that time, namely the West Indies (which shocked the whole world), India and Sri-Lanka. However, after 2003 is when Kenya started to see the decline that it lost quite a few consecutive games till 2011 where ICC then had to snatch away the ODI status of the country (Gemmell, 2013)which made Kenya to pass through qualifiers which till date they have not even been able to qualify to get a spot in the world cup(Stoddart, 2017).

Currently, there are many players in the country who want to play Cricket as a profession but due to less scope, corruption, poor management, and outdated technology in the country for cricket the players opt to choose for other occupations for better pay and better opportunities and play cricket as a hobby on the side (Hideg, 2021).

From the above information it suggests that Cricket in Kenya is certainly declining, however there is a chance of reviving it with the help of a proposed system. The system will be able to provide up-to-date and relevant information that will help fight corruption, poor management and more so use the technology to analyse players well and make informed and reliable decisions to improve the outcome of the matches played and also when selecting the national team.

1.2 Problem Statement

The proposed system attempted to deal with the problems identified in order to help grow cricket in Kenya and any other regional countries that are playing in the division two or lower than that to enable the players to use the latest technology to improve the results of the game in favour of the team. The web application will do scoring of all domestic matches being played and store and maintain all the data(Hossen et al., 2023). The scoring was be done on match-to-match basis where the scores were updated after every match. Then after the season is over batsman scores per match are tallied up and total runs added up and a ranking leader board was be generated with the highest run scorer as number 1 and the rest follow. The same for the bowlers, the wickets per match were accumulative and were added at the end of the season with the highest wicket taker being ranked as first and the order follows. This system allowed the teams to have centralized data that is stored and managed by them(G et al., 2023).

Once the leader boards were generated, this report was downloaded and then used to predict the performance for the players in the upcoming matches so that the predictions can help select the best starting eleven for the next match that will benefit the team. This prediction was based on the players' previous data on previous 5 performances of that player and their batting average. This data was trained by checking the previous performances for the last five matches and checked the trend and gave an outcome. The players with the good performance prediction were selected in the starting eleven of the matches. The players in the starting eleven were predicted to perform well and if the prediction was accurate then the outcome of the match will be favourable to for the team (Singla & Shukla, 2020). This led to development of the game and of players as well and deal with the issue of corruption and management since the players in the leader boards will only be allowed to play for the national team.

However, the game has different formats and categories and due to time constraints only one of the formats will be able to be focused on and that is the 5 over format for the main men's team and using a web application. This is because 5 overs are the shortest format of the game and it still takes 1 and a half hour which already is a lot of time and thus the other formats are even longer and have more complexities with the rules. 5 overs format will be 5 overs a side and thus is preferred by most players(Maunder et al., 2017).

All in all, the proposed system was dealing with the issue of corruption, data storage and was helping in selecting the best possible starting eleven for a match being played. The issue of corruption was such that only the players in the leaderboards were selected in the starting eleven since the catch is that you cannot be performing well and still be playing for the national team. This was an issue overlooked by other systems. Moreover, most regional teams rely on external software or system to do their scoring and thus at time importing the data for analysis becomes a challenge and thus with the proposed system they will have the data stored with them and this leads to centralisation of data which definitely benefits the cricketing body since they can save time wasted on importing of data then inputting it in their system and analysing it which also could lead to typo errors while inputting. Moreover, the proposed system also helped in selecting the starting system based on the previous performances of the players in the squad. These were some of the issues that have been overlooked by other similar systems.

1.3 Aim

To design and develop a web application and prediction model that predicts player's performance to enhance the selection of starting eleven and deal with the corruption issues to revive cricket in Kenya.

1.4 Specific Objectives

i. To examine and understand the current player selection process.
ii. To investigate the challenges faced by the existing selection process.
iii. To review machine learning algorithms used in player selection.
iv. To design and develop a classification model that is going to help predict player performances for future matches.
v. To test the proposed system using unit and integration testing.

1.5 Research Questions

i. What is the player selection process?
ii. What are the challenges faced by using the existing process?
iii. What are the various machine learning algorithms that can be used to select players?
iv. How effective will the option of developing the model using a classification algorithm? If not, which other algorithm could be used to meet the system requirements?
v. What type of testing will be used to check on the validity of the system to be developed?

1.6 Justification

The current selection process of the starting eleven is based on a few assumptions and gut feelings and how close a player is to the coach(Agarwal et al., 2017). Also, the issues of corruption and use of outdated technology are still in place and also the collection of data is through external apps like CricHQ and CricHeroes which leads to inconsistency in the data collected(Saikia et al., 2019).

In order to bridge the issues brought up by the current processes the proposed system that deals with the issues of corruption, outdated technology and centralization of the data storage and all this will be achieved by developing an application that does scoring for all the matches being played in regional countries which at the end updates a leader board for top performers who are only the ones allowed to play for the national team unlike the players who initially were in the team and with poor recent performance. From the leader boards a 20-man squad is generated including 10 bowlers, 10 batsmen (Malek et al., 2018). The starting eleven will then be selected using a prediction model that will predict the players performance for future endeavours based on previous match performances and the relative features that is for a batsman it is batting average and the economy for the bowler. The players who are predicted with a verdict of good performance. This will both enhance the process of selection of the starting eleven and return a favourable outcome of the match which in the long run will benefit the game in the country and help in development of players skills (Lascu et al., 2020) and also be an opportunity for other players with natural talent be able to choose cricket as a profession unlike currently since there is no scope, people are fearing to take cricket as an option for their career paths(Philpott, 2018).

1.7 Scope

As there are many formats of cricket the proposed project will focus on the 5 over side format which is very interesting as well. Together with only matches being played locally that is in Kenya only. The reason for going with the 5-over format is because the data capturing will be quick and efficient with the resources that a university students have (Denscombe, 2017).

1.8 Delimitations and Limitations

1.8.1 Delimitation

Although countries like Kenya does not have the status to play the fifty-fifty format and the Test format, they still play these formats under the supervision of ICC however no points are given to Kenya for playing these matches(Premkumar et al., 2020). This is therefore a delimitation as there is then no point on focusing what the country does not play. This could be a project for the future when cricket in Kenya reaches its highest again and get the Test format status.

1.8.2 Limitation

The limitations of the project will be that the project will only be able to base the country to Kenya since Kenya has not played many international matches and the data for the international played are not available and thus the project will be limited to Kenyan boundaries.

Chapter 2: Literature Review

2.1 Introduction

This chapter explains all the current player selection processes and the challenges faced by the current problems and reviewing the existing systems to identify and learn the methods used by them, list out the gaps in these existing projects and what can be done to reduce the bridge between these gaps identified and finally the conceptual framework of the proposed system to suggest how the system will work and the roles of the major users of the system.

2.2 Current player selection process for the starting eleven

With more than 100 countries playing professional cricket (Harris, 2022), all the countries have to think well through the selection process for the starting eleven for the matches so as to get the best outcome for the match. Even after a lot of brainstorming, the reason why most of teams fail to choose the best combination for their starting eleven is due to lack of intelligence. Mostly, the selection is done by the captain, the coach and the selectors and sometimes the decision made on specific players does not come out as expected and because of this the outcomes of the match sometimes turns out to be unfavourable (Jhansi Rani et al., 2020).

Usually, the selection is based on experience, that is the most experienced players are given priority in their respective roles and after them follow those who are currently in form. However, usually the captain of the team has the final call on the selection after listening to the advice of certain players from the coaches and the selection panel after doing all the analytics (P. Shah & Patel, 2018). Sometimes the analytics could be wrong and thus the combination comes out to be unfit for the situation and thus the match turns out to be unfavorable for the team.

2.2.1 Challenges faced by using the current player selection process

From time-to-time existing processes need to be reviewed in order to keep up with the rest of the world and their advanced processes. Kenya over the years has been reviewing their selection process since 2010 and still is struggling to get the right process(Jhansi Rani et al., 2020). Again in 2021 they wanted a few changes in the process, this shows that the process has been very unreliable and constantly needs to be reviewed(A. Shah, 2020).

Another challenge is that in a pre-selected squad there almost two to three players for every slot with almost the same experience and thus it gets hard to select one player over the other and then this leads to gut feeling selection which not always is the correct decision. Also, the players who have a good relationship with the captain have a higher chance of getting selected in the team which is not fair to other deserving players (Newman et al., 2019). Finally, the conditions of the pitch and the other factors mentioned above also affect the choosing of the players, these conditions are evaluated then a decision is made and this decision can sometimes be based on a false evaluation and this is leads to maybe even losing the match (Son et al., 2018).

2.3 Related Works

2.3.1 Data Mining System for Selecting a Winning Cricket Team

This is a system that predicts players performance to help select a starting eleven team. Their focus was based on certain factors such as previous performance, height, style, average, strike rate, number of matches and number of innings played. From the above-mentioned factors, they used the data and predicted if a specific player will perform well in the upcoming matches. This was achieved by developing a regression model specifically with Random Forest regression, Decision tree regression, KNN regression and Linear regression. The final outcome turned out to be accurate using the random forest regression however one of the issues was that some of the features used did not make sense to the prediction such as batting style or bowling style which has no direct relation to performance as what matters is skills which both a left hand or a right-hand style player can accumulate through practise and training. Thus, this leads to over fitting of data which can bias the outcome of the match or even the performance of the player (Hasanika et al., 2021).

Player Combination	Support	Confidence	Combined Average	Individual Sum
HH Pandya, V Kohli, S Dhawan, RG Sharma	0.1607	0.7500	187.63	171.93
MS Dhoni, V Kohli, KL Rahul, RG Sharma	0.1071	0.8000	182.53	177.27
KM Jadhav, V Kohli, S Dhawan, RG Sharma	0.2053	0.7931	173.72	173.31
Kuldeep Yadav, V Kohli, KL Rahul, RG Sharma	0.1071	0.7059	173.65	152.19
V Kohli, S Dhawan, YS Chahal, RG Sharma	0.1696	0.7600	173.04	154.85

Figure 2.1: Correlation of combined average of batsmen (Hasanika et al., 2021)

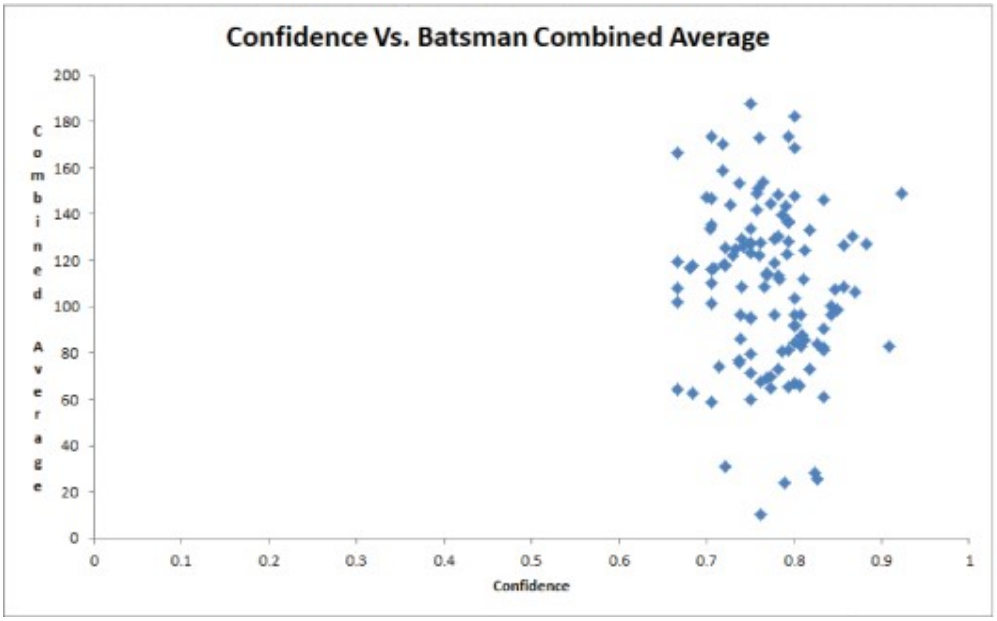


Figure 2.2: Distribution of confidence association rules against combined average (Hasanika et al., 2021)

2.3.2 Player's Performance Prediction in ODI Cricket Using Machine Learning Algorithms

The aim of the above paper by the authors was to predict performance of a player to allow them to enable selecting a team with the best combination to help with the satisfactory outcome for the match. The algorithms used to develop this system were Linear regression and SVM with Linear and Polynomial Kernel. One of the few limitations in their project is that the features they are using to consider the prediction are not accurate and they require more accurate features such as previous few scores to know where a player is in current form or not. Despite this limitation the project has turned out to be accurate with minimum errors (Anik et al., 2018).

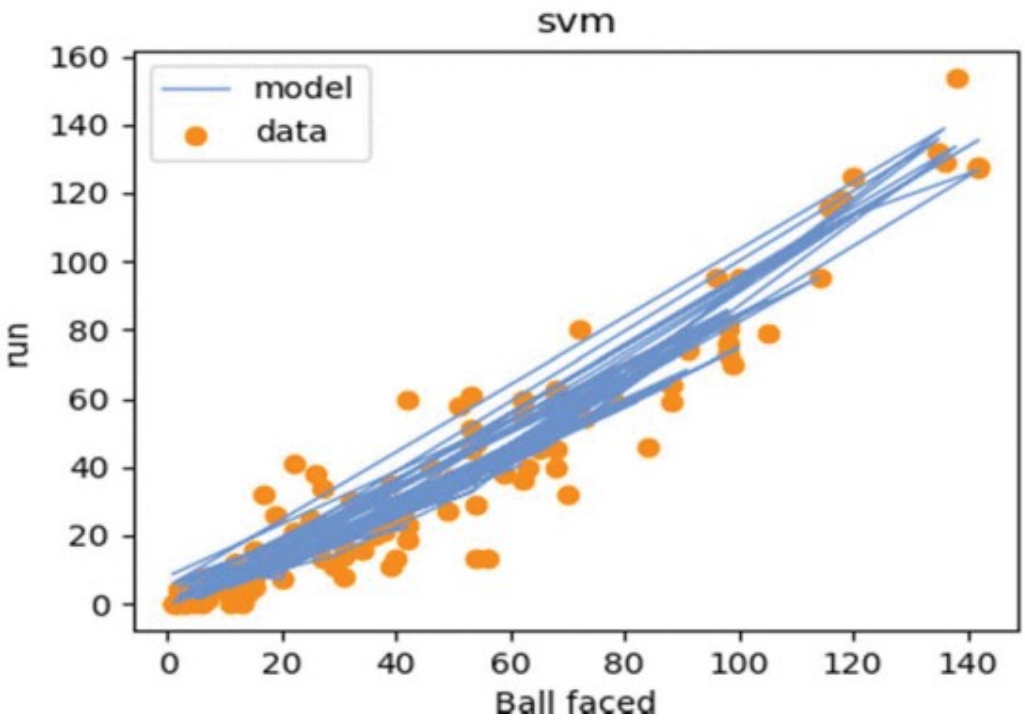


Figure 2.3: Best fitted svr line for Run and Ball face (Anik et al., 2018)

2.3.3 Machine learning-based Selection of Optimal sports Team based on the Players Performance

The objective of this study and project was to help in selecting the team with the perfect combination of players to win the match. The technologies and algorithms used for this project are logistic regression, SVM and random forests. The data was obtained from IPL matches and trained for Indian players. Since it was trained and tested in Indian conditions and players it is then limited to Indian boundaries thus limiting the scope of the whole project to the whole world(Shetty et al., 2020).

