Identification and Selection of Rising Performers in T20 Cricket Using Machine Learning Algorithms

MSc Research Project
Data Analytics

Arkoprovo Sarkar
Student ID: x19148038

School of Computing
National College of Ireland

Supervisor: Aaloka Anant

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Arkoprovo Sarkar
Student ID:	x19148038
Programme:	Data Analytics
Year:	2022-2023
Module:	MSc Research Project
Supervisor:	Aaloka Anant
Submission Due Date:	15/12/2022
Project Title:	Identification and Selection of Rising Performers in T20 Cricket Using Machine Learning Algorithms
Word Count:	6335
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	31st January 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Identification and Selection of Rising Performers in T20 Cricket Using Machine Learning Algorithms

Arkoprovo Sarkar
x19148038

Abstract

Local twenty20 cricket leagues such as the Indian Premier League and others have seen an increase in the amount of money franchise owners are willing to spend on quality players between the ages of 20 and 30. In cricket, the most important and time-consuming job is player recruitment since it determines the team's chances of winning. If a team overpays for the incorrect player, they risk losing the championship and maybe millions of dollars. Because of this, there has been a lot of work put into developing machine learning models that can forecast a cricket player's performance. This research is being done with the intention of determining whether it is possible to utilize machine learning to pick out potential young and middle aged players aged 20 to 30, on the basis of their past statistics. In this particular study, two distinct approaches to machine learning have been used. In the course of research, machine learning algorithms performance of Random Forest and Naive Bayes has been measured using a variety of metrics like accuracy, precision, and so on. Both of these models have been used to make predictions about the number of runs scored by the batsmen and the number of wickets taken by bowlers. It was discovered that the Random Forest Classifier was the one that was the most accurate for predicting both runs scored and wickets taken.

1 Introduction

Cricket is played between two teams of eleven players each. It was initially played in England in the sixteenth century. (Kamble et al.; 2021). One team bats to score runs, while the other bowls to stop them from scoring. Team trying to score maximum amount of runs are known as batting team while the team trying to restrict the batting team from scoring runs by taking as many wickets as possible are called bowling team. On a worldwide scale, it is competed in three distinct formats: test match, one-day international (also known as an ODI) and twenty20 match (T20). Cricket has seen a resurgence in its popularity since the twenty20 format has been launched. This format has helped cricket restore its prior level of popularity. People of South Asia like watching and participating in this game the most right now. The adoption of twenty20 format has been a significant factor in crickets growing appeal in a number of European countries. In fact, the Cricket Council has disclosed the new format that will be used for twenty20 world cup in 2024. According to this new format, each of the continents of Africa, Asia and Europe will be allotted two qualification spots, while both East-Asia Pacific region and America will receive one qualification spot. Due to its meteoric rise in popularity, our research will entirely focus on twenty20 format. In the quest of victory, the most essential thing to do is to make sure you choose the right players. As a result, it would be useful to have a prediction model that can assist in determining whether or not a player has the potential to become a star performer on a consistent basis in the foreseeable future.

1.1 Research Motivation & Project Background

Cricket teams and the administration of the sports have spent a significant amount of time trying to find a method that would allow them to choose the suitable individuals. A poor choice of players may result in a multitude of issues, including loss of games and significant sums of money. Finding cricket player between the ages of 20 and 30 who has the potential to become a high performer in the near or distant future calls for extensive research on the qualitative and quantitative characteristics of the available players. These characteristics include the number of runs a batsman has scored in a variety of situations and settings, against a wide range of opponents, under various conditions. Therefore, the thing that would be the most beneficial would be a prediction model that could yield more objective outcomes in a shorter length of time.

Using machine learning techniques to predict how the players will perform and how the game will turn out to be, has been the subject of a lot of research. However, the vast majority of attention has traditionally been concentrated on the longer version of the game which are ODIs and test. Passi and Pandey (2018) predicted the performances of players by examining parameters such as the number of runs scored, the number of wickets taken, and so on; however, their methodology was only relevant to one-day international cricket. Brown et al. (2022) presented their work that was most pertinent to this topic. In their research, the authors analyzed the performance trajectories of bowlers and batsmen in the sport of cricket from junior level to the older professional level by analyzing the performances of players belonging to different age groups. The number of games or the data that were gathered for the study was, however, somewhat limited, which might have had an effect on the findings and the overall conclusions of the study.

1.2 Research Question

"Is it possible to identify young and middle-aged cricket players, between the ages of 20 and 30 who have the potential to become great stars in twenty20 cricket in future, using machine learning technology?"

1.3 Research Objective

- To analyze previously published works and find methodologies and formulas that may be adopted and utilized for proposed research and evaluation metrics.

- Players data collection from a reliable source.

- Implementation of Machine Learning models that have been proposed for forecasting future performance of players between the ages of 20 and 30.

- Evaluation of the suggested methods, including a comparison of its output to that of previously completed research and an assessment of whether or not it yields better results.

2 Related Work

In this part, the influence of a variety of attributes on the performance and value of players has been explored first. Following that, a discussion on how predictive models have been used to anticipate the future performances of players in a variety of sports industries has been presented and then this section has been concluded with a brief discussion on several models that have been used to make predictions on the outcomes of matches.

2.1 Analysis of various attributes that affects players performance

Jayalath (2018) presented a method for analyzing the predictors of one-day international cricket matches. Occasionally, external variables affect the result of a match or a team's performance. In this study, variables like the home ground advantage, toss outcome, batting first or chasing were examined. A logistic regression model was devised, which, when applied to previous outcomes, can determine the likelihood of a team's victory against a certain opposition. According to this model, having the game played at one's own stadium has a significant advantage for the vast majority of teams. Additional study was performed on the data by classifying it according to the time in which the match was played, and it was discovered that the toss is a major influence. They came to the conclusion that the outcome of the coin toss would be the best factor to employ while developing a classification tree model to forecast the result of a match. In addition,the predictor was altered, which was the result of the match, to the margin of victory and conducted a regression tree approach as it helped provide a clearer interpretation of the results.

