BEST 11 SELECTION USING MACHINE LEARNING

Inspiring Excellence

SUBMISSION DATE: 05.04.18

SUBMITTED BY:
Aminul Islam Anik(14301039)
Sakif Yeaser(14301045)
A.G.M. Imam Hossain(14301046)
Department of Computer Science and Engineering

Supervisor:

Amitabha Chakrabarty, Ph.D
Assistant Professor
Department of Computer Science and Engineering

Declaration

We, hereby declare that this thesis is based on results we have found ourselves. Materials of work from researchers conducted by others are mentioned in references.

Signature of Supervisor

Amitabha Chakrabarty, Ph.D
Assistant Professor
Department of Computer Science and Engineering
BRAC University

Signature of Authors

Aminul Islam Anik
(14301039)

Sakif Yeaser
(14301045)

A.G.M. Imam Hossain
(14301046)

ABSTRACT

Cricket has become the most popular game not only in Bangladesh but also around the world. Day by day it is gaining more people's attention. The name and glory of team Bangladesh has trespassed the country's border and it is creating a huge impact in the world cricket. To make sure the continuous development of the team, a more professional approach and help from IT sector is needed. Remembering this fact, we used machine learning to select the best players from the standby list based on the previous playing statistics and then from that players we have found out the winning team combination. We have collected the data from websites that offers trustable sports statistics. Feature selection algorithms like Recursive feature elimination and univariate selection are used to find out the attributes that are more related to the output feature. Machine learning Algorithms such as linear regression, support vector machine with linear and polynomial kernel was used to predict the runs scored by a batsman and runs given by a bowler. Later on, we have also used fully connected neural network to find out the performance comparison of different algorithm. We have selected the players for the team according to their performances and experiences. Our goal was to form a well-balanced team through our approach.

ACKNOWLEDGEMENT

We are deeply thankful to BRAC University for providing us such a wonderful environment to peruse our research. We would also like to express our utmost gratitude and appreciation to our supervisor Amitabha Chakrabarty for his attention and time. We would also like to thank him for giving us the opportunity to work on this topic and assisting us throughout the process. We found the websites helpful and rich in data such as: ESPNcricinfo.com and Howstat.com from where we collected our data for our research work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF TABLES

CHAPTER 1

Introduction

As machine learning aims to focus on larger and complex tasks, the importance of providing a potential amount of relevant data has become the most crucial part of the field. In supervised machine learning for the development and to run a relevant algorithm it is important that the data records should be organized.

We have collected our data for the purpose of our research work from some credible web-based documents such as, howstat.com [7] and ESPNcricinfo.com [8] individually for each player for the selection thrust of best eleven players from a given pool of players. We have used decision tree algorithm to compare players based on their past records against a particular opposition, in particular venue and current form and built a model to predict the best possible performance of each selected players. Finally, we used the result of our experiment to propose the best team formation.

1.1 Objective

The main objective of this work is to find out the best possible team combination from the standby list. At the same time, we intend to predict the run score of each batsman and the run each bowler will give away in the next match. We used different supervised machine learning technique and compared the performances of the different model.

1.2 Motivation

In our country, cricket is now the most popular game. People from all ages and occupations are fans of cricket because the Bangladeshi team has been doing well recently. But

sometimes we see that many confusions arise before a match about team combination, example: which player to pick or drop, which batsman should play in which position and so on. From here we have got the motivation of our work to make the team combination automated using machine learning.

1.3 Methodology

The overview of the methodology of the work is given below in Figure 1:

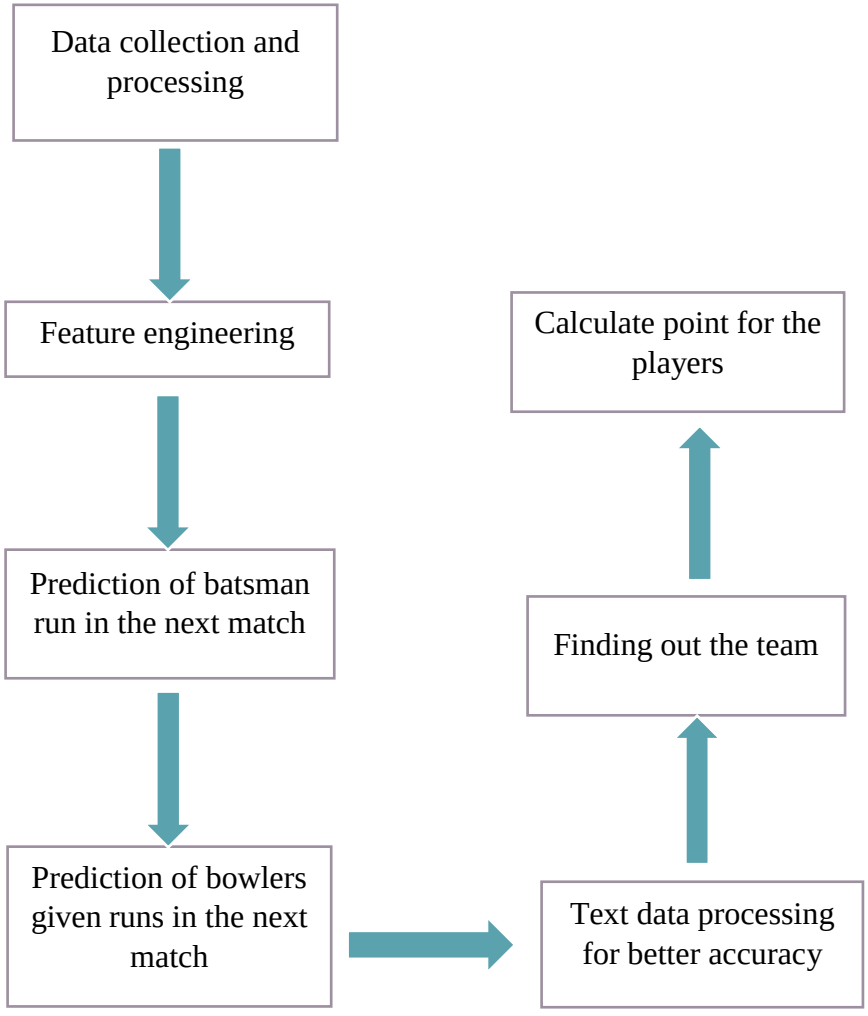

Figure 1: Methodology of research

CHAPTER 2

Literature Review

Kusiak, Kern, K. H. Kernstine and B. T. Tseng, suggested, in order to select the best combination of players for a team, it is necessary to develop a better selection tool. In their research, they have applied K-Nearest neighbor to find a score which is closer to the standard score or threshold of a player. These threshold scores and performance are recorded for of all players [1]. Mickey Arthur has summed up the power to participate or responsible in a game victory by addressing some influencing factors of ODI cricket such as physical strength, scoring, home ground advantages, day/night situation, toss and decision of bating first or second. How these factors can affect the performance of players in a cricket match is determined by understanding and estimation of the values used in Bayesian classifiers [2]. R. L. Holder and A. M. Nevill suggested to use Logistics and normal regression models to find the ranking and home ground advantages for players as according to his research he has found that players perform well when they play in their home grounds. Moreover, in order to determine the weak points of opposite teams, Reverse data mining technique is used, which can plan defensive strategy for the next game they play [3]. In terms of average score, strike rate and consistent performance, the form and the selection of players are defined. To classify a batsman, Lemmer has given emphasize on the consistency curve and in which he has mentioned that the higher accuracy of the predicted output would be achieved if the consistency co-efficient is higher. On the other hand, with a lower consistency co-efficient it is more likely to get a low accurate predicted

result. Here, author has used three subsets of data and has given weights to each by the consistency level with the predicted feature. Then he has made a combined set of these weighted subsets, which he used to predict performance of each batsmen. Thus, he compared performance between players. After that, a classification algorithm is used here which assigns classes to the batsmen based on their predicted performance where higher assessed batsmen would fall in higher classes and others will be sorted accordingly [4]. Machine learning technique is also used in tennis game to predict the match outcome. Using historical data, they extract 22 features and applied two supervised machine learning algorithms (logistic regression and artificial neural networks). Evaluating over a test set of 6315 ATP matches (played during 2013-14) they found the result which outperform Knottenbelt's Common-Opponent model and the current state-of-the-art in stochastic modelling [5]. Renato Amorim Torres, used machine learning algorithms (linear regression and Maximum Likelihood Classifier) are to predict the match outcome and both of the algorithm achieved a result which is much better than the conventional approach. Here previous match records (2006-2012) have used as data source and feature vectors were selected through some graphical chart or from some simple prediction [6].

CHAPTER 3

Data description and Methodology

This chapter contains the data description and its graphical analysis. This also contains the methodology used for extracting and selecting the important features.

3.1 Data and Analysis of data

In every machine learning based work data is the most valuable part and can be treated as the heart of the whole process. So, collecting relevant data from a trustable source is the first and foremost duty. We have collected our data from the howstate.com [7] & espncrickinfo.com [8] which are the most recognized and authentic source of the cricket statistical data.

3.1.1 Data collection

After getting the data source the first challenge is to collect the data. For our work we have used the Microsoft excel worksheet as a platform to keep our data. We directly imported our intended data from the website using the Microsoft excel. For that we opened a new excel and then went to “DATA” option in the main menu. Inside that there is an option called “From web”. After clicking that a window will be opened and there is an option to write the web site address from where we want to import any data we want.

3.1.2 Data representation

To work with the data, we need to represent data in a way so that we can do our further experiment. As we are using Python to implement all the algorithms we have used the csv format in the Excel which is easy to work with and compatible with python coding. As our

initial experiment we have collected the game statistics of the Bangladeshi team captain Shakib Al Hasan from all the 180 matches he played.