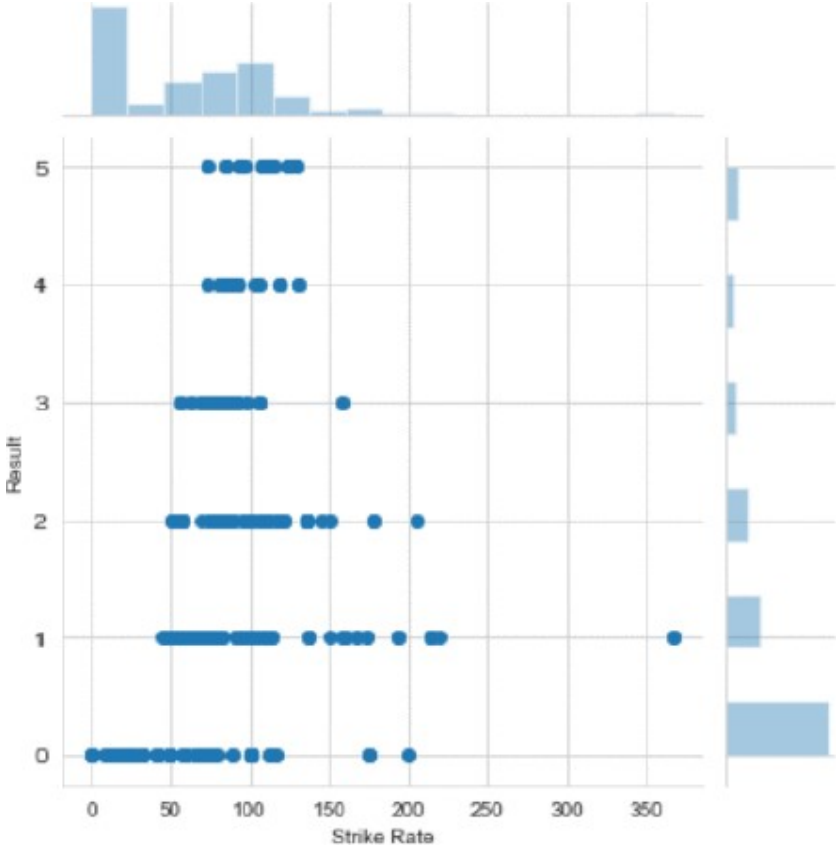


Figure 2.4: Joint plot of result vs strike rate (Shetty et al., 2020)

2.4 Gaps in existing systems

Majority of the existing systems relating to player performance prediction have an issue with the features used to predict. Some of these features missing and need to be there for the prediction and some features that were used are not accurate in predicting since they do not affect the outcome of the prediction of performance in the upcoming matches. A few examples of such features are physical body factors such as height which are not directly proportional in usage for the prediction of the performance since cricket is a sport that is played by all types of people that is short people, tall people and the factor of height does not affect the performance(Hasanika et al., 2021). In fact, one of the best players the game as ever seen named Sachin Tendulkar was called the Little Master because of his short height and another player who is very tall is also doing well in his cricketing career that is Macro Jansen.

Therefore, in order to have the system working well there is need to have all the appropriate features being used for the prediction and be tested and trained with data and use the most appropriate algorithm to give the most accurate prediction (Bunker & Thabtah, 2019).

2.5 Conceptual Framework

The proposed system has 2 users that is the admin and the club. The main role of the admin is to register clubs, start or end a season and view leaderboards to generate a leaderboards file that will be sent to the AI model to train the players and their previous performance and so that a suggested starting eleven can be predicted and used for the upcoming matches. The main role of the club is to register the players laying for their club and doing the scoring for the matches being played by their club in order to provide the performance data of players to the admin or Cricket Kenya Board for them to use this data to predict the starting eleven.

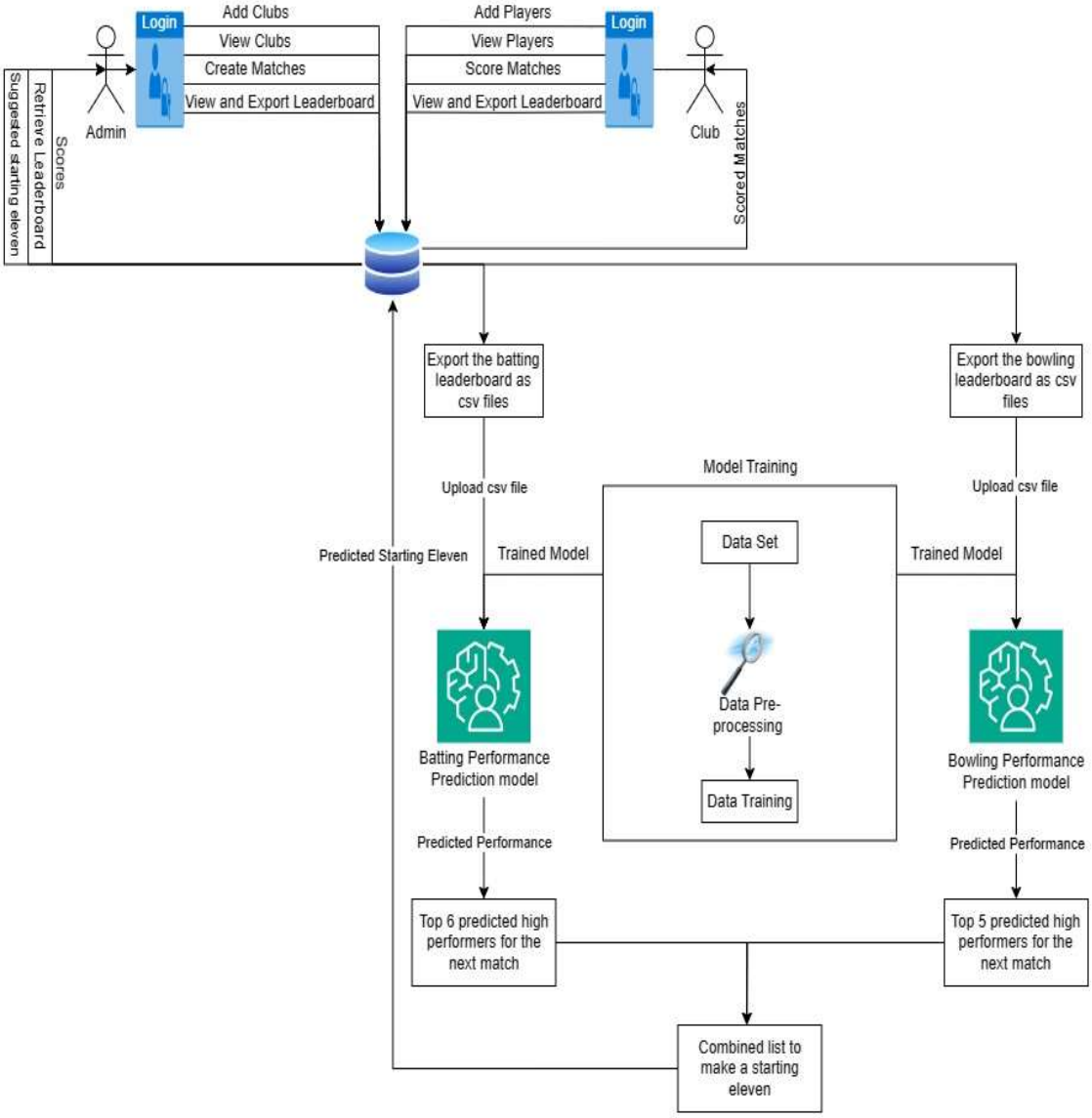


Figure 2.5: Conceptual Framework

Chapter 3: Methodology

3.1 Introduction

This chapter talks about the specific methodologies that were used to develop the system that is OOAD with the help of waterfall with feedback methodology which is predictive methodology. Moreover, this chapter mentions how the requirements of the system were collected and the specific diagrams that were drawn to facilitate development. Furthermore, the chapter mentions the specific tools and techniques used to develop this system. This chapter also explains the various tests that were conducted to make sure that the system is functioning as expected. Finally, this chapter also shows the user modules and their specific roles in the system.

3.2 Methodology

The specific methodology used was Waterfall with Feedback under Waterfall methodology as this methodology allows the correction of errors moving from one phase to another and also because the requirements of the system were known before the development started thus it just following the requirements(Fagarasan et al., 2021).

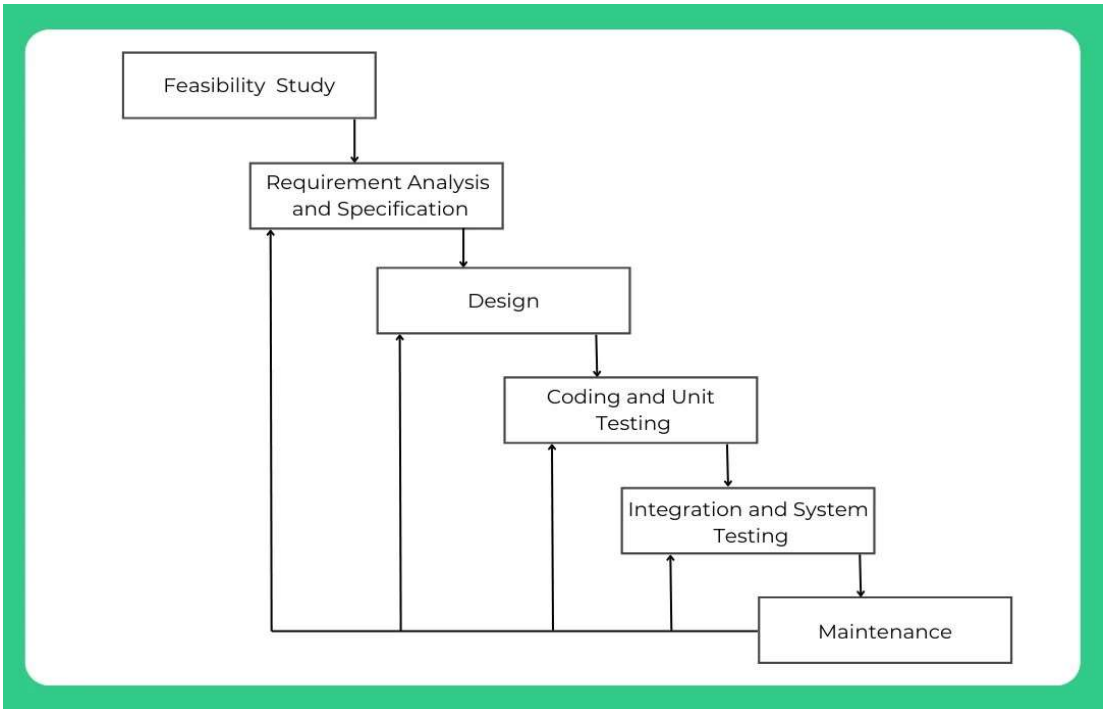


Figure 3.1: Waterfall with feedback methodology (Fagarasan et al., 2021)

3.2.1 Feasibility Study

This methodology starts with the analysis and the evaluation on the proposed project. This is done to check if the project is achievable and viable to undertake the project. This study acts as a tool to allow making a decision on whether the project needs to be undertaken or not(Fagarasan et al., 2021).

3.2.2 Requirements Analysis and Specification

This is the phase where the capabilities features and constraints of the system are identified in detail in order for the system to run smoothly. These requirements are usually gotten from following software engineering and requirements gathering process mostly by interviews as a form of data collection with ideation (Medeiros et al., 2018). And another way of getting the requirements is by identifying the gaps in existing system and filling those gaps to make the predictions more accurate (Fagarasan et al., 2021).

The training data and the actual data will be received from the scoring model and will be used to predict the performance in upcoming matches.

3.2.3 Design

In the second phase of the waterfall with feedback methodology a basic design of the system is drawn in order to facilitate development of the system. This is done by drawing design diagrams that helps the developer of the system to understand the requirements well and helps in fastening the development process of the prototype(Mitsuyuki et al., 2017). The paradigm approach chosen is Object Oriented Analysis and Design (OOAD) this is because OOAD turns out to be more suitable with the waterfall with feedback methodology. Moreover, this approach divides the system into use cases thus understanding the interests of specific users of the system well, thus understanding each user requirements well, this is easier as this approach is also data oriented. Moreover, if there are any changes or upgrades required for the system it is easy to implement(Mitsuyuki et al., 2017). Also, OOAD uses bottom-up approach which allows developing complex systems in a simpler manner easing up the development process(Fagarasan et al., 2021). Below are the analysis diagrams that will be drawn in order to help in facilitate development.

The proposed model is a machine learning model that will use supervised learning and form a classification model that learns the patterns and relationships from the input of data provided and the same will be used for the proposed system with the specific technique of decision trees

and random forest this is because these techniques allow handling non-linear relationships and missing data moreover allowing interpretability of the system and this then allows informed and reliable decisions to be made and accurate predictions of the players performance. However other techniques will also be used to check if they return a better accuracy than decision trees and random forest those ones will be used(Song et al., 2017).

3.2.3.1 Use Case Diagram

The reason for drawing this diagram is to help understand the functional requirements of the system and specific users of the system that is the admin and the club and developing a system keeping the interests of users in mind(Vachharajani & Pareek, 2019).

3.2.3.2 Sequence Diagram

The importance of a sequence diagram is to show the interactions between the users and the system by the order in which they occur. For the proposed system, the sequence diagram helped to know the processes, and which one occurs first. For example, the club admin logs in so the process starts with them entering the credentials, the credentials being sent to the database to verify if they are matching sends a message of successful login and gives access to the index page. If the credentials do not match send a message saying incorrect email or password(Setiyawan et al., 2023).

3.2.3.3 System Sequence Diagram

This is a graphical representation of the system that shows how the components of the system interact with each other and generally explaining how the overall system works. They are used hand in hand with Activity diagrams (Setiyawan et al., 2023).

3.2.3.4 Entity Relationship Diagram

An ERD is drawn to help the developer understand how the data is stored in the database with the specific details related to that data hence this is the basis on which the generic database is developed(Frantiska, 2018).

3.2.3.5 Class Diagram

This diagram shows the visuals of the system and the relationship of data stored in the database as the relationships explained using the ERD are not well understood thus a class diagram brings out the clear relationships of data collected. This diagram allowed the developer to understand the exact responsibilities of each user of the system(Setiyawan et al., 2023).

3.2.3.6 Activity Diagram

This is a diagram that is used to describe the detailed flow from stage to another and how each activity is carried out at each stage to dynamically explain the major aspects of the system(Setiyawan et al., 2023).

3.2.4 Coding and Unit testing

In this stage system is coded and developed with the help of the diagrams drawn above. Once the system is coded then each module of the system is independently tested to check if it is functioning well or not. If it is not working as required the necessary debugging and troubleshooting is done and the module is tested again till works as expected(Setiyawan et al., 2023).

3.2.5 Integration and System testing

In this phase the two or more models are combined and then tested if they are working as they are supposed to work just like they were before the integration. Once all the modules are integrated and tested the whole system is tested at once to again check if everything is working as expected(Setiyawan et al., 2023).

The types of tests carried out to test the final system will be the ones stated below.

Table 1: Tests to be used

Type of test	**Reason.**	**Tool to be used.**
Black box testing.	To check if the system is outputting correct data by data input (Romdhana et al., 2022).	Microsoft Word – to help come up with test cases. UFT Automating tool.
Alpha testing.	To validate the system before rolling it out to the market or before submitting it(Romdhana et al., 2022).	Google. VSCode.
Regression testing.	To check if any additions in the code have affected the output of other code(Dahiya et al., 2022).	Micro Focus Unified Functional testing.
Testing of code.	To check if the code written out is accurate such as there are no errors and not many white lines wasted(Romdhana et al., 2022).	WebdriverIO
Automated testing	To help fully check the code and the system with minimum effort(Dahiya et al., 2022).	Burp Suite Application Security Testing Software

3.2.6 Maintenance

This is the final phase of the waterfall with feedback methodology where once the system is tested for final errors and then deploys the website to the user's server and maintains it with any updates required in future(Turnip et al., 2020).

3.3 Deliverables

These refer to the specific end results that are measured and submitted or presented to the supervisors or the seniors in order to complete the project and checking the progress of the project(Jaber et al., 2018).

3.3.1 Documentation Document

This is a document that outlines everything in detail about what the whole system will be anticipating doing and how it will be done. This needs to be submitted as a deliverable since it needs to be assessed by supervisors so that they can give a go ahead on developing the actual system(Team, 2022).

3.3.2 Cricket Scoring System

This will also be a deliverable since this system will be required for the prediction model to work as it is from here that the data for the top performers will be obtained from.

3.3.3 Prediction Model

Finally, this deliverable will be required to complete the project since the entire project is based on prediction model of a starting eleven of a cricket team.

3.4 Tools and Techniques

3.4.1 Language

The languages used for developing the website will be in a combination of HTML, CSS, JavaScript and PHP for the frontend and backend of the website and for the prediction model python will be heavily used. The combination of these languages used will return the best possible outcome of the project(Tatroe & MacIntyre, 2020).

3.4.2 Code Editor and IDE

The code editor and IDE that will be used to compile and run the rode will be VSCode for the frontend and backend development of the web application.

3.4.3 Output

For the output of the project Chrome will be used since it is fast, secure and reliable and also widely used thus will be easy to present using Google Chrome(Berham & Morris, 2022).

3.4.4 Framework

CodeIgniter framework will be used for development of the web application this is because it has a free and simple set up configuration file that facilitates development using PHP and other languages(Valarezo & Guarda, 2018).

3.4.5 Database

MySQL allows systems to store data, especially usernames and passwords, and other data with appropriate security and safety stored data. MySQL is also used to clean the data and is used as a data warehouse or data lake and is better for storage(Patil et al., 2017).

3.4.6 Backup

GitHub is the best place to back-up incomplete code since it is a secured form of storage with easy usage(Han et al., 2019).