Metelski (2021) performed a research based on the Ekstraklasa which is Poland's top division football league. The analysis examined all transfers which were made throughout the history of the league. It was observed that close to 100 transfers were made, with a total expenditure of more than one hundred million euros on them. The primary objective of this study was to identify the qualities that set football players apart from one another in terms of their market worth using descriptive statistical analysis. It was observed that age along with performance plays a very important role in determining a players market value. According to the findings, the majority of players who were getting transferred to different leagues were between the ages of 21 and 24. It was also observed that group of players aged 21 and under received the highest transfer prices overall.

2.2 Forecasting performance of the players

One of the most well-known instances of player performance forecasting would be Anik et al. (2018). Here the researchers attempted to anticipate a cricket player's performance by using several machine learning methodologies, such as linear regression and support vector machines, as well as feature selection strategies, such as recursive feature deletion and univariate selection. Information on the members of the Bangladesh national cricket team, including player statistics, was compiled from reliable sports sources. The gathered statistical data was first translated into numerical values and then applied in the algorithms in order to implement the specified model. Thereafter, machine learning algorithms were used to forecast how many runs the Bangladeshi batsmen will score and how many runs their bowlers will concede in the upcoming match. The authors were able to achieve 91.5% accuracy with the SVM Kernel model for batter Tamim. Bowler Mahmudullahs data obtained 75.3% accuracy using the support vector machine model. However very limited attributes like balls faced, runs scored per match, number of boundaries, sixes, opposition, etc were considered for this research.

Iyer and Sharda (2009) used neural networks to make predictions about each cricket player's future performance based on the analysis of their previous results. The players on the cricket team were ranked in one of three categories: top performer, average, or failure. Player performance data from 1985 onwards up to the 2006–2007 season was gathered first. After that, the neural network models were gradually trained and evaluated utilizing a total of four different sets of data. Subsequent to that, the trained neural network models were used to provide a prognosis of the cricketer's performance in near future. Cricketers were suggested for the 2007 World Cup based on the ratings produced.

Machine learning models were implemented by Oytun et al. (2020) to forecast female handball players' performance and to explore the elements that significantly influenced those models predictions. Multiple techniques like linear regression, decision tree, support vector regression, etc were used and for every machine learning model a total of 118 occurrences of training patterns and 23 parameters were recorded. With the help of radial basis function neural network, the researchers were able to demonstrate that nonlinear relationships can be established between a variety of physical and exercise parameters in female handball players.This model outperformed other models and was able to predict the kinds of athletic performance that were evaluated with R2 values ranging from 0.86 to 0.97.

Brown et al. (2022) conducted an analysis with the intention of developing age-specific benchmarks for young cricketers in order for them to achieve senior professional status. This was accomplished by performing one-way ANOVA tests in order to analyze data collected on match performance based on county and age bracket. This work was done in order to more accurately characterize bowler and batsman's growth trajectories over a period of time. Very young male cricketers from professional English first-class county club were the only focus of the study. Each player's performance in games were used to assign them to one of two skill groups. Statistics like bowling and batting averages, wickets, runs scored and so on were analyzed. Differences in the points at which bowling and batting performance statistics concur with professional status were discovered. Despite this, it is possible that the findings and conclusions of the research were influenced by the relatively small size of the sample data.

2.3 Forecasting outcome of matches

Bunker and Thabtah (2019) have offered a critical analysis of the existing research in the field of machine learning, with a particular emphasis on the use of artificial neural networks (ANN) to forecast the outcomes of sporting events. They were able to recognize the learning methods that were used, data sources, suitable means of model evaluation and key challenges of forecasting sport results. As a result, they were able to propose a novel framework for sport prediction through which machine learning can be used as a learning approach. Some current ANN-based sports prediction studies were critically analysed in this article. Following that, a 'SRP-CRISP-DM' framework for sport result prediction was presented to address the complicated challenge of sport outcome prediction. Furthermore, the difficulties encountered by the sport prediction software were shown in order to identify future work for researchers in this critical application.

Using players data of a specific tournament, Jayanth et al. (2018) suggested a model for predicting cricket match outcomes and ideal team structure. They also suggested creating a system that nominates or recommends player for a certain position on a team based on their previous performances.After being gathered from a variety of open-source sources, the unstructured data was then stored into the database. A number of different statistical methods were then utilized to quantify the players performances in order to establish where each player stood in the rankings. The machine learning algorithm used these player rankings as input in order to make predictions about the result of matches and the structure of ideal teams. Linear, poly and RBF were used to train the SVM model. While testing the model, it was seen that RBF had done much better than the other two approaches with an accuracy of 75%. The overall accuracy was the on lower side.

For the purpose of forecasting which side will win a match, Vistro et al. (2019) presented a model that was built using machine learning techniques. The researchers focused on a variety of aspects including the performance of individual players and teams as well as the venue and the weather. In both testing phase and training phase, several machine learning classifiers were used. These are the ones that were utilized: random forest, support vector machine, naive bayes, logistic regression, and decision tree. The accuracy of the results that the decision tree predicted was 76% out of the multiple classifiers that were employed for this model. After some minor adjustments were made to the model's parameters, the performance increased from 76% to 94%. In the same fashion, random forest initially had an accuracy of 71%. The parameters were tuned more precisely, which led to an improvement in accuracy to 80%. In the end, XGBoost had an accuracy of 94.2% when it came to its predictions of the outcomes.

In this study, Danisik et al. (2018) provided the results of a number of experiments that were carried out by them in accordance with their recommended design, which was based on the data set containing both match related information and the player related information. All the player related information were also collected from a video game known as FIFA football and were combined with the ones which were previously gathered from real match data set. Post that, in order to provide the most accurate results possible, a number of configurations using cross-validation were tried and tested. LSTM regression model was observed to be the one that performed the best with an accuracy of 52.4%.

Cornman et al. (2017) attempted to forecast the results of future tennis matches by analyzing data from previous tennis matches. They also attempted to leverage the forecasts generated by the resultant model to get an advantage over the odds that were already being offered. After putting the prediction and gambling models in place, they were successful in predicting the results of 69.6% of the matches which were held in the year 2016 and 2017. They were also able to make a profit of 3.3% per game. Variety of machine learning models were used in this research. SVM was tested with multiple kernels. In the end, Linear kernel provided the best performance.

3 Methodology

This section provides an in-depth look at the technique that has been used for this project. In addition, it provides justification for using KDD technique rather than any other methodology.

Cleaning, analyzing, and interpreting data have always been required steps in the process of transforming data into knowledge D'Oca et al. (2015). Research that is driven by data mining demands extensive processing and thoughtful decision making. As a result, during our study, we have made use of Knowledge Discovery in Databases (KDD). KDD outlines the complete process of information retrieval and draws significant conclusions from the collected data. Because of this, the majority of its applications may be found in the academic and business worlds. Its a closed-loop continuous feedback mechanism, which iterates between phases as needed by algorithms and pattern interpretations. Because KDD is primarily concerned with data mining rather than project management, it is well-suited for classification and forecasting purposes. The steps involved in the KDD process are shown in figure 1.

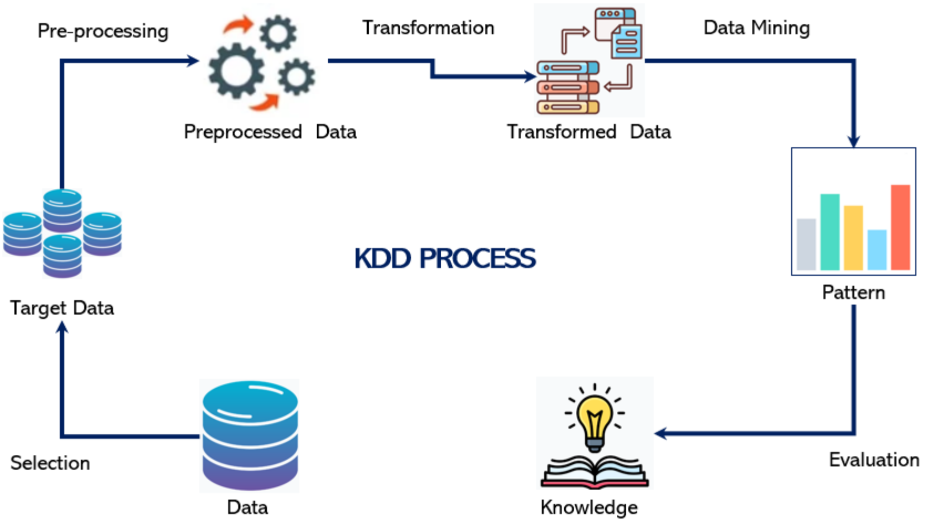


Figure 1: Knowledge Discovery Database Methodology

Establishment of goals is a fundamental prerequisite to the KDD process. It is the initial phase in the process and calls for an existing understanding as well as expertise in the field for it to be used in. This is the stage where we determine how the transformed data and discovered patterns could be utilized to derive knowledge. This part is incredibly essential, and if it is not stated correctly, it may lead to incorrect interpretations, which in turn can have a detrimental influence on the end-user. Steps involved in KDD process are shown below:

3.1 Data Selection

During the phase known as "data selection," task-relevant information is gathered and retrieved. Because they serve as the framework for data mining and have an impact on the types of data models that are developed, the parameters selected or extracted during this phase are very important.

3.2 Data Cleaning & Pre-Processing

This phase covers looking for data that is missing and deleting data from the data collection that is noisy, repetitive, or of poor quality. Feature engineering is an essential component of this process. The goal of this step is to enhance the data's dependability and its capacity to accomplish its intended purpose. The accuracy of the machine learning model is susceptible to fluctuations if the data is not refined properly. This phase often accounts for 70 to 80 percent of the total effort that is spent.

3.3 Data Transformation

In this stage, the data that will later be processed by the data mining algorithm are prepared. As a result, the data have to be presented in a consolidated and aggregated manner. If we supply categorical data to machine learning models that are dependent on numerical computations, then these models may transform the data into the improper format, which may impact the accuracy of the models. therefore, it is necessary to transform all categorical data into a format that can be represented by numbers.

3.4 Data Mining

This step is the most fundamental part of the overall KDD process. At this stage, algorithms are used to retrieve patterns from the final data, which are then utilized to assist in the development of prediction models. The insights unearthed by data mining have the potential to be beneficial in the areas of marketing, sports and various other industries. Following data mining algorithms have been utilized for this study:

3.4.1 Random Forest

When random forest was first implemented it was observed that the issue of over fitting and excessive variance that emerges as tree length grows can be reduced to a minimum by its usage. Bags are produced in this procedure, which involves selecting a subset of records and columns and each bag has a decision tree fitted into it. Instead of merely being able to extract patterns from significant features or variables, bagging offers a model that can extract patterns from all of the features.

3.4.2 Naive Bayes

Naive Bayes classifier can be implemented without any rigorous coding Yang (2018). This model is a general-purpose toolkit that can be used in a wide range of categorization fields. It is fundamentally based on Bayes's theorem, with the strong assumption that features are independent of one another. It prioritizes each feature, which distinguishes it from all other models.