3.4.8 API

This API is used to hold in building and testing the accuracy of reports created by the system and also to allow testing of the email integration(Subramanian & Raj, 2019).

3.4.9 Diagram Tools

Draw.io and MS Project is used to help in drawing the design diagrams and the Gantt chart(Technologies (CSIT) & MM, 2021).

3.4.10 SciKit – Learn

This is a free machine learning library for python that allows access to many functions that ease up development of the model(Szymański & Kajdanowicz, 2018).

Chapter 4: System Analysis and Design

4.1 Introduction

This chapter mostly deals with the analysis and design of the system. To be specific this chapter divides the requirements into functional and non-functional requirements to understand the user requirements and the system requirements. Moreover, the chapter also includes details about the specific system architecture used for handling requests and also the design diagrams that help the developers to know what to code. The specific diagrams include use case diagram, class diagram, system sequence diagram, sequence diagram, entity relationship diagram, activity diagram and lastly the database schema. The list of diagrams mentioned above are those that are required if the analysis and design approach chosen is Object-oriented. However, on top of the diagrams there are a few wireframes also created that shows blueprints of the system.

4.2 System Requirements

System requirements is required for a software engineer to help in facilitate designing the diagrams for the developers so that they can know what to code in order to develop a system that meets the user requirements. Thus, the same applies to the player performance prediction model. These requirements are divided into functional requirements and non-functional requirements. Specific functional and non-functional requirements are mentioned below.

4.2.1 Functional Requirements

Functional requirements refer to the specific abilities of a system specifically conducted by a specific user. The player performance prediction model will have 2 users that is the admin and the club. Starting with the specific requirements of the admin that are:

i. To login and access the admin module after inputting the correct credentials. This happens after a verification and validation process with the credential match.
ii. Register a club so that the club can add their players.
iii. Edit and view clubs and their associations.
iv. Start and end a season or a tournament.
v. View leaderboards of players.
vi. View scores of ongoing matches and previous matches.
vii. Generate leaderboards of players at the end of a season or a tournament.

The second user is the club and the requirements of this user are:

i. To login and access the club's module after entering the correct credentials.
ii. To register new players to their club.
iii. Edit the registered players information.
iv. To score for their matches basically for matches played by this club.
v. View scores of ongoing and previous matches.
vi. View leaderboards of players in the current and previous seasons and tournaments.

4.2.2 Non-Functional Requirements

Non-Functional requirements refer to the intangible requirements of the system that define the user experience. Although the systems can function without these requirements, but it is necessary to attract new users and retain the existing users. For the player performance prediction model, the non-functional requirements are:

i. The system needs to be reliable meaning that the data stored in the database needs to be accurate, like for scores of players they need to be accurately stored since the prediction of the players performance will be done based on this score so if the scores stored are inaccurate then the predictions will be inaccurate. This is done by picking the exact score of a player and once a match is over the scores should not be editable and should be exported to the database the way they are.
ii. Availability is also required so that whenever the user wants to register or log in to the system, he/she is able to do so without facing any difficulties.
iii. The aspect of security also comes in although the player performance prediction model has no integration of money hence less prone to be attacked by the hackers but still, they can hack into the system and play around with the scores and statistics which is very risky if the data stored is used for decision making. Thus, the viewing module should be restricted to registered users and should be only view only data and not editable.
iv. Should be user friendly for a non-IT user or a naïve user to ease their activity with the system. There can be a recorded tutorial on how to use the model and also on the main page there can be a question button that can provide any help to users.

4.3 System Analysis Diagrams

4.3.1 System Architecture

System architecture refers to the structural design of the system trying to show what happened at the back end when a request to access a service from the system is made by a client. The proposed player performance prediction model has a 3-tier system architecture which is a client-server architecture. As evident from the figure below from the system architecture diagram the user can make a request to access service from the system by send a request from the web browser after being connected to the internet. The web browser then requests the web server for the request via the internet from which the request is taken to the database server that responds accordingly to the request made(Arteaga et al., 2019).

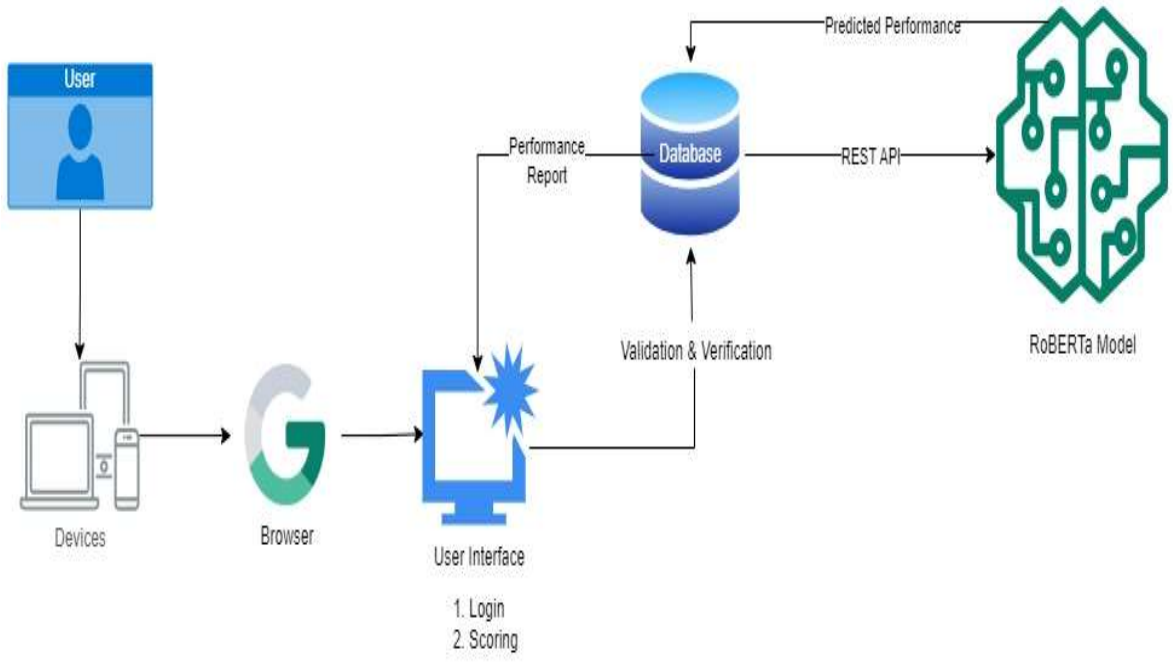


Figure 4.1: System Architecture

4.3.2 Use Case Diagram

As evident from the diagram below the use case is used to show the roles and abilities of all the users in the system. There are two users of the system that is the Admin and the Club. The admin is the primary actor followed by the club. The main role of the club is to do scoring for the matches so that the leaderboard is updated with the correct statistics and then this data can be retrieved by the admin so that he/she can feed this data into the prediction model that will analyse the data and make a prediction is the performance of the player will be for the future matches(Vachharajani & Pareek, 2019).

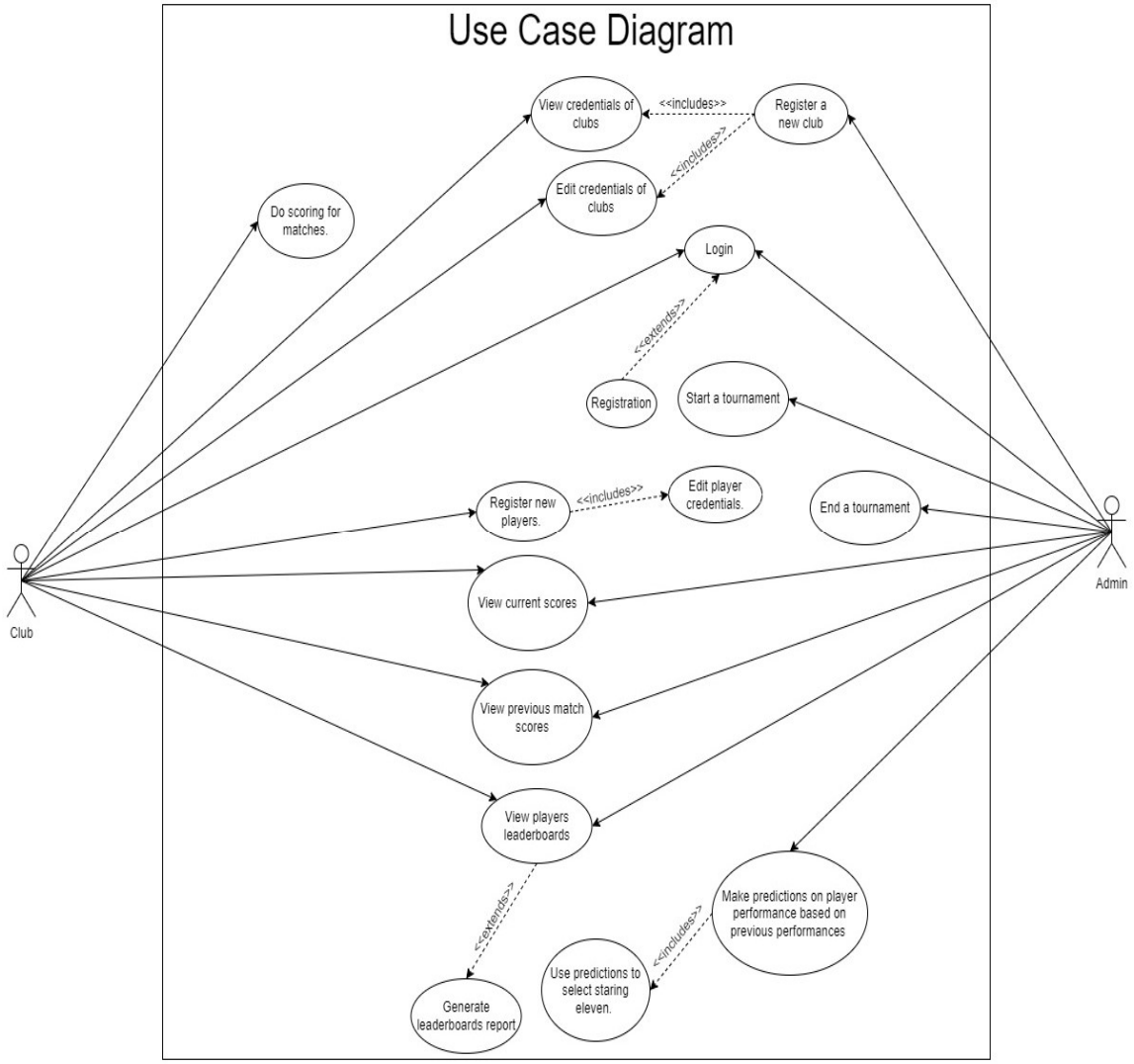


Figure 4.2: Use Case Diagram

Table 2: Use Case Scenario 1

Use Case	Values
Actor	Admin
Pre - Condition	Must be logged in to the system.
Post - Condition	All input fields should have appropriate data before clicking the predict button or before the prediction is done.
Alternate Condition	No input field should be empty otherwise the user will be forced to fill it in first.

Table 3: Use Case Scenario 2

Use Case	Prediction
Actor	Admin
Pre - Condition	No input field should be empty
Post - Condition	A prediction verdict of good, bad and average should be displayed
Alternate Condition	In case of invalid or inconsistent information the system should display an error message prompting the admin to correct the input before generating the prediction

Table 4: Use Case Scenario 3

Use Case	Report Generation
Actor	Admin
Pre - Condition	A prediction must have been made
Post - Condition	The report should be displayed in a presentable format such as a pdf
Alternate Condition	If there are errors in the prediction data, the system should provide an error message and prevent the generation of the report

4.3.3 Sequence Diagram

A sequence diagram illustrates the sequence of message between object and how those objects interact with each other. The diagram below shoes the requests made by each other user to retrieve or access data from the database through the user interface and the series of steps followed by the users. If the actions are not followed as evident from the diagram, then there could be issues in accessing what is needed by the user(Setiyawan et al., 2023.