4 Design Specification

First, both ball by ball and match details datasets have been loaded into the environment. Following that, both sets of data were combined and subjected to preprocessing. The merged dataset only had a small number of variables, which did not provide us with sufficient information for making predictions. Therefore feature engineering was carried out on the data using the information that was available in the combined dataset, we generated multiple batting and bowling attributes using various mathematical formulas. After that, those attributes were classified into separate bins. Post that, 2 different models Random Forest and Naive Bayes were used for the prediction. Finally, classification assessment methods were used in order to analyze the outputs of the models. The entire design specification is shown the form of an image in figure 2

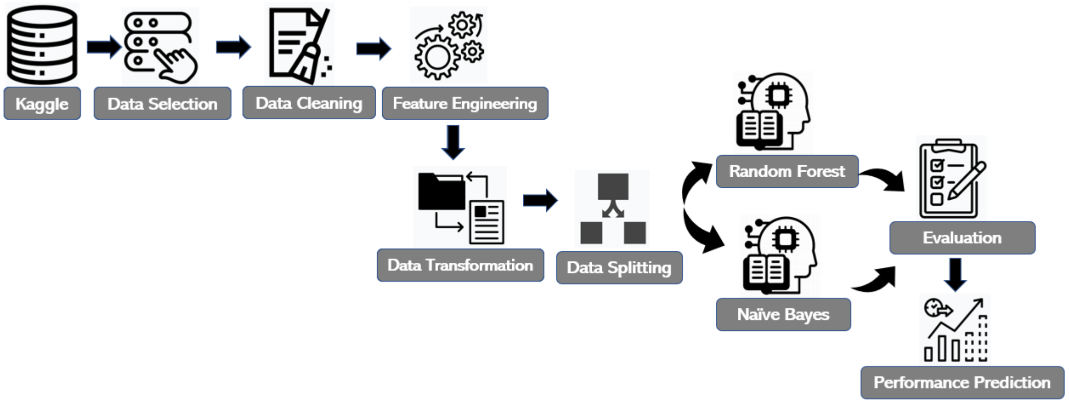


Figure 2: Research Structure

The following hardware and software specifications serve as the environment in which this study was conducted.

Hardware Specification: Core i5 10th Generation processor having a windows 10 operating system with a hard-disk storage capacity of 256 GB has been used for this research.

Software Specification: Version 3.8 of Python has been used for the implementation of this research and Jupyter Notebook is utilized for the coding of the project. The data for both training and testing were stored using the Numpy package. The data was divided or split with the help of the sci-kit learn library. Sci-kit Learn libraries have been used for both naive bayes and random forests. We have also used the Power BI software to generate charts and graphs from the collected data. The tools and technologies used for this research have been shown in the figure 3 below :

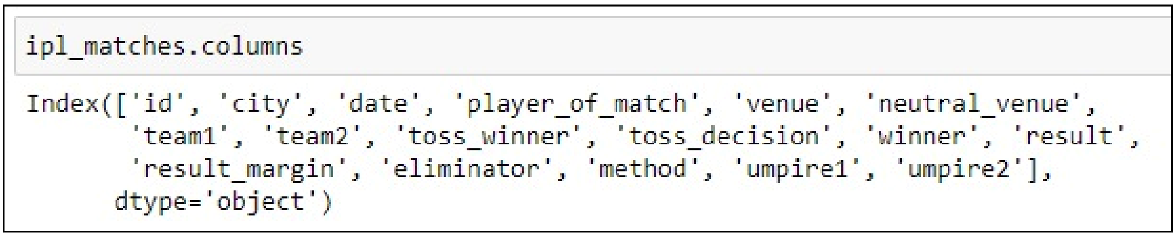
Figure 3: Tools and Technologies used for this research

5 Implementation

This section contains full explanation of the dataset and the attributes which have been derived and used for the prediction. In addition to that, models that were used have also been addressed.

5.1 Data Selection

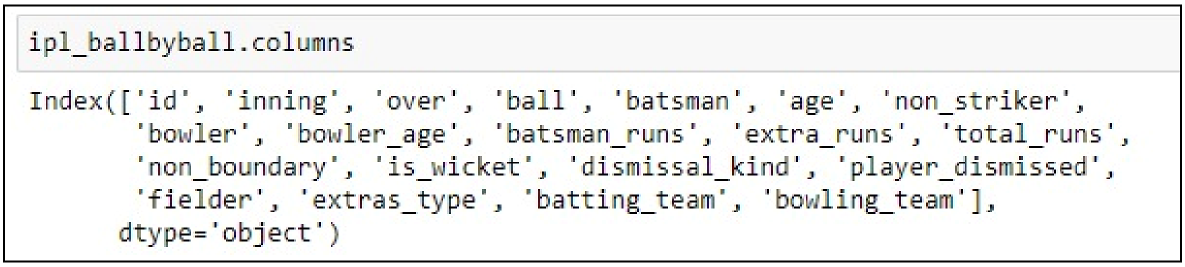
The dataset of the Indian Premier League (IPL) has been used for this research. The CSV dataset [1] was obtained from Kaggle. This dataset includes information of 800 plus games that were played between 2008 and 2020. The retrieved dataset had two files: a match overview file, which includes columns like city, date, player of the match, venue, toss winner, etc as shown in figure 4, and a ball-by-ball detail file that includes columns like over number, ball number, batsman name, non-strikers name, bowler name and other player relevant information shown in figure 5. Information from both sets of data were included in the study that we conducted.



```
ipl_matches.columns

Index(['id', 'city', 'date', 'player_of_match', 'venue', 'neutral_venue',
       'team1', 'team2', 'toss_winner', 'toss_decision', 'winner', 'result',
       'result_margin', 'eliminator', 'method', 'umpire1', 'umpire2'],
      dtype='object')
```

Figure 4: Columns in Match_Details dataset



```
ipl_ballbyball.columns

Index(['id', 'inning', 'over', 'ball', 'batsman', 'age', 'non_striker',
       'bowler', 'bowler_age', 'batsman_runs', 'extra_runs', 'total_runs',
       'non_boundary', 'is_wicket', 'dismissal_kind', 'player_dismissed',
       'fielder', 'extras_type', 'batting_team', 'bowling_team'],
      dtype='object')
```

Figure 5: Columns in Ball_by_Ball dataset

5.2 Data Preprocessing

It is necessary for the dataset to be in the suitable format in order to get better outcomes from the machine learning models. The transformation of unprocessed, raw data into processed, clean data is what's meant by "data pre-processing." Data fetched from Kaggle were preprocessed using the following steps:

[1] https://www.kaggle.com/datasets/patrickb1912/ipl-complete-dataset-20082020

5.2.1 Handling Missing and Incorrect Values:

In this phase, we examined the dataset for incomplete or incorrect data, a key obstacle in machine learning models. If data is fed without managing missing values or correcting the incorrect values, there is a substantial likelihood of erroneous output. Therefore, it is essential to handle missing values correctly. In our dataset, there was not a single cell that was empty. However there were few old team names which had to be updated with the new ones.

5.2.2 Data Engineering

Very limited number of variables were included in the ball by ball dataset. These were the number of overs, deliveries, runs per delivery, extras conceded, boundaries scored, sixes scored and wickets. These factors alone would not have been sufficient for us to make an accurate forecast. Therefore, the purpose of data engineering was to make use of this basic data in order to include characteristics that would assist the model in properly making predictions.

First, the age of all the players (as of 2020) has been gathered from a website known as espncricinfo, which is widely acknowledged to be the most reliable and genuine source for cricket data. Following that, age column has been inserted manually in the dataset. This has been done so that batsmen and bowers in the age range of 20 to 30 years could be screened out, who are the primary target for this project. Using aggregate functions and mathematical formulas, large number of batting and bowling features were derived from data accessible in the ball-by-ball dataset. Typically, these features are used to evaluate the performance of a player. Some of the most important batting attributes are batting average, batting strike rate, number of ducks, number of single digit scores,thirties, fifties, etc. while some important bowling attributes are overs bowled, wickets taken, bowling average, bowling strike rate, bowling economy rate and number of 3,4 or 5 wicket hauls.

Batting attributes likes batting averages, batting strike rate, boundaries scored and sixes hit have positive impact on the performance of a batsman. Higher the values, the better. On the other hand, attributes like ducks, single digit scores are bad aspect of the game since it shows the number of times the batsman has failed. The main batting attributes used in this research are as follows:

- **Batting Average:** The amount of runs scored over the course of an innings by the batsman.
- **Batting Strike Rate:** The average number of runs scored per 100 balls by the batsman.
- **Duck:** When a batsman is dismissed without having scored any runs.
- **Single Digit scores:** The number of times a batsman has scored less than ten runs in a game.
- **Centuries (100s):** A score of 100 or above by a batsman in an innings.
- **Fifties (50s):** A score of 50 or above by a batsman in an innings.
- **Thirties (30s):** A score of 30 or above by a batsman in an innings.

- **Boundaries Scored:** The number of times a batsman has scored four runs in one delivery throughout the game.
- **Sixes Hit:** The number of times a batsman has scored six runs in a single delivery.

The main bowling attributes used in this research are as follows:

- **Overs Bowled:** A group of six balls that are delivered to the batsman by the bowler.
- **Bowling Average:** The total number of runs that a bowler has conceded for each wicket that they have taken.
- **Bowling Strike Rate:** The amount of balls bowled per wicket, by the bowler.
- **Bowling Economy Rate:** Average runs conceded in an over by the bowler.
- **Three/Four/Five Wicket Haul:** Indicates if a bowler has claimed 3, 4, or 5 wickets in a single innings.

Besides prior performance, there are some additional aspects that influence performance of the players, such as venue and opposition. Passi and Pandey (2018) developed formulas for determining the stability and form of batsmen and bowlers. They also derived formulas to calculate the the performance of batsmen & bowlers against a specific opponent and in a specific venue. We have used the equations that were derived by them and with the help of python script we have determined the values of consistency, form, opposition & venue. These attributes were also considered for our analysis.

Consistency: This aspect of the player's game demonstrates how consistent the player has been in his entire career.

- **Batting Consistency** = 0.4262*average + 0.2566*no. of innings + 0.1510*SR + 0.0787*Centuries + 0.0556*Fifties – 0.0328*Ducks
- **Bowling Consistency** = 0.4174*no. of overs + 0.2634*no. of innings + 0.1602*SR + 0.0975*average + 0.0615*five wicket haul

Form: Form highlights the yearly performance of the player.

- **Batting form=** 0.4262*average + 0.2566*no. of innings + 0.1510*SR + 0.0787*Centuries + 0.0556*Fifties – 0.0328*Ducks
- **Bowling form** = 0.3269*no. of overs + 0.2846*no. of innings + 0.1877*SR + 0.1210*average + 0.0798*five wicket haul

Opposition/Opponents: Shows the performance of player against various teams.

- **Batting Opposition=** 0.4262*average + 0.2566*no. of innings + 0.1510*SR + 0.0787*Centuries + 0.0556*Fifties – 0.0328*Ducks
- **Bowling Opposition=** 0.3177*no. of overs + 0.3177*no. of innings + 0.1933*SR + 0.1465*average + 0.0943*five wicket haul.

Venue: This feature highlights the venue wise performance of a player.