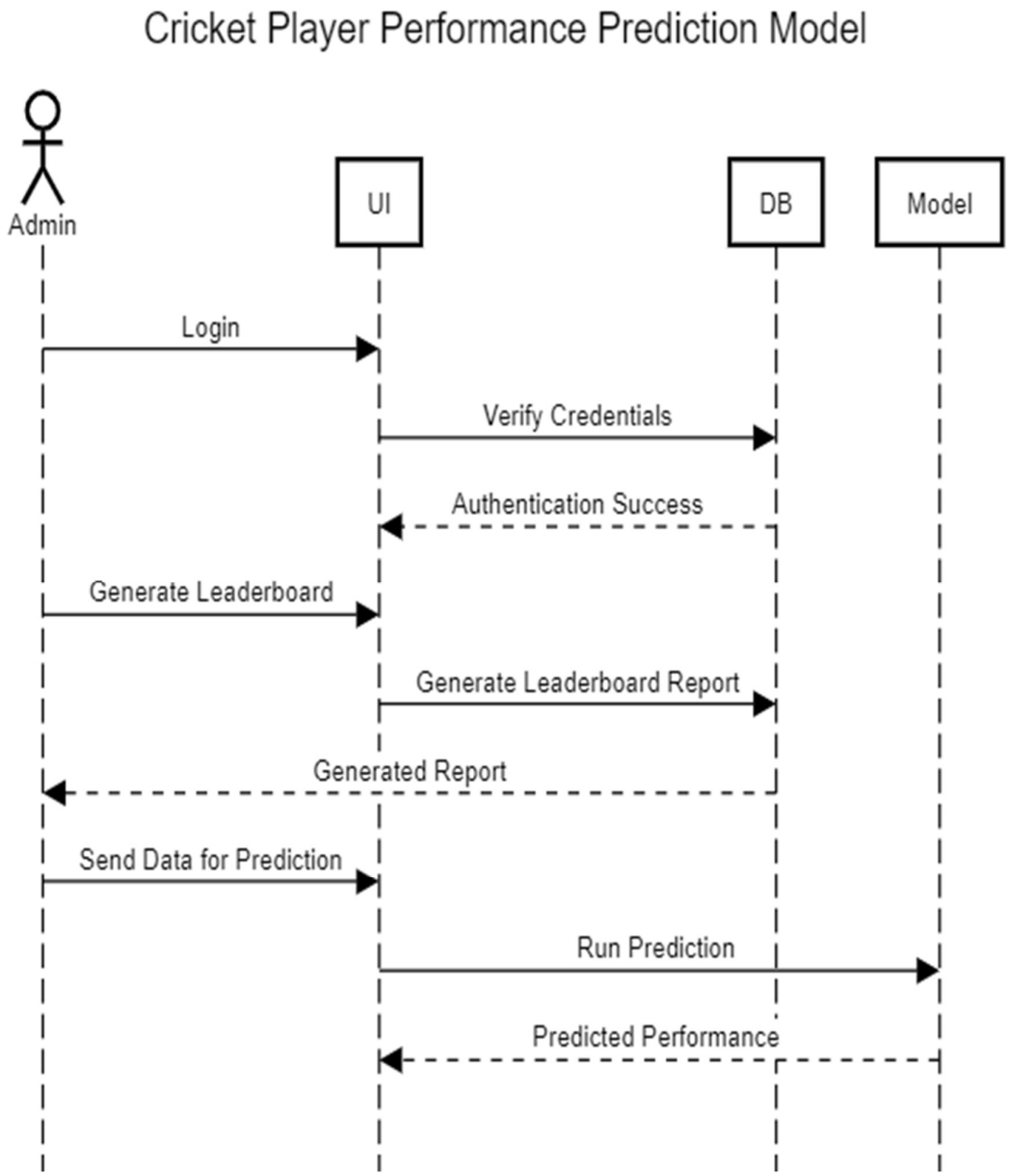


Figure 4.3: Sequence Diagram - Admin

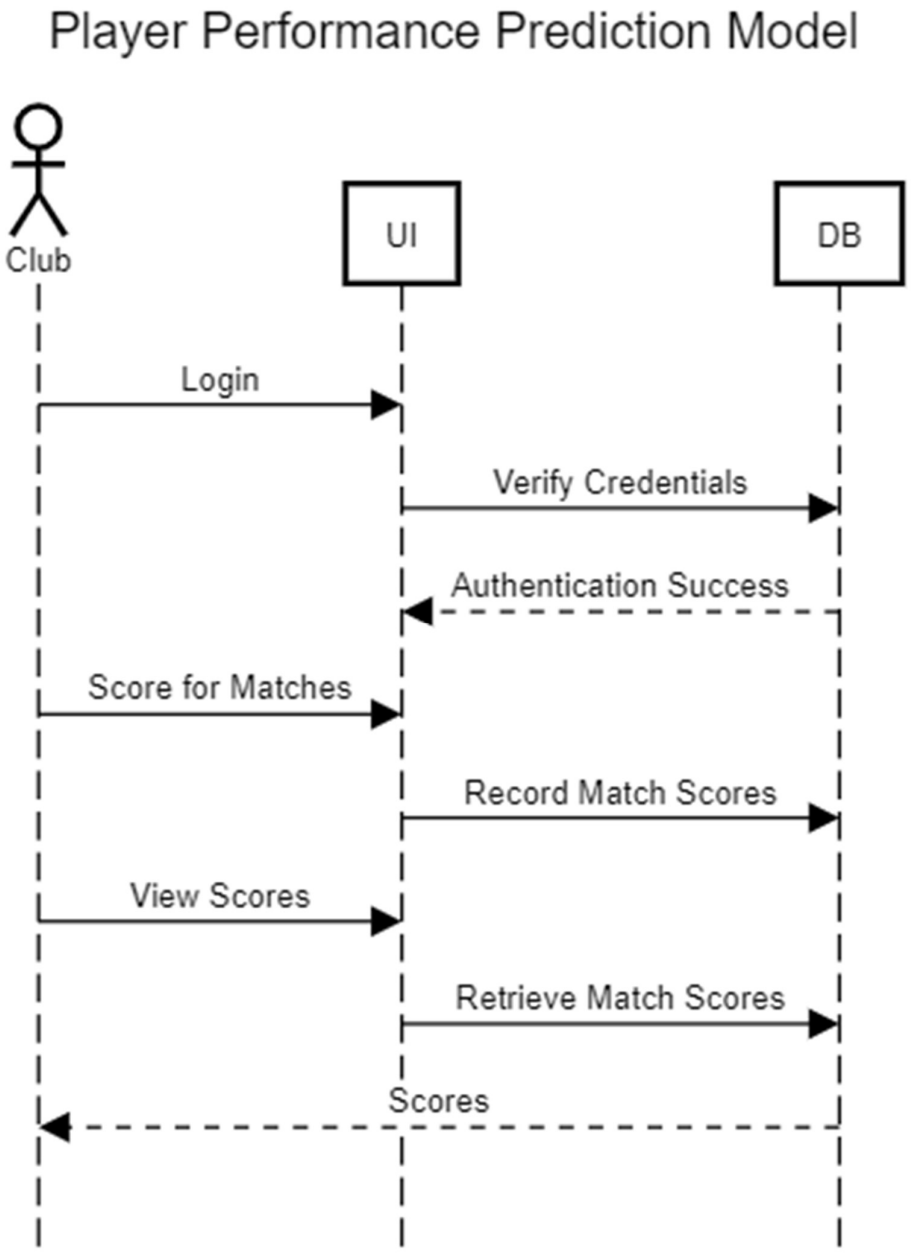


Figure 4.4: Sequence Diagram - Club

4.3.4 System Sequence Diagram

This diagram is used to display the interactions of between an external user and a system. It mostly focuses on the events exchanged between the system and the user to achieve a specific goal. As evident from the diagram and mentioned earlier that the admin is the main user of the system who gets the data of players performance from the leaderboards which is updated by the clubs and the admin feeds this data to the prediction model. So first the data needs to be cleaned meaning it should be made consistent to avoid any errors during the prediction. Then select the features needed for the prediction and choose a model technique that suits the data prediction model which is the decision tree or the random forest algorithm for the player performance prediction model. Use the algorithm to train the data and once the data is trained it can then successfully predict the performance and download the report.

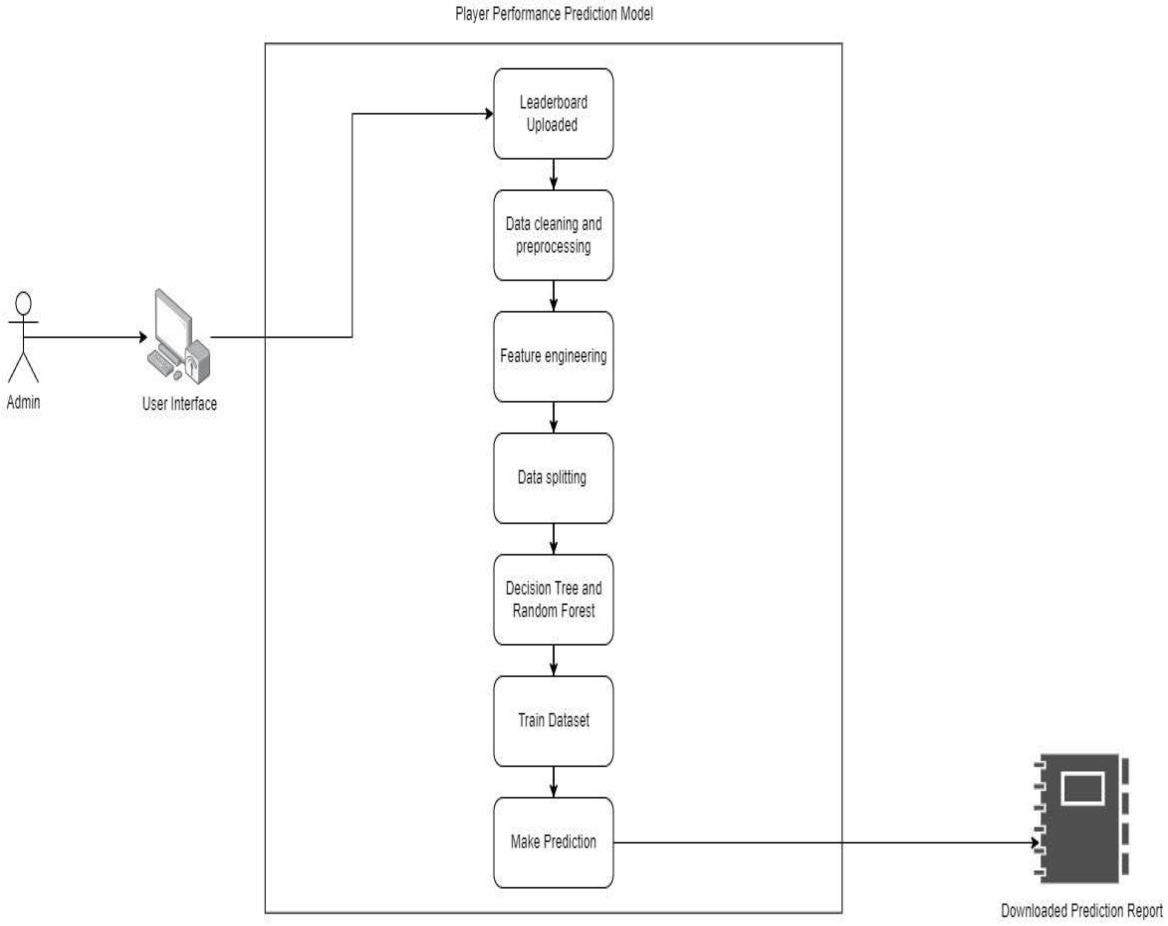


Figure 4.5: System Sequence Diagram

4.3.5 Entity Relationship Diagram

This diagram helps to understand all the tables in the database and the relationships they carry with each other. As evident from the ER diagram below the system has nine entities represented in the rectangles. The rhombus defines the relationship between the entities, the oval represents the attributes the oval that has an underline is known to be the primary key. Moreover, the diagram also has several types of arrows pointing to entities that shows the type of relationship or cardinality of the relationship. The most used cardinalities are one to one, one to many and many to many. As evident from the diagram below one to one and one to many cardinalities are mostly used. An example of a one-to-one relationship is one user who has one role in the system either the admin or the club(Setiyawan et al., 2023).

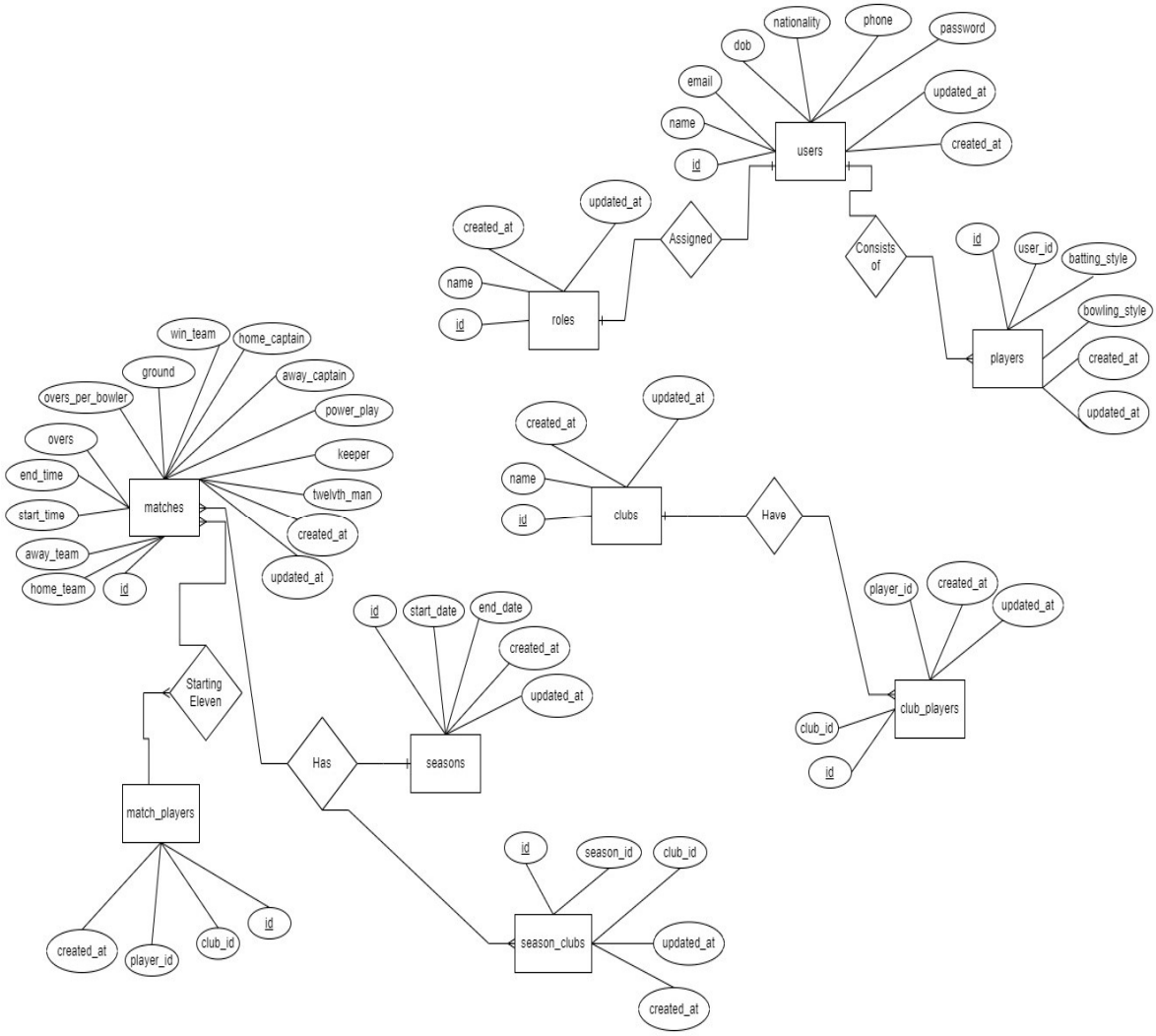


Figure 4.6: Entity Relationship Diagram

4.3.6 Class Diagram

A class diagram illustrates the classes, entities, its methods, and the relationship amongst the classes. As evident from the diagram below the class diagram has a superclass or the parent class that is the user class that has subclasses such as the players, the match_players and the club_players. Just like the ER diagram and the Database schema this diagram also shows relationships based on their cardinalities. As evident from the diagram below the relationships are also shown with several types of arrows firstly to show a parent class there is an inheritance and pointing towards the user meaning that the user is the parent class, a composition arrow showing that the classes from which the arrows are coming out cannot exist without the class to which they are pointing at. For our case, a club can exist without players thus there is an aggregation relationship. For the composition relationship as evident that there is a relationship between the season and the clubs playing that season that is the season_clubs thus if the season is closed or deleted then the clubs playing in that season will also be removed. (Setiyawan et al., 2023

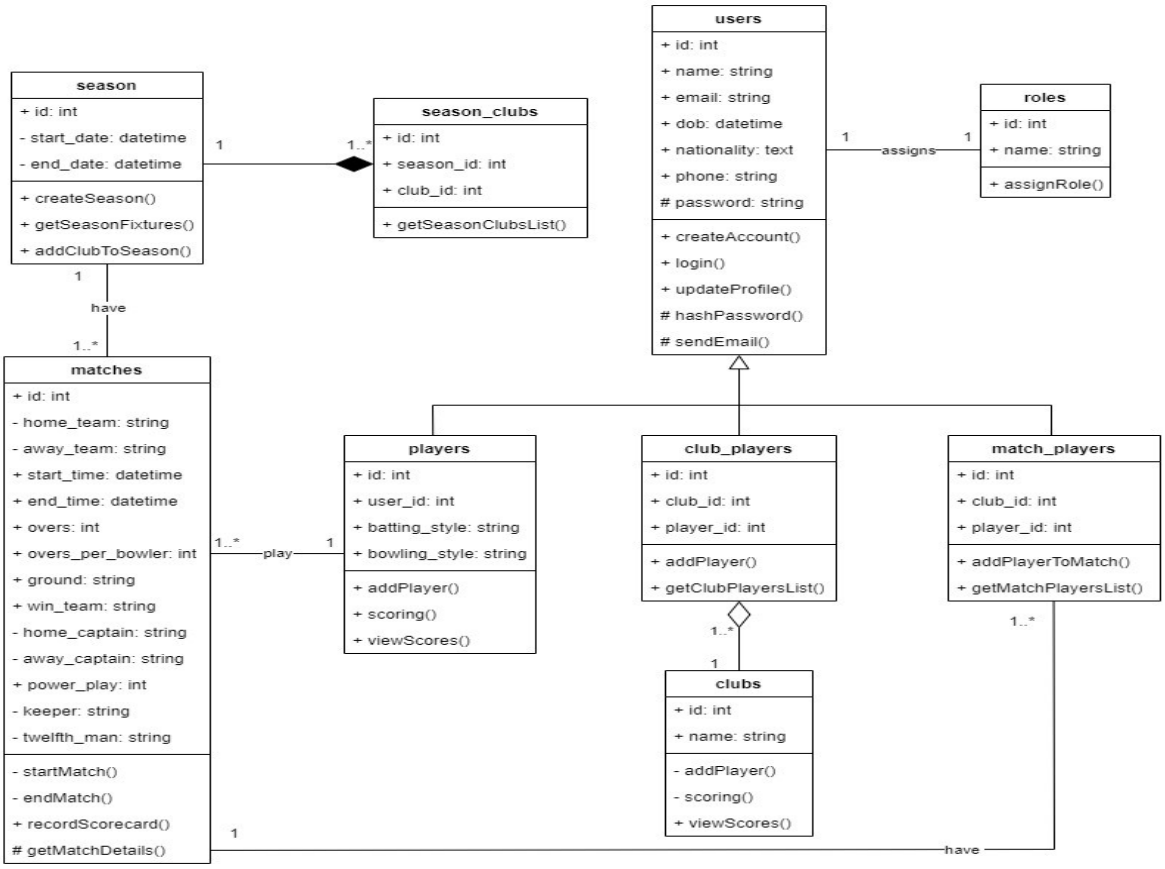


Figure 4.7: Class Diagram

4.3.7 Activity Diagram

An activity diagram is used to illustrate the steps followed to meet the purpose of a use case. They are used to model all the steps undertaken in a system. For the prediction model the steps start by a login system which all the users have to go through to access their authorized interface and carry out their assigned roles. The rhombus is a checkpoint where the appropriate steps are taken to reach the desired output. The most activity is carried out by the admin who generates the leaderboards model and feeds it into the RoBERTa model which is a deep learning model used to help in predictive tasks and help to come up with a prediction for the players performance in the upcoming matches (Setiyawan et al., 2023.

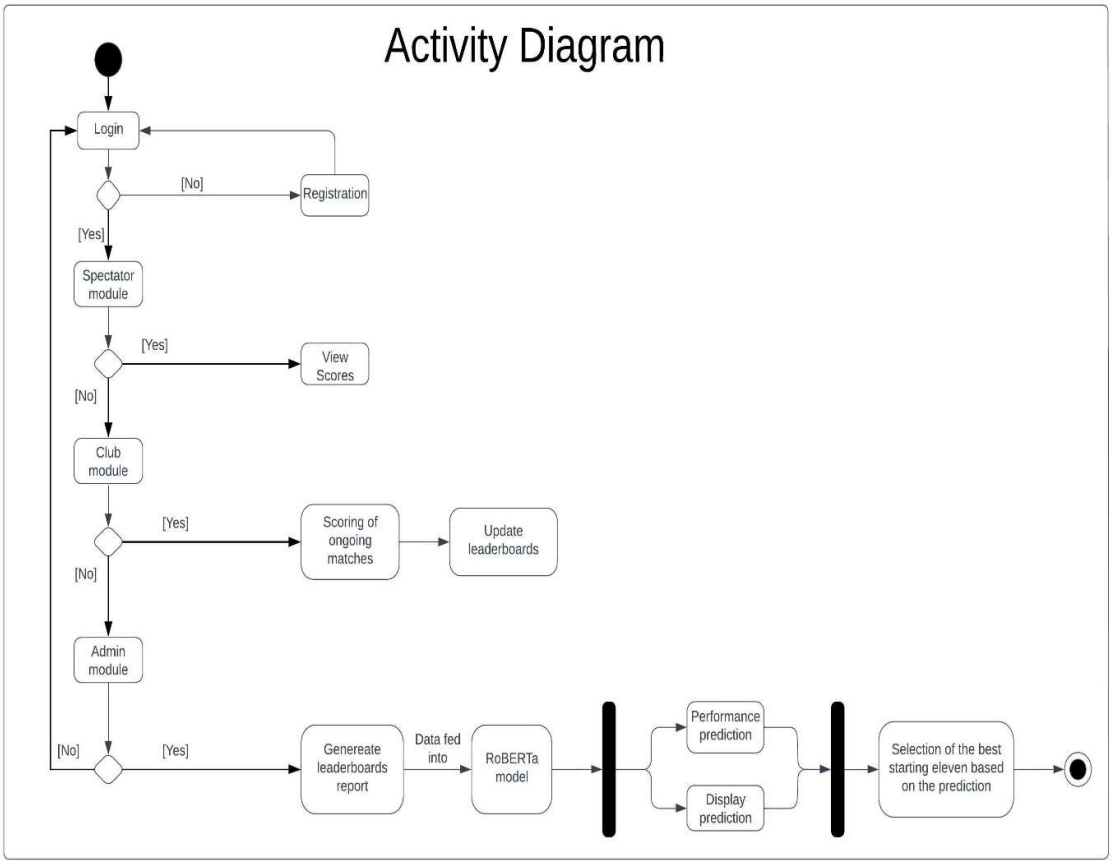


Figure 4.8: Activity Diagram

4.3.8 Database Schema

This diagram helps the developer with creating the database as the developer knows what tables to create and with what attributes with their key attributes. The player performance prediction model has 9 tables with 9 primary keys and 8 foreign keys. The foreign keys are the primary keys of other tables or entities that needs to be linked to in order to have a relationship. The primary key uniquely identifies the information required. Just like the Entity Relationship Diagram the database schema also has the cardinality relationships exactly similar to those displayed in the ER diagram(Delplanque et al., 2018).

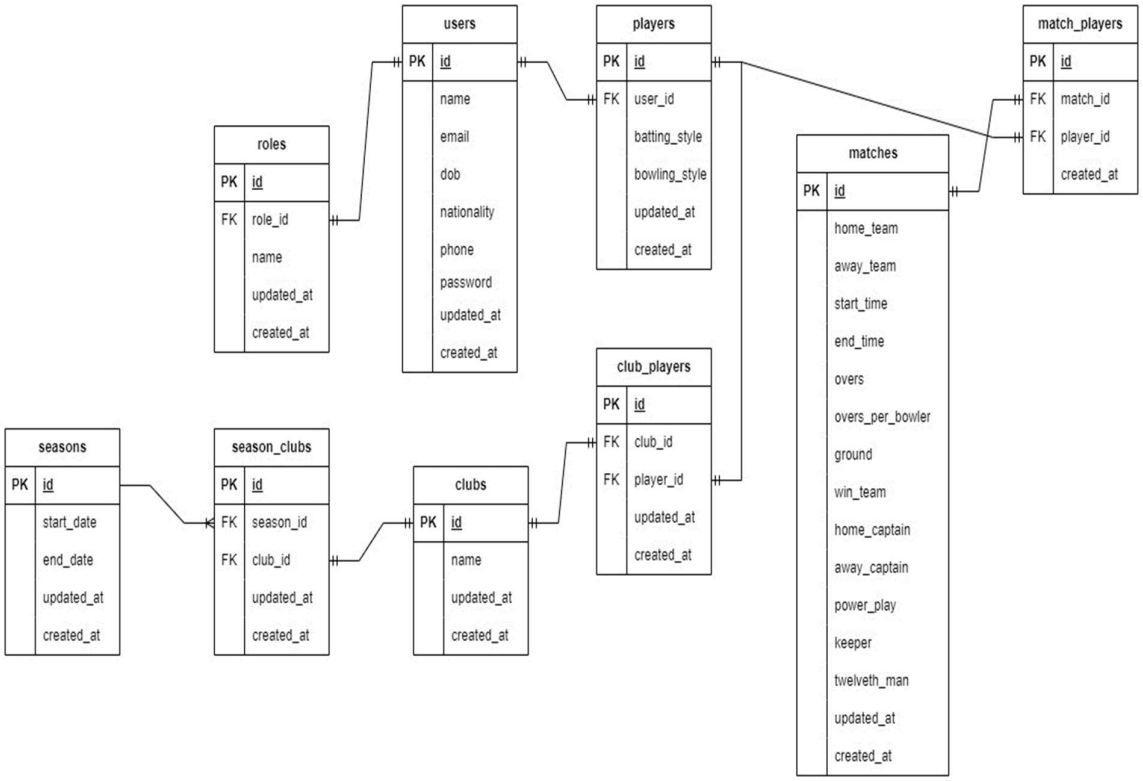


Figure 4.9: Database Schema

4.3.9 Wireframes

These diagram helps with blueprints of how the system looks like so that the users and the developers can anticipate and have an early feel of what the actual system is going to look like(Garverick, 2018).

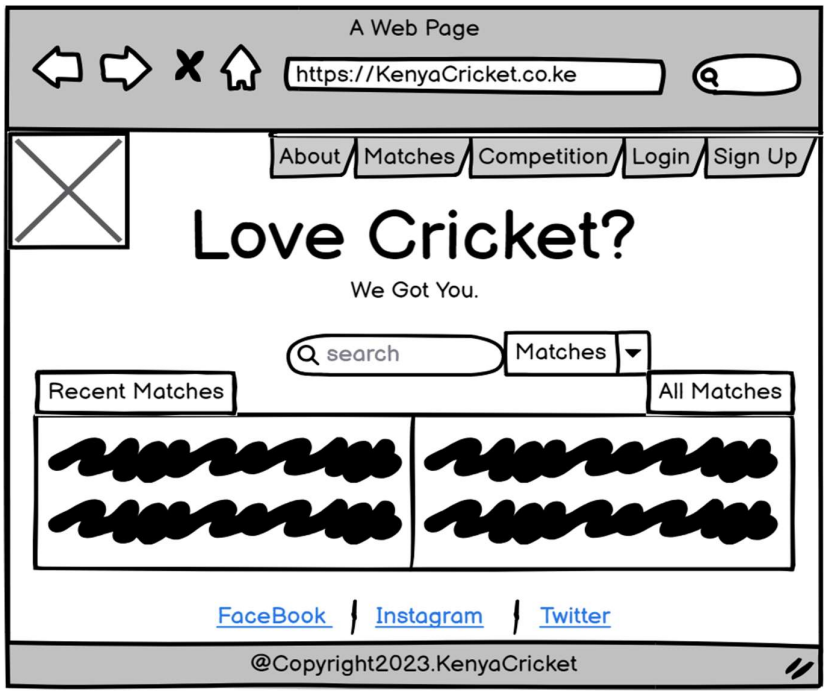


Figure 4.10: Landing Page Wireframe

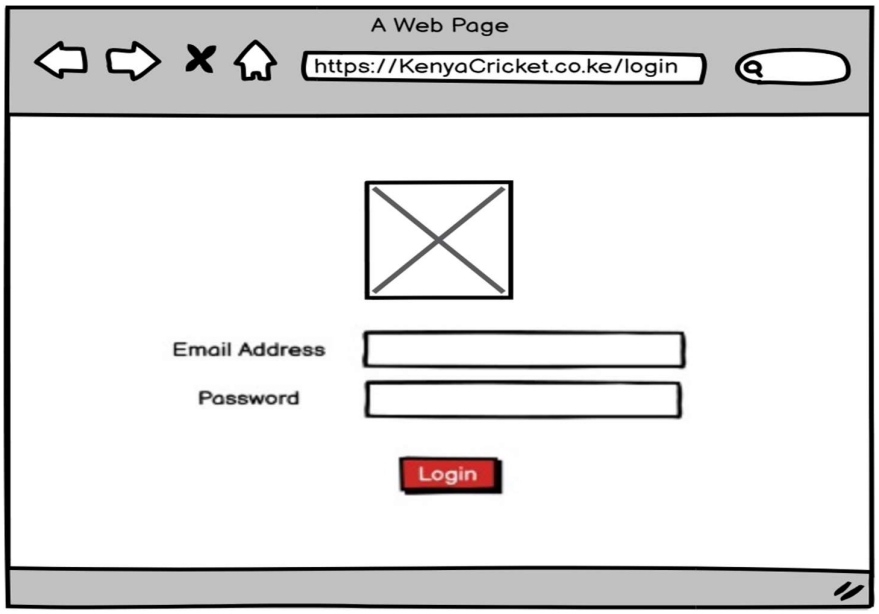


Figure 4.11: Login Page Wireframe

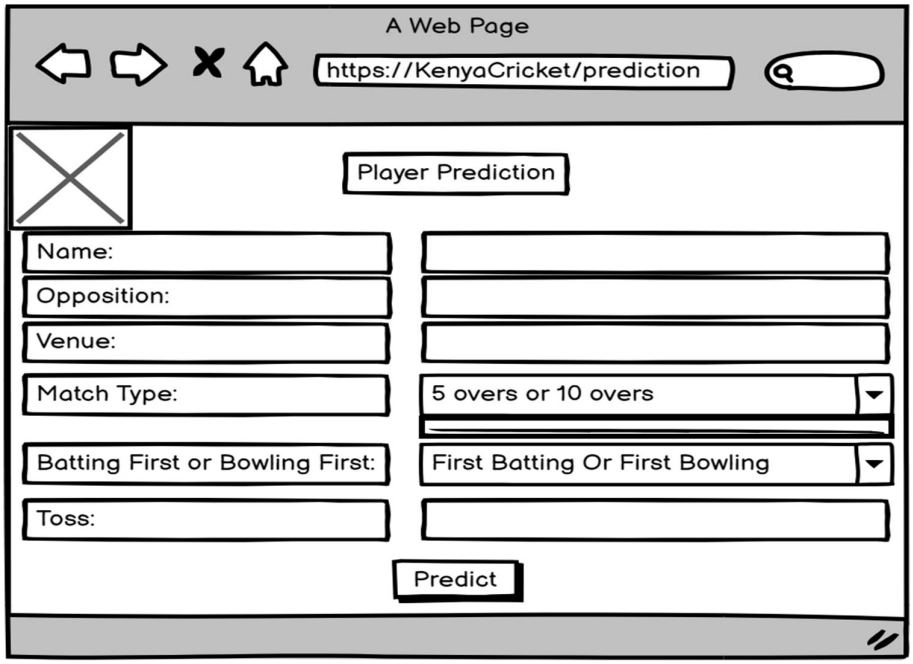


Figure 4.12: Prediction Model Wireframe

Chapter 5: System Implementation and Testing

5.1 Introduction

This chapter talks about the various hardware and software requirements for the system to work properly. It gives the minimum requirements and the recommended requirements for the system to work perfectly that is for the hardware requirements and for the software requirements it gives the required software and gives justification why that software is required. It also gives a brief description of about the dataset and how it was trained into the prediction model to get the desired output. The testing paradigms also included that is the black box testing where a table of information was generated to display the specific tests that were carried out and the results of the test cases with sufficient evidence.

5.2 Description of the Implementation Environment

5.2.1 Hardware Requirements

The table below show the hardware requirements for the system to work properly and also suggests the recommended requirements for the system to work perfectly (Hazelwood et al., 2018).

Table 5: Hardware Requirements

Component	Minimum	Recommended
Processor	1.9 gigahertz (GHz) x86- or x64-bit dual core processor	2.8 gigahertz (GHz) x86
Storage	256 GB SSD or 512 HDD	512 GB SSD or 1TB HDD
RAM	8 GB (7.88 GB useable)	16.0 GB (15.9 GB usable)
System Type	64-bit operating system, x64-based processor	64-bit operating system, x64-based processor

5.2.2 Software Requirements

Hardware only is never enough to run a project there is always need of software and below are the software requirements for the system that will ensure smooth running (Amershi et al., 2019).

Table 6: Software Requirements

Component	Description	Justification
Operating System	Any Windows 7 or above and any Mac OS 10.8 or newer	A device with any of the specified operating system makes it easier to access the system securely and running the system on a stable operating system
Internet or Broadband Network compatibility and connection	5 Mbps or better	Most of the systems are developed and run on the internet and this makes it easier to share code and also to save the code and for this internet connection is required to connect to the internet connection
Web Browser	Google Chrome or Microsoft Edge	Web browsers are the platforms that allow accessing most of the developed systems. The aim to have a web browser compatible device is to allow sharing the system with many people who can then access and use the system.

5.3 Description of the dataset

The player performance prediction model has two datasets that is working on that is the batting dataset and the bowling dataset.

5.3.1 Batting data

This is data that is downloaded from the scoring system from the leaderboards report, specifically the batting leaderboards report. This is report that is generated from the scores that is updated after every match and it stores all the batting data of all the players. The batting data was downloaded in csv format and loaded into the prediction model where the data is then cleaned and pre-processed such that any row that has null values is removed so that the final data will only predict the performance of all the players that have all the data. The dataset is then split into training and testing data using the ratio 0.8:0.2 where 0.8 is the training data and 0.2 is the test data. So once the dataset was split it then the dataset with 0 missing values is trained and then it uses 6 features to do the prediction that is the previous five performances for the batsman and the average of the batsman. So, it is such that if the performance for the next match is high and the average of the player is high then it selects that player in the dataset. This means that the prediction of the next match that is the runs and the average have a direct relationship where if both are high then that player is selected. Top six players who are predicted to perform well in the batting data are selected and starting eleven of the next match. This dataset had a few null values and thus there was need to clean the dataset before actually training and testing the data.

id	Name	Matches	Runs	Bowls	Outs	Average	Performance Runs 5	Performance Runs 4	Performance Runs 3	Performance Runs 2	Performance Runs 1
1	Ali Connah	8	146	132	6	18	0	9	24	12	18
2	Rubina Jansik	6	50	164	5	8	0	10	0		
3	Barret Penwright	6	48	167	5	8	17	24	0	14	14
4	Minna Darrigone	7	115	174	5	16	9			5	18
5	Lilly Cherrington	7	96	165	6	13	25	13	17	16	5
6	Sallie Beaford	5	57	123	5	11	7	17	17	17	
7	Bone Rhymes	5	132	160	5	26	1	3	18	7	17
8	Galvin Novello	7	70	167	6	10	6	0	13		25
9	Abran O' Cloney	8	93	86	7	11	17	3	20	24	
10	Orbadiah Rotham	8	53	84	7	6	12	18	6	3	9
11	Lauren Playdon	6	75	115	5	12	23	12		25	12
12	Burtie Jonczyk	6	50	64	5	8	12	25	25	9	0
13	Sidoney Eckley	7	122	122	5	17	6	17	1		9
14	Claudius Martinho	5	130	150	5	26	2		15	7	20
15	Ania Chiommienti	5	99	104	5	19	0	14	0	0	9
16	Randi Cattrall	5	53	153	5	10	24	10	7	3	
17	Emerson Belk	7	128	79	6	18	15	25	19	10	0
18	Clarita Tunnicliff	5	90	157	5	18	20	15	11	0	
19	Brigham Karlolak	8	54	113	7	6	8	20	9	25	