- **Batting Venue=** 0.4262*Average + 0.2566*No. of innings + 0.1510*SR + 0.0787*Centuries + 0.0556*Fifties + 0.0328*Highest Score
- **Bowling Venue=** 0.3018*No. of overs + 0.2783*No. of innings + 0.1836*SR + 0.1391*Average + 0.0972*five wicket haul

5.3 Data Binning

The values of batting and bowling features lie within fairly large ranges, and tiny deviations in these values do not differentiate one player from another in the game of cricket. Along with the target variables runs scored and wickets taken, the values of averages, strike rates, consistency, form, opponent-wise performances and venue-wise performances were allocated to separate bins based on the value. While the experiment was being carried out, it was discovered that binning contributes to huge improvement in the accuracy of the models.As a result, we assigned a value between 1 and 5 to each batting features, with 1 being the least important and 5 being the most important, depending on the range over which its value falls.

Bin	Consistency	Form	Opponent	Venue
1	0-20	0-20	0-20	0-20
2	21-40	21-40	21-40	21-40
3	41-50	41-50	41-50	41-50
4	51-60	51-60	51-60	51-60
5	61 and above	61 and above	61 and above	61 and above

Table 1: Derived Batting Attribute Bins

Batting Average				
Bin	Overall	Yearly	Opposition Wise	Venue Wise
1	0-10	0-10	0-10	0-10
2	11-20	11-20	11-20	11-20
3	21-30	21-30	21-30	21-30
4	31-40	31-40	31-40	31-40
5	41 and above	41 and above	41 and above	41 and above

Table 2: Batting Average Bins

Batting Strike Rate				
Bin	Overall	Yearly	Opposition Wise	Venue Wise
1	0-25	0-25	0-25	0-25
2	26-75	26-75	26-75	26-75
3	76-125	76-125	76-125	76-125
4	126-150	126-150	126-150	126-150
5	151 and above	151 and above	151 and above	151 and above

Table 3: Batting Strike Rate Bins

In a similar manner, in order to categorize bowling properties, we gave each of them a value that ranged from 1 to 5, with 1 being the least significant and 5 being the most significant. Because of this, the model's accuracy in predicting the number of wickets increased drastically.

Bin	Consistency	Form	Opponent	Venue
1	1-20	1-10	1-10	1-10
2	21-40	11-20	11-20	11-20
3	41-60	21-30	21-30	21-30
4	61-80	31-40	31-40	31-40
5	81 and above	41 and above	41 and above	41 and above

Table 4: Derived Bowling Attribute Bins

Bowling Average & Strike Rate				
Bin	Overall	Yearly	Opposition Wise	Venue Wise
5	0-10	0-10	0-10	0-10
4	11-20	11-20	11-20	11-20
3	21-30	21-30	21-30	21-30
2	31-40	31-40	31-40	31-40
1	41 and above	41 and above	41 and above	41 and above

Table 5: Bowling Average & Strike Rate Bins

Both predictions were classified as classification problems. Hence runs scored and wickets taken were both divided into 5 different bins.

Target variables		
Bin	Runs Scored	Wickets Taken
1	0-15	0-1
2	16-30	2
3	31-50	3
4	51-70	4
5	71 and above	5 and above

Table 6: Target Variable Bins

5.4 Data Encoding

Models of machine learning often include mathematical computation; thus it is essential that the values are presented in numerical form. After the data were partitioned into their respective bins, we inspected the data types each column. Data that was originally presented in a format other than numerical were converted into numerical values. This was done in order to reduce the likelihood of receiving inaccurate outcomes.

5.5 Data Mining

In part 2, a wide range of research articles were examined in order to create machine learning models for forecasting a cricket player's future performance. Random forest and Naive Bayes were identified by the researchers as the ones that have shown the greatest track record of successfully predicting future performance. Cross validations have been performed on all of the models so that we can be certain that the model is accurate.

5.5.1 Random Forest

Random forest is a machine learning method that is versatile and simple to implement, and it generates an excellent result the majority of the time even when the hyperparameters are not tuned. In addition to this, it is one of the most frequently used algorithms due to the ease with which it can be implemented as well as the versatility with which it can tackle classification and regression problems. Random Forest generates many decision trees and then combines them in order to get a more precise and consistent forecast, which helps to ensure that the resulting models are not over fitting. A important step in the process of performance prediction is the separation of features into those that are relevant and those that are not relevant. One of the most significant benefits of using random forests is the fact that it can determine the relevance score of each feature, therefore learning the influence that each feature has on the prediction of the classes Uddin and Uddiny (2015).

5.5.2 Naive Bayes

Bayesian classifiers are a kind of statistical classifier that determines the likelihood of a given tuple belonging to a certain category based on its characteristics. This classifier operates on the presumption that each feature has its own unique impact on the class label, irrespective of the values of the features that are used to describe the other classes. This whole process is referred to as class-conditional independence. Jarecki et al. (2013) explored the significance of the assumption of class conditional independence of object attributes on human classification learning. This classification method is based on Bayes Theorem which was named after Thomas Bayes.

$$\textbf{Bayes Theorem: } P(X \mid Y) = \frac{P(X) * P(Y \mid X)}{P(Y)}$$

6 Evaluation

The experiment was performed with a variety of training and test set sizes in order to identify the one that produced the highest level of accuracy. Two different machine learning algorithms were used in the experiment: Random Forest and Naive Bayes. In order to assess the models, metrics like Accuracy, Precision F1 score and recall were used.