Figure 5.1: Batting Dataset Overview

```
[ ]  df_bt.describe()
```

	id	Matches	Runs	Bowls	Outs	Average	Performance Runs 5	Performance Runs 4	Performance Runs 3	Performance Runs 2	Performance Runs 1
count	600.000000	600.000000	600.000000	600.000000	600.000000	600.000000	600.000000	600.00000	600.000000	600.000000	600.000000
mean	300.500000	6.585000	96.690000	120.865000	5.790000	14.775000	12.771667	12.61000	12.545000	12.691667	12.246667
std	173.349358	1.108222	31.060777	34.152561	0.945546	5.762788	7.726947	7.66763	7.485856	7.759098	7.538160
min	1.000000	5.000000	44.000000	60.000000	5.000000	5.000000	0.000000	0.00000	0.000000	0.000000	0.000000
25%	150.750000	6.000000	70.750000	91.000000	5.000000	10.000000	6.000000	6.00000	6.000000	6.000000	5.750000
50%	300.500000	7.000000	97.000000	121.000000	5.000000	14.000000	13.000000	13.00000	13.000000	13.000000	12.000000
75%	450.250000	8.000000	125.000000	150.000000	6.000000	19.000000	20.000000	19.00000	19.000000	20.000000	19.000000
max	600.000000	8.000000	150.000000	180.000000	8.000000	30.000000	25.000000	25.00000	25.000000	25.000000	25.000000

Figure 5.2: Batting Data Description

5.3.2 Bowling data

This data was also downloaded from the scoring model where just like the batting data it the bowling data was also updated after every match. The bowling data was downloaded in csv format then loaded into the prediction model where the data was then cleaned and pre-processed since there was presence of null values in some of the players data so the final data that was used to predict the performance had all the rows will the null values removed thus only players with full performance data was used to predict the performance. The dataset was then split into two that is the training data and the testing data using the ration 0.8:0.2 where 0.8 was the training data and 0.2 was the testing data. Just like the batting data the bowling data also used 6 parameters to predict the performance, that is performance 5, performance 4, performance 3, performance 2, performance 1 and the economy. So, the previous 5 performance data was used to predict the performance of the players together with the economy such that the economy and the performance prediction has an inverse relationship where the higher the performance and the lower the economy this enabled a player to be selected in the starting eleven. So, 5 players from the bowling data were selected and added to the 6 players that were predicted from the batting dataset and this returned a list of eleven players that were predicted to perform well for the upcoming matches.

id	Name	Matches	Overs Bowled	Bowls	Runs Given	Wickets Taken	Economy	Average	Performance Wickets 5	Performance Wickets 4	Performance Wickets 3	Performance Wickets 2	Performance Wickets 1
1	Geralda Lacroutz	6	6	36	50	6	8.23	8	4	2	4	1	
2	Beverley Cicculi	5	5	30	90	4	4.99	22	1	3	3		
3	Rachele Annice	7	7	42	47	1	4.75	47	1	1	3	1	
4	Merilyn Blue	7	7	42	31	7	7.77	4	1	3	2	1	6
5	Louise Luxon	5	5	30	64	9	5.97	7	2	4	1	1	
6	Allie Yeates	8	8	48	88	7	6.27	12	1	2		2	6
7	Horst Challenor	8	8	48	51	9	7.67	5	4	1	1	2	2
8	Arch Ruben	8	8	48	68	10	8.93	6	3	1	3	3	
9	Lalo Norssister	7	7	42	39	2	4.23	19	1	2	2	2	
10	Anya Lande	5	5	30	64	5	5.12	12	1	3			2
11	Angelika Petty	6	6	36	74	6	7.92	12	4	1	4		
12	Brit Cockley	5	5	30	83	12	7.66	6	4	3	4	2	1
13	Martainn Rivalant	6	6	36	48	6	7.6	8	1	3	2	2	2
14	Leland L'Hommeau	5	5	30	84	12	7.3	7	2	4	1	4	5
15	Lizzy Hounsom	7	7	42	90	8	6.04	11	4	4	3	4	1
16	Beauregard Gatward	5	5	30	39	2	4.53	19	1	2	2	3	5
17	Gilda Bicknell	6	6	36	53	10	5.73	5	3	2			
18	Solly Corteney	6	6	36	50	10	5.62	5	3	3	4	2	2
19	Alma Raye	5	5	30	50	4	5.09	12	1	3	1	2	4
20	Dari Overland	6	6	36	79	7	5.83	11	3	2	3	4	6
21	Rikki Ollerenshaw	7	7	42	76	6	4.12	12	3	4	3	1	3

Figure 5.3: Bowling Data Overview

```
df_bw.describe()
```

	id	Matches	Overs Bowled	Bowls	Runs Given	Wickets Taken	Economy	Average	Performance Wickets 5	Performance Wickets 4	Performance Wickets 3	Performance Wickets 2	Performance Wickets 1
count	600.000000	600.000000	600.000000	600.000000	600.000000	600.000000	600.000000	600.000000	600.000000	600.000000	600.000000	600.000000	600.000000
mean	300.500000	6.473333	6.473333	38.840000	60.531667	6.593333	6.512350	14.576667	2.516667	2.500000	2.518333	2.478333	3.515000
std	173.349358	1.120140	1.120140	6.720839	17.437961	3.471028	1.419011	15.162694	1.123310	1.115979	1.115828	1.148214	1.703314
min	1.000000	5.000000	5.000000	30.000000	30.000000	1.000000	4.000000	2.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	150.750000	5.000000	5.000000	30.000000	45.000000	3.000000	5.295000	5.000000	1.000000	2.000000	2.000000	1.000000	2.000000
50%	300.500000	7.000000	7.000000	42.000000	61.000000	7.000000	6.495000	9.000000	3.000000	2.000000	3.000000	2.000000	4.000000
75%	450.250000	7.000000	7.000000	42.000000	75.000000	10.000000	7.812500	18.000000	3.000000	4.000000	4.000000	4.000000	5.000000
max	600.000000	8.000000	8.000000	48.000000	90.000000	12.000000	9.000000	90.000000	4.000000	4.000000	4.000000	4.000000	6.000000

Figure 5.4: Bowling Data Description

5.4 Data Training

As mentioned above the data was obtained from the web application where the two datasets are downloaded that is in csv format. Once downloaded there are two models that is the batting prediction model and the bowling prediction model so since there are two datasets the batting model will now predict the top six players suitable for the next match and the bowling model will predict the top 5 players for the next match and then these players will then be selected in the starting eleven of the next match. So, the process started when the data was downloaded and fed into the system where the data first was cleaned and pre-processed such that any data with null values was removed so that the data that is trained is consistent so that the accuracy returned is the highest possible.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Build and Train the Model
from sklearn.linear_model import LinearRegression

model = LinearRegression()
model.fit(X_train, y_train)

# Make Predictions for "Average"
predictions = model.predict(X_test)

# Evaluate the Model (Optional)
from sklearn.metrics import mean_squared_error, r2_score

mse = mean_squared_error(y_test, predictions)
r2 = r2_score(y_test, predictions)

print(f'Mean Squared Error: {mse}')
print(f'R-squared: {r2}')
```

Figure 5.5: Data Splitting and Training