6.1 Case Study 1: Predicting Run Scored

During the process of predicting runs scored from batting dataset, it was found that the accuracy of both Random Forest and Naive Bayes declined as the size of the training dataset was increased. When the models were trained on 70% of the dataset, Random

Forest achieved a prediction accuracy of 92.2%, whereas Naive Bayes was able to obtain an accuracy of prediction of 84%. In a similar manner, when the models were trained on 80% of the dataset, Random Forest was able to reach a prediction accuracy of 91.7%, whilst Naive Bayes was only able to acquire a prediction accuracy of 83.1%. These results have been highlighted in a chart 6 below:

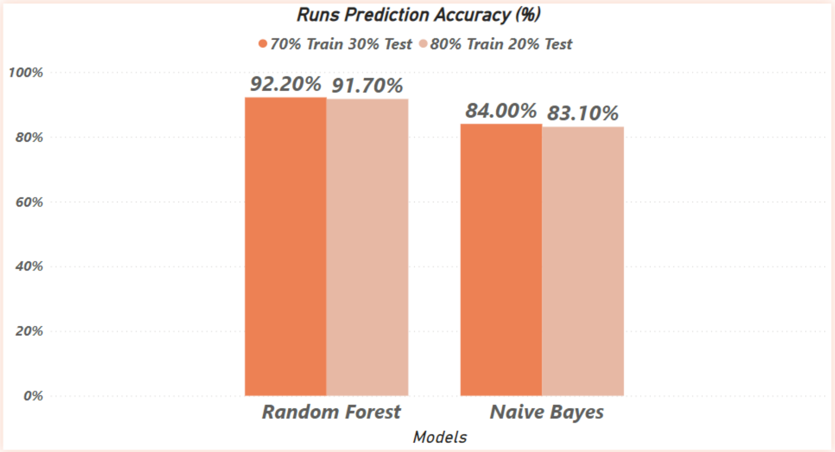
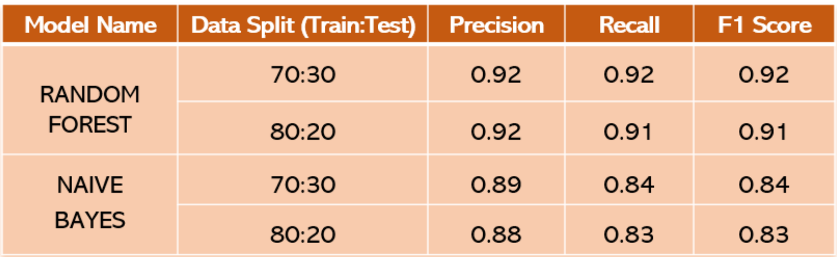


Figure 6: Model Wise Runs Prediction Accuracy

When trained on 70% of the dataset, Random Forest had the highest performance in terms of all other parameters while forecasting runs. It had a precision score of 0.92, a recall score of 0.92 and an F1 Score of 0.92. These are all outstanding numbers for a classifier. Naive Bayes, on the other hand, obtained a precision score of 0.89, recall and F1 scores of 0.84 when trained on 70% of the data. When tested on 80% of the dataset, both models produced lower results. This is shown clearly in the following figure 7.



Model Name	Data Split (Train:Test)	Precision	Recall	F1 Score
RANDOM FOREST	70:30	0.92	0.92	0.92
	80:20	0.92	0.91	0.91
NAIVE BAYES	70:30	0.89	0.84	0.84
	80:20	0.88	0.83	0.83

Figure 7: Assessment Table of Runs Prediction

Anik et al. (2018) were able to attain an accuracy of 91.5% when forecasting runs. However, their research was only conducted with a small number of players and the number of factors considered for the analysis of performance were also minimal. The method that was used in this study is effective for every player in the globe and it is also relevant to new players who will be playing for their respective club or countries in the near future.

6.2 Case Study 2: Predicting Wickets Taken

While estimating wickets based on bowling data, it was discovered that increasing the amount of the training dataset led to an improvement in the accuracy of the Random Forest model, while Naive Bayes' performance became worse as the quantity of the training sample got bigger. When the Random Forest was trained on just 70% of the data, it attained an accuracy of 82.33%; however, when it was trained on a bigger dataset consisting of 80%, it boosted its accuracy to 84.26%. On the other hand, when Naive Bayes was trained on only 70% of the data, it achieved an accuracy of 69.1%, but when it was trained on 80% of the data, the accuracy dropped to 68.3%. A chart 8 has been used to draw attention to these findings.

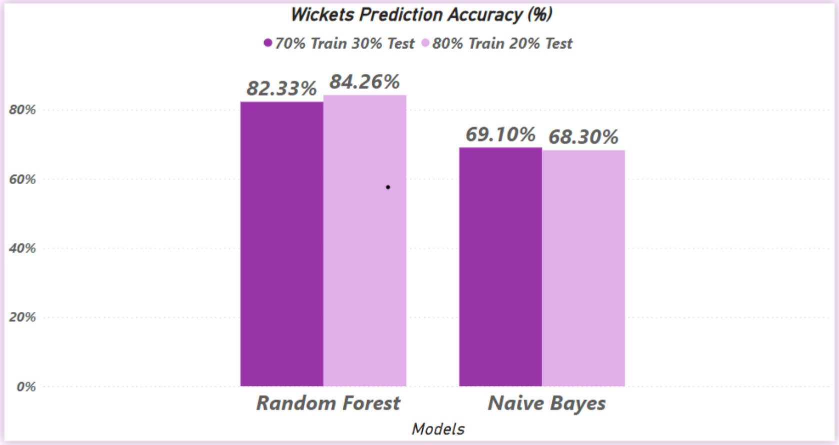
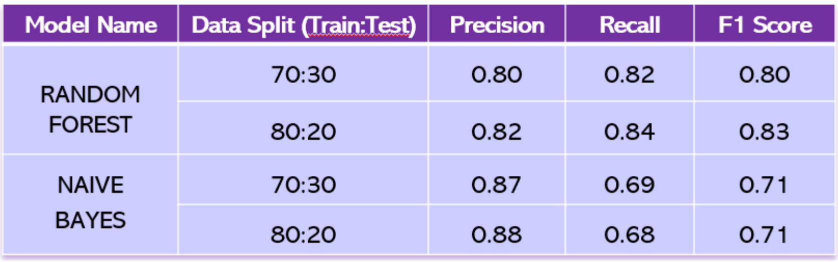


Figure 8: Model Wise Wickets Prediction Accuracy

When it came to forecasting wickets, Random Forest once again performed more accurately than Naive Bayes. When trained on 80% of the dataset, it got a precision score of 0.82, a recall score of 0.84, and an F1 score of 0.83. Naive Bayes algorithm achieved an precision score, recall score and F1 score of 0.88, 0.68 and 0.71 respectively. The figure 9 below clearly demonstrates this.



Model Name	Data Split (Train:Test)	Precision	Recall	F1 Score
RANDOM FOREST	70:30	0.80	0.82	0.80
	80:20	0.82	0.84	0.83
NAIVE BAYES	70:30	0.87	0.69	0.71
	80:20	0.88	0.68	0.71

Figure 9: Assessment Table of Wickets Taken

6.3 Discussion

The difficulties that were experienced while doing the study has been discussed in this part. It has also been discussed whether or not the objectives were fulfilled.

With the use of knowledge gathered from past researchers and existing understanding of cricket, this study was carried out. Feature engineering was a fundamental need to be able to create an accurate prediction, since the information that was already there in the initial dataset was not adequate on its own. Even player ages weren't included in the data. In order to select out batsmen and bowers between the ages of 20 and 30, who are the project's main focus, age of all the players were manually inserted into the dataset. Random forest classifier and naive bayes classifier were utilized for this study since both the problems were classification-based. Following the implementation and analysis of both models, it was found that the Random Forest Classifier was the one that was the most accurate in terms of forecasting both the number of runs scored and the number of wickets taken.

7 Conclusion and Future Work

Mega auction, where players are bought, is of tremendous assistance to teams that want to rebuild their squads in high profile tournaments such as the Indian Premier League. Indian Premier League has its mega auction once every three years. The retention criteria for this tournament allows teams to keep up to a maximum of four players out of a total pool of twenty-five players. As a result, recognizing and investing in high performing players between the ages of 20 and 30 is essential for the team management each time there is an auction. The study that has been implemented will be of assistance to the owners and management of the clubs in making the appropriate player selection. Results of this research would be most useful to scouts working for domestic Twenty20 teams, if they were used to assist in the process of identifying players who have the potential to be valuable additions to the teams. Moreover, this will assist gambling firms in increasing their overall revenue. In this research, the batting and bowling datasets were analyzed using the past statistics and attributes of the players.Two distinct classification techniques were made use of and assessed. In the end, Random Forest classifier proved to be the most accurate classifier for predicting runs and wickets. A success rate of 92.2% and 84.26% was achieved while predicting the number of runs scored and the number of wickets taken respectively.

This technique of analysis that's been suggested in this study may also be adapted to be used by the international or domestic team in any other format, however it will need some alterations to be suitable for other formats. Additional research can be conducted in future to extract other bowler related attributes from open source websites that might assist in making more accurate predictions on bowlers wicket taking ability, hence enhancing the accuracy overall.

7.1 Acknowledgement

I would want to use this opportunity to express my gratitude to my supervisor Mr. Aaloka Anant for his guidance throughout the duration of this research. His recommendations and continuous support were really helpful to me. I would also like to thank my parents for all the support and encouragement they gave me during this research.

References

Anik, A. I., Yeaser, S., Hossain, A. I. and 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)*, IEEE, pp. 500–505.

Brown, T. W., Gough, L. A. and Kelly, A. L. (2022). Performance trajectories of bowlers and batters from youth level to senior professional status in cricket, *International Journal of Performance Analysis in Sport* **22**(1): 1–15.

Bunker, R. P. and Thabtah, F. (2019). A machine learning framework for sport result prediction, *Applied computing and informatics* **15**(1): 27–33.

Cornman, A., Spellman, G. and Wright, D. (2017). Machine learning for professional tennis match prediction and betting, *Stanford Unverisity* .

Danisik, N., Lacko, P. and Farkas, M. (2018). Football match prediction using players attributes, *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, IEEE, pp. 201–206.

D'Oca, S., Corgnati, S. and Hong, T. (2015). Data mining of occupant behavior in office buildings, *Energy Procedia* **78**: 585–590.

Iyer, S. R. and Sharda, R. (2009). Prediction of athletes performance using neural networks: An application in cricket team selection, *Expert Systems with Applications* **36**(3): 5510–5522.

Jarecki, J., Meder, B. and Nelson, J. D. (2013). The assumption of class-conditional independence in category learning, *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 35.

Jayalath, K. P. (2018). A machine learning approach to analyze odi cricket predictors, *Journal of Sports Analytics* **4**(1): 73–84.

Jayanth, S. B., Anthony, A., Abhilasha, G., Shaik, N. and Srinivasa, G. (2018). A team recommendation system and outcome prediction for the game of cricket, *Journal of Sports Analytics* **4**(4): 263–273.

Kamble, R. et al. (2021). Cricket score prediction using machine learning, *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* **12**(1S): 23–28.

Metelski, A. (2021). Factors affecting the value of football players in the transfer market, *Journal of Physical Education and Sport* **21**: 1150–1155.

Oytun, M., Tinazci, C., Sekeroglu, B., Acikada, C. and Yavuz, H. U. (2020). Performance prediction and evaluation in female handball players using machine learning models, *IEEE Access* **8**: 116321–116335.

Passi, K. and Pandey, N. (2018). Increased prediction accuracy in the game of cricket using machine learning, *arXiv preprint arXiv:1804.04226* .

Uddin, M. T. and Uddiny, M. A. (2015). A guided random forest based feature selection approach for activity recognition, *2015 international conference on electrical engineering and information communication technology (ICEEICT)*, IEEE, pp. 1–6.

Vistro, D. M., Rasheed, F. and David, L. G. (2019). The cricket winner prediction with application of machine learning and data analytics, *International Journal of Scientific & Technology Research* **8**(09).

Yang, F.-J. (2018). An implementation of naive bayes classifier, *2018 International conference on computational science and computational intelligence (CSCI)*, IEEE, pp. 301–306.