```
Mean Squared Error: 0.907280885253129
R-squared: 0.9754574755602298
```

Mean Squared Error (MSE): MSE is a measure of how well the model's predictions match the actual values. In your case, a MSE of 0.907 is relatively low, suggesting that, on average, the model's predictions are close to the actual values.

R-squared (R2): R-squared is a measure of how well the independent variables explain the variability of the dependent variable. An R2 of 0.975 is very high, indicating that the model explains a large portion of the variance in the target variable.

These results suggest that the model, based on the selected features, is able to predict the 'Average' well.

Figure 5.6: Outcome of the Data Training

5.5 Testing

Testing is process of checking the system whether it is working as expected and to identify and solve any errors in the system before handing it over. This enables to solve the identified errors beforehand so that the system works as expected.

5.5.1 Black Box testing

Black box testing was used to check if the data input was correct and verified it and validated it against the database. This type of testing mostly allows the tester to know the top surface of what is happening and nothing deep that is it allows the tester to check if the system is outputting or doing whatever is expected after giving the correct data input. So, basically while carrying out black box testing the tester does not how the system is coded, they only focus on what output the system is returning when a data input is done. This black box testing paradigm was mainly used to confirm what happens when the tester inputs certain values for data when checking if the system will act accordingly or not(Romdhana et al., 2022).

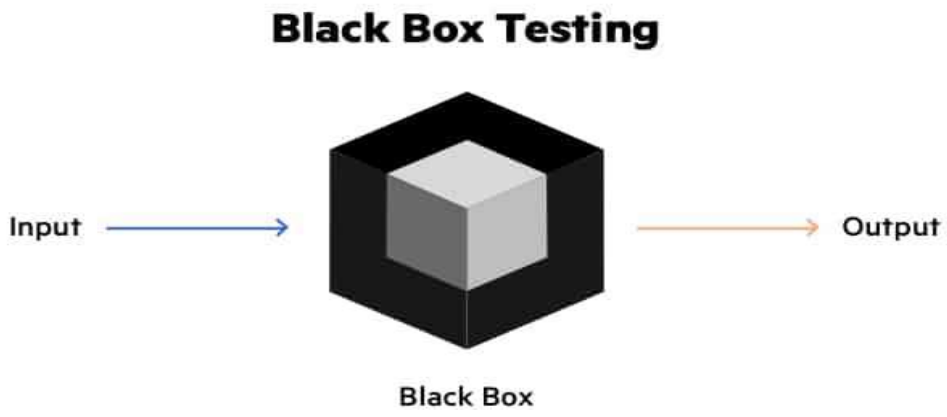


Figure 5.7: Black Box Testing (Romdhana et al., 2022)

5.5.2 Testing results

The following tests were carried out and the results and verdicts were as shown below.

Table 7: Test Case and Results

Test ID	Details	Test Data	Outcome	Verdict	Evidence
001	Trying to access the login system with an empty field. This should block the user entry into the system unit the correct data is entered and no fields are blank	Username: admin@admin.com Password: left blank	Error message displayed stating password field is required	Pass	Appendix 7.1
002	Trying to access the system by logging in with incorrect password. This should not allow access into the system until the correct credentials are entered.	Username: admin@admin.com Password: p'ssword Correct Password: password	Error message displayed stating incorrect password	Pass	Appendix 7.2
003	Registering a new club with an email that does not have the @ symbol. This should not allow to add or create a new club.	Club Name: Premier Email: 123 Ground: Jamhuri	An error message displayed suggesting to enter the symbol '@' in the email address	Pass	Appendix 7.3

004	Moving on to another form for data entry with some values not entered. This should not allow to access the other forms without one field being empty.	Moving to bowling first form from batting first form with a runs left empty for one of the players	An error message being displayed asking the tester to fill in the form where the required fields are not left empty before moving to the next form.	Pass	Appendix 7.4
005	Trying to fill in the batting form for the players that been selected did not bat. This should block out input for data entry if did not bat is selected.	Did Not Bat selected for a player.	It did not allow data entry until the selection was changed from Did Not Bat to Batted.	Pass	Appendix 7.5

Chapter 6: Conclusion, Recommendations and Future Works

6.1 Conclusion

The project idea was thought after identifying the gap in the cricketing body here in Kenya. As a player playing the Nairobi Provincial Cricket Association (NPCA) this gap was identified eyeing that the current system used to store data is an external app that is CricHQ which then again needs to be brought in by the cricketing body of Kenya that is Cricket Kenya who again has to import or manually add it to their system. Also, the basis of choosing the starting eleven did not seem efficient and that is why the results of the matches were mostly against the national team. Thus, the proposed system tried to help resolve this issue where there is a system that saves the data and generates the leaderboards reports which essentially is the top performing players. From this leaderboard report the top ten players from the batting side and the top ten players from the bowling side are generated such that their previous five performance records are saved and the average. This data is then used to predict the performance in the sixth match then compare it to the average to see which players will be performing well in the upcoming matches so that so that the best starting eleven is selected who will be performing well and so that the outcome of the match turns in favour of the team.

6.2 Recommendations

The main challenge faced during the development of the system was to update the leaderboards report. As much as the report looks easy it is not since it is just a view only report and that the report is a dynamic report meaning it changes every time matches are scored. So, after every match the leaderboard report is updated. So once the leaderboard report was working as expected then the issue of exporting also lasted for a while since the exported data that is the batting data or the bowling data needed to be in the exact format that is column names so that is can be read by the model a small error in the name was returning an error in the model.

6.3 Future Work

In vision to complete the project with limited resources the project had to be forced to work with a limited scope so that the deliverables can be met on time. With this in mind the project had a lot of compromises that needed to be done. Thus, for anyone else who wishes to undertake a similar project, they can add the prediction for the wicket keeper and fielders. Moreover, the system can be developed such that the whole system does scoring from to ball to ball unlike the developed system that is doing scoring on match-to-match basis. Ball to ball basis allows to deal with errors arising from transposition of data from the scorecard into the system unlike when the system is directly used to do the scoring. This will improve efficiency and accuracy of scoring done for the match and also will be able to store more details about players which can lead to better analysis of the player.

References

Agarwal, S., Yadav, L., & Mehta, S. (2017). Cricket Team Prediction with Hadoop: Statistical Modeling Approach. *Procedia Computer Science*, *122*, 525–532. https://doi.org/10.1016/j.procs.2017.11.402

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software Engineering for Machine Learning: A Case Study. *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, 291–300. https://doi.org/10.1109/ICSE-SEIP.2019.00042

Anik, A. I., Yeaser, S., Hossain, A. G. M. I., & Chakrabarty, A. (2018). Player's Performance Prediction in ODI Cricket Using Machine Learning Algorithms. *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT)*, 500–505. https://doi.org/10.1109/CEEICT.2018.8628118

Arteaga, D., Arenas, J., Paz, F., Tupia, M., & Bruzza, M. (2019). Design of information system architecture for the recommendation of tourist sites in the city of Manta, Ecuador through a Chatbot. *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–6. https://doi.org/10.23919/CISTI.2019.8760669

Berham, S., & Morris, S. (2022). A CRITICAL COMPARISON OF BRAVE BROWSER AND GOOGLE CHROME FORENSIC ARTEFACTS. *Journal of Digital Forensics, Security and Law*, *17*(1). https://doi.org/10.15394/jdfsl.2022.1752

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, *15*(1), 27–33. https://doi.org/10.1016/j.aci.2017.09.005

Dahiya, O., Solanki, K., Rishi, R., Dalal, S., Dhankhar, A., & Singh, J. (2022). Comparative Performance Evaluation of TFC-SVM Approach for Regression Test Case Prioritization. In V. Goyal, M. Gupta, S. Mirjalili, & A. Trivedi (Eds.), *Proceedings of*

International Conference on Communication and Artificial Intelligence (pp. 229–238). Springer Nature. https://doi.org/10.1007/978-981-19-0976-4_20

Delplanque, J., Etien, A., Anquetil, N., & Auverlot, O. (2018). Relational Database Schema Evolution: An Industrial Case Study. *2018 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 635–644. https://doi.org/10.1109/ICSME.2018.00073

Denscombe, M. (2017). *EBOOK: The Good Research Guide: For Small-Scale Social Research Projects*. McGraw-Hill Education (UK).

Fagarasan, C., Popa, O., Pisla, A., & Cristea, C. (2021). Agile, waterfall and iterative approach in information technology projects. *IOP Conference Series: Materials Science and Engineering*, *1169*(1), 012025. https://doi.org/10.1088/1757-899X/1169/1/012025

Frantiska, J. (2018). Entity-Relationship Diagrams. In Jr. Frantiska Joseph (Ed.), *Visualization Tools for Learning Environment Development* (pp. 21–30). Springer International Publishing. https://doi.org/10.1007/978-3-319-67440-7_4

G, S. R. K., Kalakota, S., & Harshavardhan, R. (2023). *Performance Prediction of Cricket Player Using Blockchain Enabled HMM Model*. 438–451. https://doi.org/10.2991/978-94-6463-252-1_47

Garverick, J. (2018). The Wireframe. In J. Garverick (Ed.), *Migrating to Azure: Transforming Legacy Applications into Scalable Cloud-First Solutions* (pp. 85–103). Apress. https://doi.org/10.1007/978-1-4842-3585-0_5

Gemmell, J. (2013). Reflections on World Cup 2011. *Sport in Society*, *16*(1), 1–4. https://doi.org/10.1080/17430437.2012.762296

Han, J., Deng, S., Xia, X., Wang, D., & Yin, J. (2019). Characterization and Prediction of Popular Projects on GitHub. *2019 IEEE 43rd Annual Computer Software and*

Applications Conference (COMPSAC), *1*, 21–26. https://doi.org/10.1109/COMPSAC.2019.00013

Harris, M. (2022, February 9). *List of ICC Cricket Playing Countries with Associate Cricket Nations*. Its Only Cricket. https://www.itsonlycricket.com/cricket-playing-countries

Hasanika, D., Dilhara, R., Liyanage, D., Bandaranayake, A., & Deegalla, S. (2021). Data Mining System for Predicting a Winning Cricket Team. *2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS)*, 92–97. https://doi.org/10.1109/ICIIS53135.2021.9660702

Hazelwood, K., Bird, S., Brooks, D., Chintala, S., Diril, U., Dzhulgakov, D., Fawzy, M., Jia, B., Jia, Y., Kalro, A., Law, J., Lee, K., Lu, J., Noordhuis, P., Smelyanskiy, M., Xiong, L., & Wang, X. (2018). Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective. *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 620–629. https://doi.org/10.1109/HPCA.2018.00059

Hossen, H., Shuvon, M. S. S., Barsha, J. B., Chy, A. A., Pathak, A., Hossen, H., Shuvon, M. S. S., Barsha, J. B., Chy, A. A., & Pathak, A. (2023). Ultimate cricket experience: Dynamic web app for a real-time scoring system in university cricket. *World Journal of Advanced Research and Reviews*, *19*(2), Article 2. https://doi.org/10.30574/wjarr.2023.19.2.1721

Jaber, H., Marle, F., Vidal, L.-A., & Didiez, L. (2018). Criticality and propagation analysis of impacts between project deliverables. *Research in Engineering Design*, *29*(1), 87–106. https://doi.org/10.1007/s00163-017-0254-7

Jhansi Rani, P., Vidyadhar Kamath, A., Menon, A., Dhatwalia, P., Rishabh, D., & Kulkarni, A. (2020). Selection of Players and Team for an Indian Premier League Cricket Match Using Ensembles of Classifiers. *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 1–6. https://doi.org/10.1109/CONECCT50063.2020.9198371

Lascu, A., Spratford, W., Pyne, D., & Etxebarria, N. (2020). Practical application of ecological dynamics for talent development in cricket. *International Journal of Sports Science & Coaching*, *15*(2), 227–238. https://doi.org/10.1177/1747954120908816

Malek, M.-B.-, Badhan, R. H., Shesir, M. I., & Fakir, N. H. (2018). *Squad selection for cricket team using machine learning algorithms* [Thesis, Brac University]. http://dspace.bracu.ac.bd/xmlui/handle/10361/12296

Matongera, T. (2021). *Sports Tourism: A documentation for economic and infrastructural development in Africa*. https://repository.up.ac.za/handle/2263/84093

Maunder, E., Kilding, A. E., & Cairns, S. P. (2017). Do Fast Bowlers Fatigue in Cricket? A Paradox Between Player Anecdotes and Quantitative Evidence. *International Journal of Sports Physiology and Performance*, *12*(6), 719–727. https://doi.org/10.1123/ijspp.2016-0389

Medeiros, J., Vasconcelos, A., Silva, C., & Goulão, M. (2018). Quality of software requirements specification in agile projects: A cross-case analysis of six companies. *Journal of Systems and Software*, *142*, 171–194. https://doi.org/10.1016/j.jss.2018.04.064

Mitsuyuki, T., Hiekata, K., Goto, T., & Moser, B. (2017). Evaluation of Project Architecture in Software Development Mixing Waterfall and Agile by Using Process Simulation. *Journal of Industrial Integration and Management*, *02*(02), 1750007. https://doi.org/10.1142/S2424862217500075

Newman, T. J., Lower, L. M., & Brgoch, S. M. (2019). Developing sport team captains as formal leaders. *Journal of Sport Psychology in Action*, *10*(3), 137–150. https://doi.org/10.1080/21520704.2019.1604590

Patil, M. M., Hanni, A., Tejeshwar, C. H., & Patil, P. (2017). A qualitative analysis of the performance of MongoDB vs MySQL database based on insertion and retriewal operations using a web/android application to explore load balancing—Sharding in

MongoDB and its advantages. *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 325–330. https://doi.org/10.1109/I-SMAC.2017.8058365

Philpott, S. (2018). The politics of purity: Discourses of deception and integrity in contemporary international cricket. *Third World Quarterly*, *39*(4), 677–691. https://doi.org/10.1080/01436597.2018.1432348

Premkumar, P., Chakrabarty, J. B., & Chowdhury, S. (2020). Key performance indicators for factor score based ranking in One Day International cricket. *IIMB Management Review*, *32*(1), 85–95. https://doi.org/10.1016/j.iimb.2019.07.008

Roberts, A. (2019). *A History & Guide to the Cricket World Cup*. Pen and Sword.

Romdhana, A., Merlo, A., Ceccato, M., & Tonella, P. (2022). Deep Reinforcement Learning for Black-box Testing of Android Apps. *ACM Transactions on Software Engineering and Methodology*, *31*(4), 65:1-65:29. https://doi.org/10.1145/3502868

Saikia, H., Bhattacharjee, D., & Mukherjee, D. (2019). Franchisee Cricket and Cricketers' Auction. In H. Saikia, D. Bhattacharjee, & D. Mukherjee (Eds.), *Cricket Performance Management: Mathematical Formulation and Analytics* (pp. 23–36). Springer. https://doi.org/10.1007/978-981-15-1354-1_2

Sen, N. (2018, August 25). *World Cup 2003—Rise of Kenya*. https://www.sportskeeda.com/cricket/world-cup-2003-rise-of-kenya

Setiyawan, N. F., Priyadi, Y., & Astuti, W. (2023). Development of Class Diagrams Based on Use Case, and Sequence Diagrams Using a Text Mining Approach in SRS Penguin. *2023 IEEE World AI IoT Congress (AIIoT)*, 0070–0076. https://doi.org/10.1109/AIIoT58121.2023.10174287

Shah, A. (2020). *Intelligent Cricket Team Selection by Predicting Individual Players' Performance using Efficient Machine Learning Technique*. *9*. https://doi.org/10.35940/ijeat.C6339.029320

Shah, P., & Patel, M. (2018). *Ranking the cricket captains using principal component analysis*. https://doi.org/10.13140/RG.2.2.33455.38564

Shetty, M., Rane, S., Pandita, C., & Salvi, S. (2020). Machine learning-based Selection of Optimal sports Team based on the Players Performance. *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 1267–1272. https://doi.org/10.1109/ICCES48766.2020.9137891

Singla, S., & Shukla, S. (2020). Integer Optimisation for Dream 11 Cricket Team Selection. *INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING*, *8*, 1–6. https://doi.org/10.26438/ijcse/v8i11.16

Son, N. T., Van Thanh, L., Duong, T. B., & Anh, B. N. (2018). A decision support tool for cross-functional team selection: Case study in ACM-ICPC team selection. *Proceedings of the 1st International Conference on Information Management and Management Science*, 133–138. https://doi.org/10.1145/3277139.3277149

Song, C., Ristenpart, T., & Shmatikov, V. (2017). Machine Learning Models that Remember Too Much. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 587–601. https://doi.org/10.1145/3133956.3134077

Stoddart, B. (2017). At the end of the day's play: Reflections on cricket, culture and meaning. In *The imperial game* (pp. 150–166). Manchester University Press. https://www.manchesterhive.com/display/9781526123824/9781526123824.00013.xml

Subramanian, H., & Raj, P. (2019). *Hands-On RESTful API Design Patterns and Best Practices: Design, develop, and deploy highly adaptable, scalable, and secure RESTful web APIs*. Packt Publishing Ltd.

Szymański, P., & Kajdanowicz, T. (2018). *A scikit-based Python environment for performing multi-label classification* (arXiv:1702.01460). arXiv. https://doi.org/10.48550/arXiv.1702.01460

Tatroe, K., & MacIntyre, P. (2020). *Programming PHP: Creating Dynamic Web Pages*. O'Reilly Media, Inc.

Team, T. oboloo. (2022, December 14). What is a Documentation? Definition. *Oboloo*. https://oboloo.com/blog/what-is-a-documentation-definition/

Technologies (CSIT), A. J. on C. S. and I., & MM, D. I. U. R., M. T. I. (2021). *APTIKOM Journal on Computer Science and Information Technologies (CSIT) Vol. 5 No. 1 March 2020*. Iaic Bangun Bangsa.

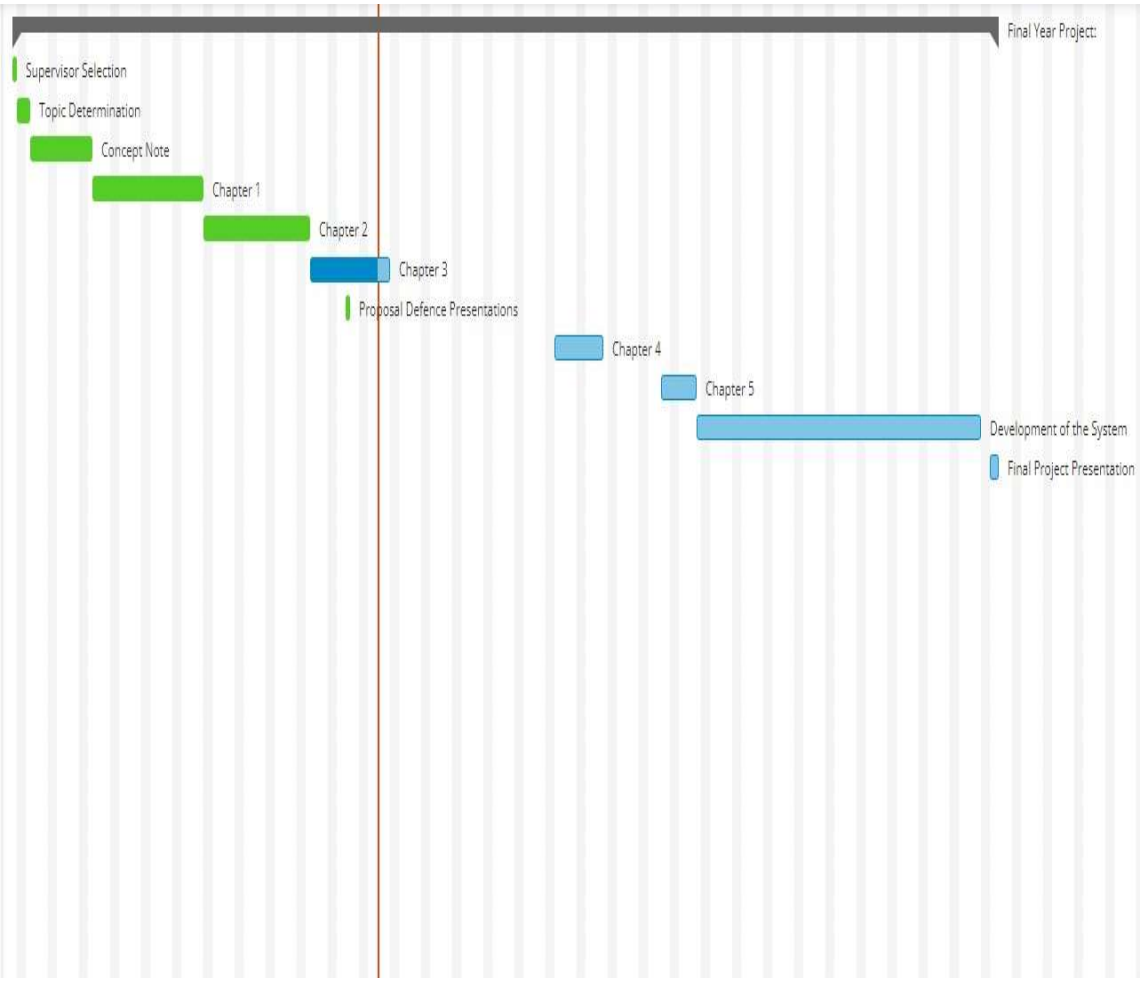
Turnip, L., Triayudi, A., & Solihati, I. D. (2020). Web Based Fixed Asset Management Information System Using the Waterfall Method (Case Study: National University): Web Based Fixed Asset Management Information System Using the Waterfall Method (Case Study: National University). *Jurnal Mantik*, *4*(1), Article 1.

Vachharajani, V., & Pareek, J. (2019). Framework To Approximate Label Matching For Automatic Assessment Of Use-Case Diagram. *International Journal of Distance Education Technologies (IJDET)*, *17*(3), 75–95. https://doi.org/10.4018/IJDET.2019070105

Valarezo, R., & Guarda, T. (2018). Comparative analysis of the laravel and codeigniter frameworks: For the implementation of the management system of merit and opposition competitions in the State University Península de Santa Elena. *2018 13th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–6. https://doi.org/10.23919/CISTI.2018.8399242

Appendix

Appendix 1



Appendix: Gantt Chart

Appendix 2

Strathmore University

School of Computing and Engineering Sciences

Information Systems Project Documentation Assessment Guide

Student Number(s) 136446

Working Title: Player Performance Prediction and Starting Eleven Selection Enhancer Using Machine Learning Algorithm For Cricket

Table 8: Assessment Guide

Evaluation Points	**Weight**	**Score**	**Notes**
Title	1		
Abstract *Updated to include chapter 1-6*	3		
Chapter 1-3 **Checking previous documentation chapters for the correctness of title and problem statement, project scope as implemented and change of tenses*			
Problem Statement	1		
Justification	1		
Scope	1		
Limitation	1		
Literature Review	2		
Methodology	2		
Chapter 4 Correct functional requirements Correct non-functional requirements System Architecture and accompanying literature 4 Design diagrams and accompanying literature	 3 3 2 4		
Chapter 5 Setup Description: Hardware, software, support libraries, frameworks, versions and compatibility	6		

<table>
<tr><td>Description of how the solution works to meet problem and business needs
Description of the test environment, data, test case
<table><tr><td>Functional Requirement</td><td>Test Data</td><td>Expected Result</td><td>Actual Result</td><td>Pass/Fail</td><td>Evidence</td></tr></table>
*Check 3 core functional requirements and evidence of test available as appendix</td><td>3

6</td><td></td><td></td></tr>
<tr><td>Chapter 6
Valid Conclusion
Sound Recommendation</td><td>2
2</td><td></td><td></td></tr>
<tr><td>Presentation
Document Structure as per template provided and grammar
Citation and References
Document Numbering and Table of Contents/figures
Existence of required appendices</td><td>
2
2
2
1</td><td></td><td></td></tr>
<tr><td>Total Marks</td><td>50</td><td></td><td></td></tr>
</table>

Comments

_____

_____

_____

_____

Examiner Name:_____ Signature:_____

Appendix 3: Plagiarism Report

Document Information

Analyzed document	136446-Proposal Document.docx (D179615650)
Submitted	2023-11-23 08:46:00
Submitted by	
Submitter email	Mirav.Bhojani@strathmore.edu
Similarity	2%
Analysis address	ttunduny.strath@analysis.urkund.com

Sources included in the report

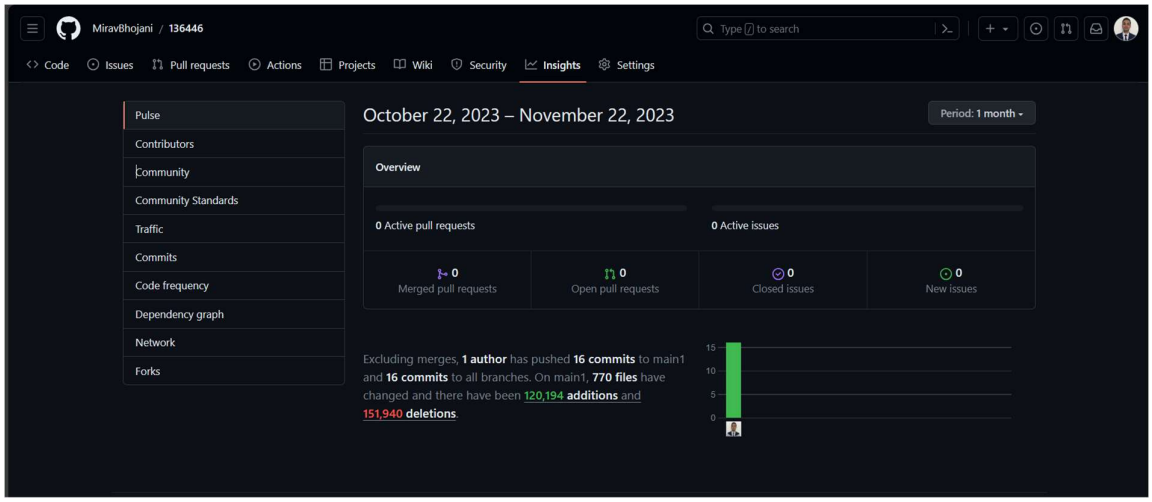
SA	**Final_Year_Project_Report-1.pdf** Document Final_Year_Project_Report-1.pdf (D168695365)	1
SA	**10609062_Thesis_Report.pdf** Document 10609062_Thesis_Report.pdf (D173232625)	3
SA	**Utilizing Machine Learning for Sport Analytics in Cricket(BJIT).docx** Document Utilizing Machine Learning for Sport Analytics in Cricket(BJIT).docx (D172045330)	3
SA	**Synopsis Big Data Sports-1.docx** Document Synopsis Big Data Sports-1.docx (D110980193)	1
SA	**Ass2_Kingshuk_Ghosh_R00208685.pdf** Document Ass2_Kingshuk_Ghosh_R00208685.pdf (D136581766)	1
SA	**CM50170_Research_Report.pdf** Document CM50170_Research_Report.pdf (D112411676)	1

Appendix 3: Plagiarism Report

Appendix 4: GitHub Analytics

Appendix4.1: Pulse

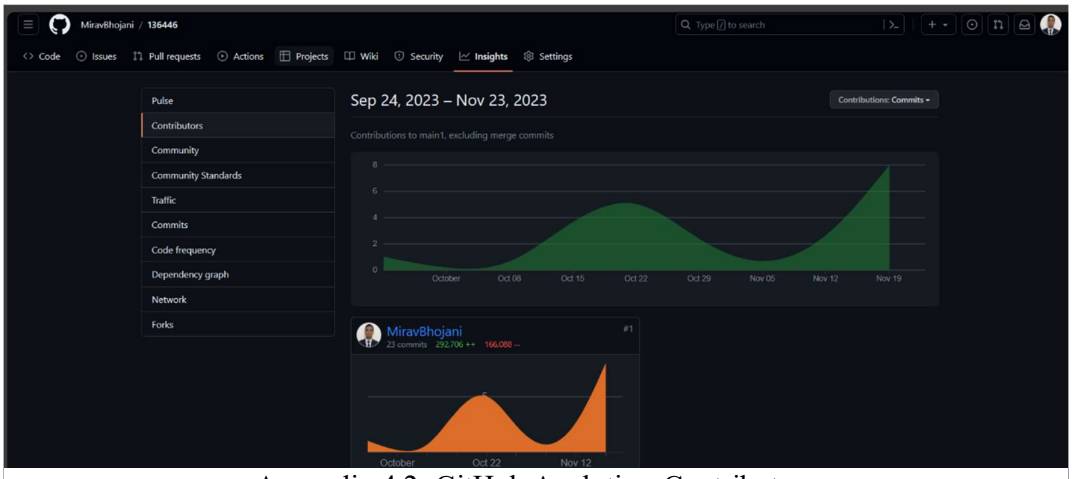
Diagram below gives analytics about the pulse which is basically the contributors activity over a period of a month that was between 22nd October 2023 and 22nd November 2023. Also shows all the active and closed issued and the pull and pull requests that were made and mostly about the number of files that have changed in the folder. As evident below there were 16 commits in the specified time period with 120,194 additions and 151940 deletions.



Appendix 4.1: GitHub Analytics: Pulse

Appendix 4.2: Contributors

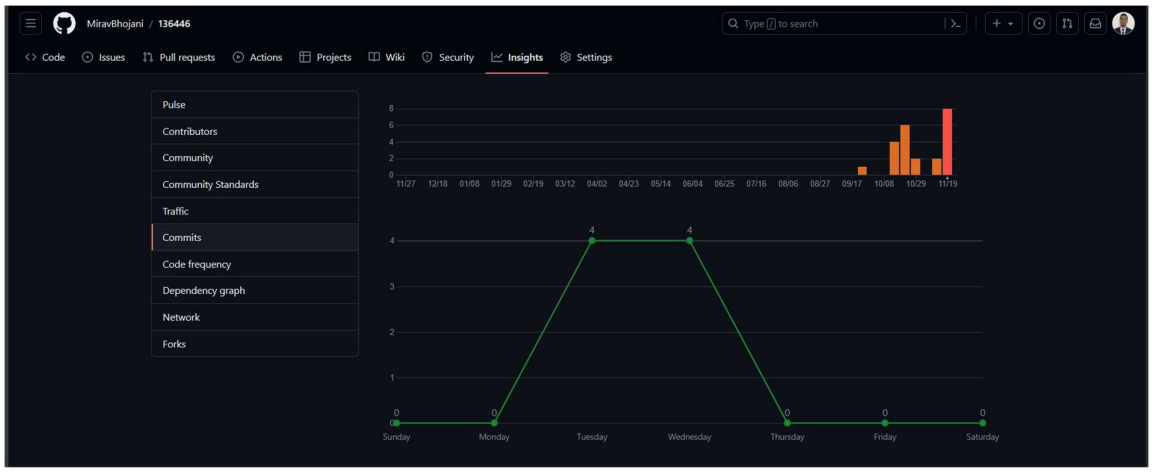
This part in GitHub shows the number of contributors in the repository and the number of commits they have done and to which branch were the commits made to. The diagram below shows evidence of the contributors' section from GitHub.



Appendix 4.2: GitHub Analytics: Contributors

Appendix 4.3: Commits

This part of the GitHub analytics insights for commits shows the number commits that were made to the repository and the changes that were made to the codebase over time. The diagram below shows the number of commits that were made to the repository.



Appendix 4.3: GitHub Analytics: Commits

Appendix 5: User Manual

The system provides access to two user roles: Admin and Club Admin. Each has distinct functionalities within the platform.

Login Credentials:

Admin:

- Username: admin@admin.com
- Password: password

Admin Dashboard

Functions Available:

1. Add New Club:
    - Access a form requiring club details (name, email, home ground) for registration.
    - An email with the login password will be sent to the registered email address for club access.
2. Create a Match:
    - Fill a form to schedule a match between existing clubs.

3. View & Export Leaderboards Report:

- Obtain team selection insights by exporting leaderboards for upcoming matches.

Club Admin Dashboard

Functions Available:

1. Add Players to Club:

- Enter player details (name, DOB, batting style, bowling style) to register players.
- Registered players are eligible to represent the club in matches.

2. Score Matches:

- Upon match schedules, access scoring forms in the "Matches" section.
- Complete four forms with match scorecard details.

3. Leaderboard Updates:

- Top ten batsmen and bowlers are updated after match scoring.
- Export these reports in CSV format to be loaded into the predictive model.

Predictive Model

- Click the choose file button and upload the exported csv files to the model to have them predict the best performing players for the next match.

Appendix 6: Developer Manual

Prerequisites

Ensure the following software tools are installed:

- VSCode: To facilitate code changes.
- XAMPP: Acts as the web server, managing both database and system runtime.
- Anaconda: Required for accessing the terminal to run the predictive models deployed on Flask.

Setup Instructions

- Repository Access:

Clone or download the system's repository from the GitHub link provided in Appendix 4.

Extract or clone the repository into the htdocs folder within Local Disk C.

- Code Configuration:

Open the system's folder in VSCode.

Modify the file paths as per your system's configuration, especially the model locations.

- Server Launch:

Run XAMPP, ensuring both Apache and MySQL modules are activated.

Launch the browser and access the system via the provided Login Page link.

- Model Deployment:
  - To run the batting model:

Open the Anaconda terminal as administrator.

Navigate to the directory storing the batting model.

Execute the command: python app.py.

Copy the returned link to access the batting model on the browser.

  - For the bowling model:

Open a separate Anaconda terminal as administrator.

Navigate to the directory storing the bowling model.

Execute the command: python app.py.

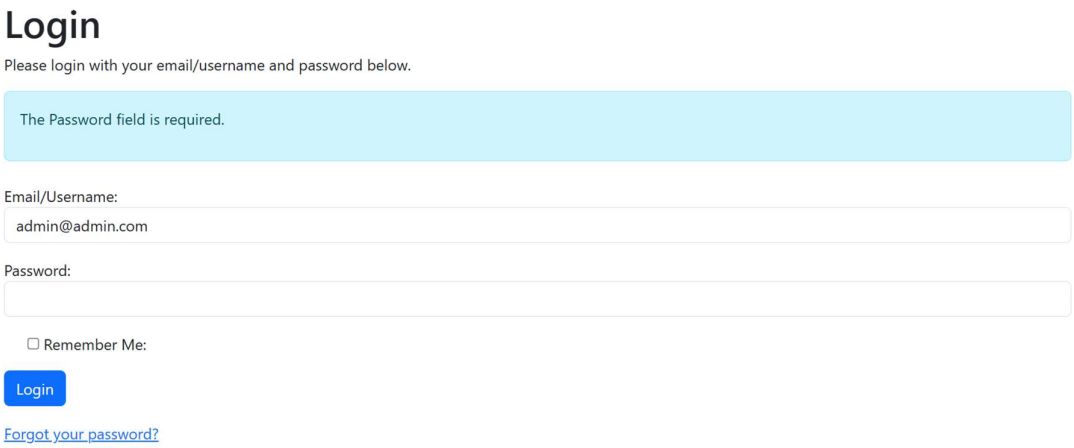
Copy the link returned and paste it into the browser to access the bowling model.

System Usage

Upon successful completion of the setup steps, users can access the system's functionalities seamlessly.

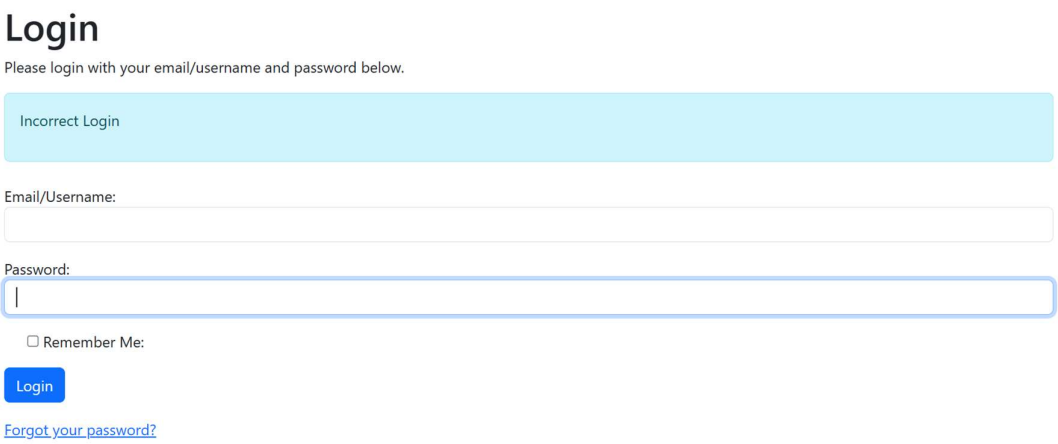
Appendix 7

Appendix 7.1



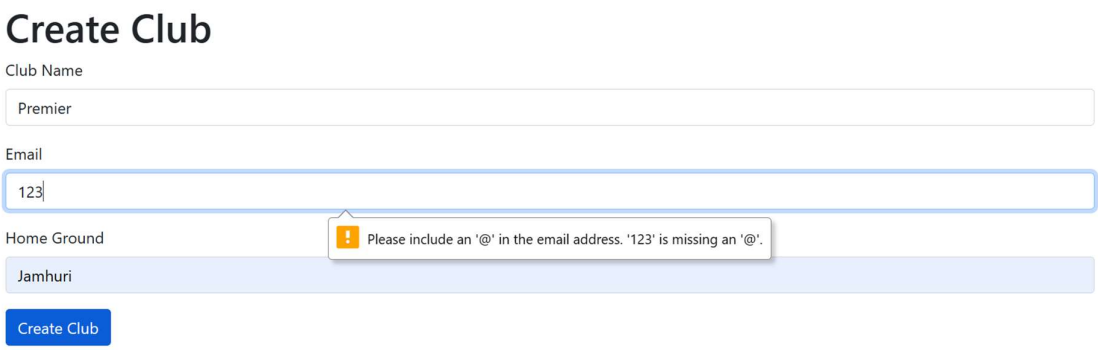
Appendix 7.1: Black Box Test ID 001

Appendix 7.2



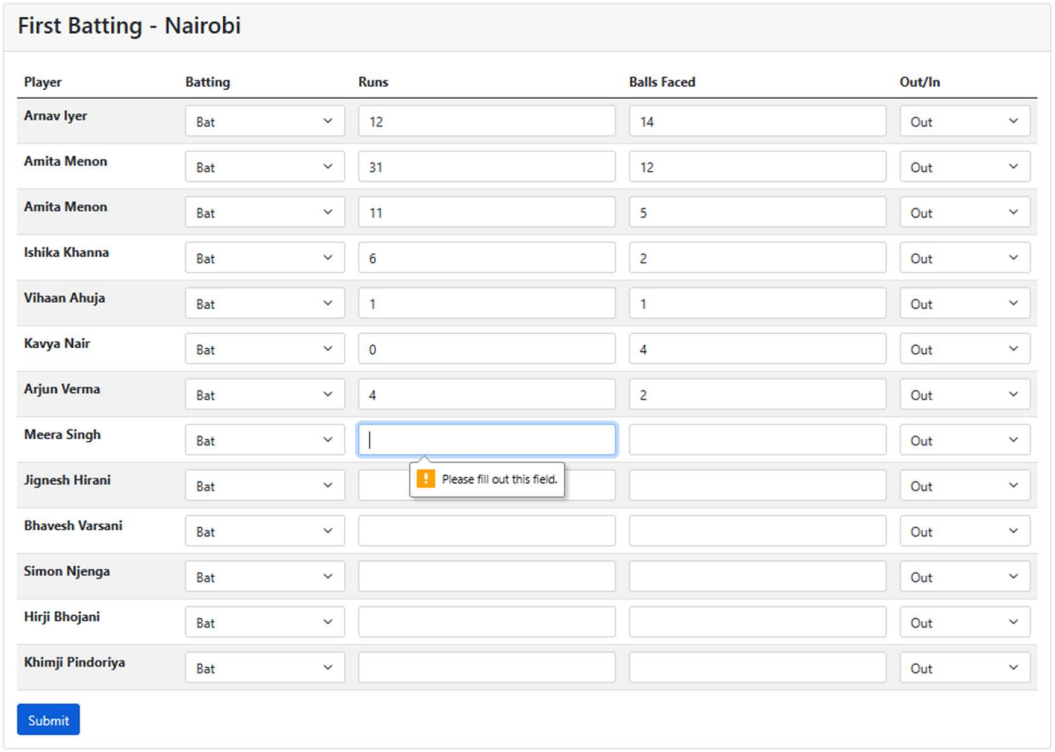
Appendix 7.2: Black Box Test ID 002

Appendix 7.3



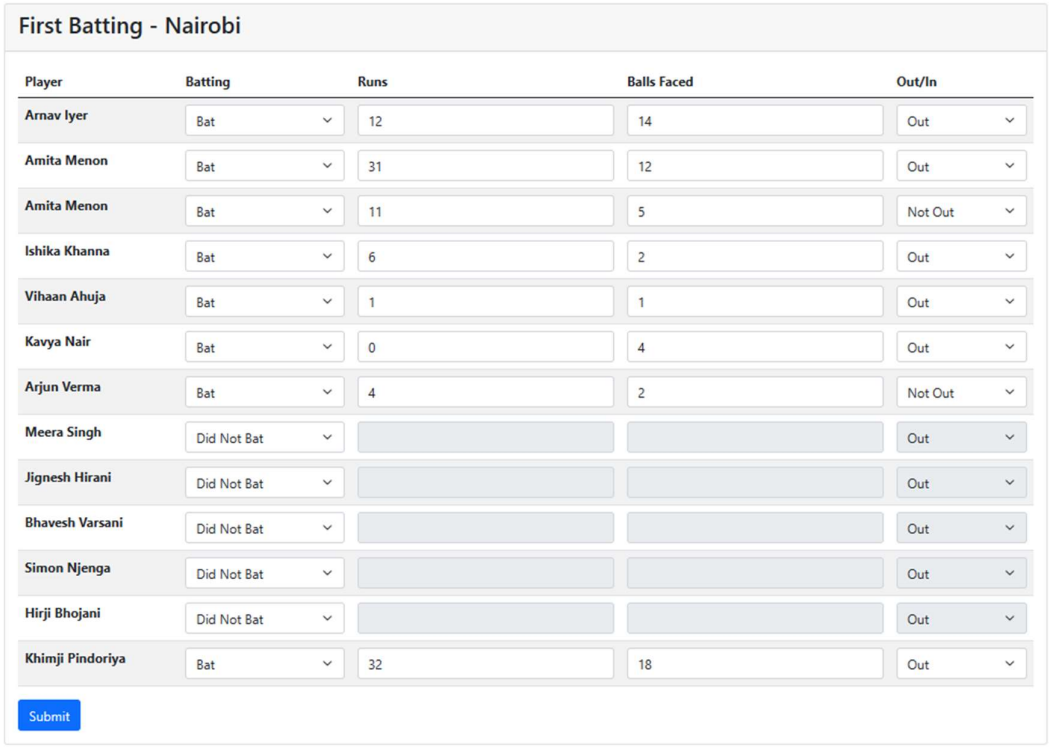
Appendix 7.3: Black Box Test ID 003

Appendix 7.4



Appendix 7.4: Black Box Test ID 004

Appendix 7.5



Appendix 7.5: Black Box Test ID 005