To represent the data, we have selected some features like: Runs, Minutes played, Ball faced, 4's, 6's, Strike Rate, playing position & opposition. Later on, we have also collected the data from various angles which involves performances of Shakib Al Hasan in different ground, against different opponents, by years. In Figure 2, 3, 4, 6, 7- we have plotted the collected data in the graph where the below mentioned abbreviations are used:

- "Mat"- Matches played against opponent
- "HS"- Highest Score
- "S/R"- Strike Rate
- "100s"- Number of 100's
- "50s" - Number of 50's
- "Runs"- Number of runs scored by the player
- "Avg"- Average runs scored
- "Inns"- Number of innings played per year
- "B/F"- Ball faced per matches in the year of 2017

3.1.3 Shakib Al Hasan Performance Analysis

To understand the nature of the data and do some initial visualization our work was concentrated with the collected data and some visual analysis conducted on the dataset of Shakib Al Hasan to make various graphs. From those graphs his performances were analyzed based on opponent, performances by years, batting records in different grounds, etc.

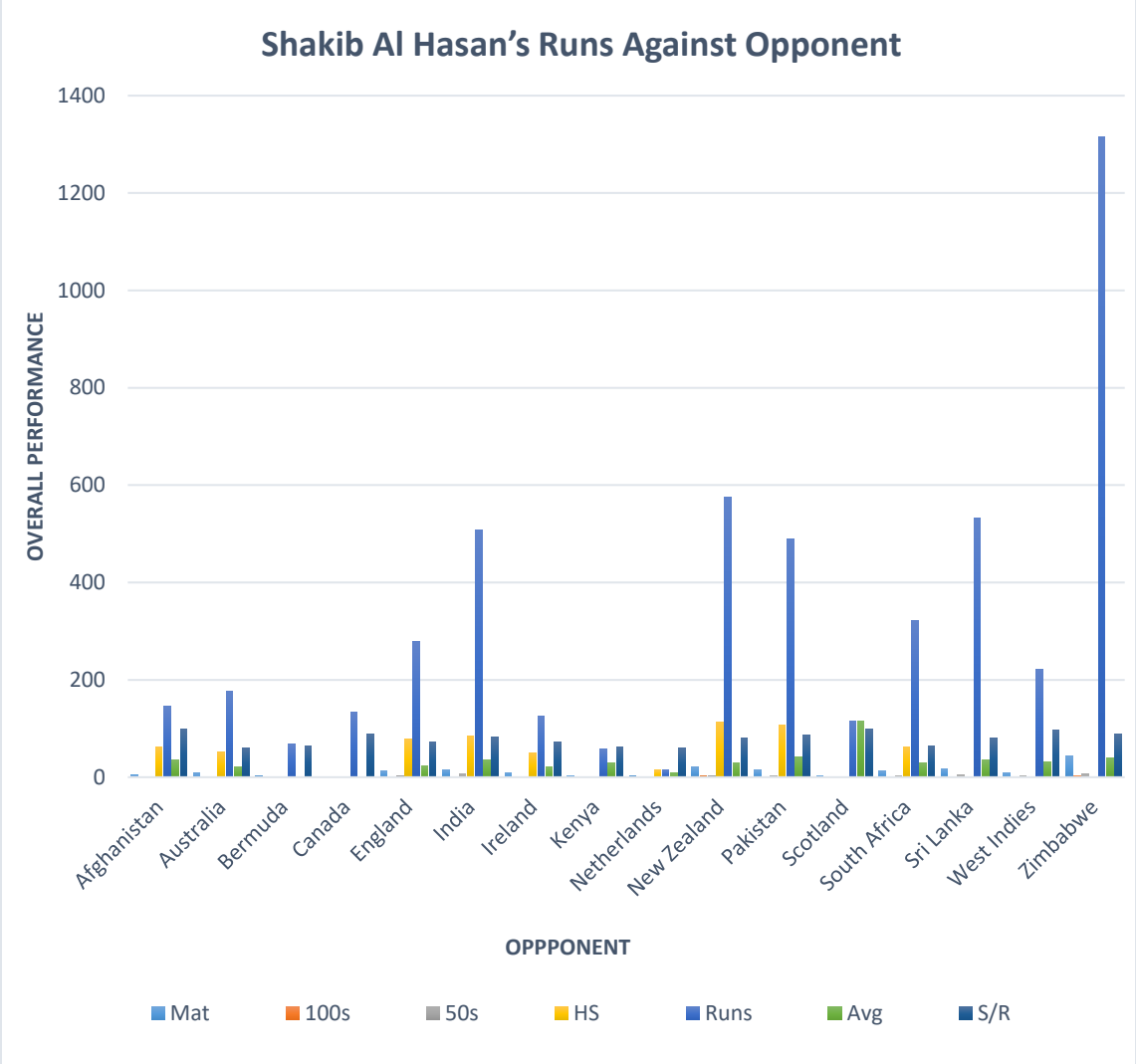

Figure 2: Shakib Al Hasan's runs against opponents

Figure 2 represents Shakib's performance against different opponents in his career. From the graph we can clearly see that he scored well against Zimbabwe, New Zealand, Sri Lanka & India but his record against Australia is not up to the mark compared to another opponent. In Figure 3, we have shown the performances of Shakib Al Hasan between the years of 2006 to 2017.

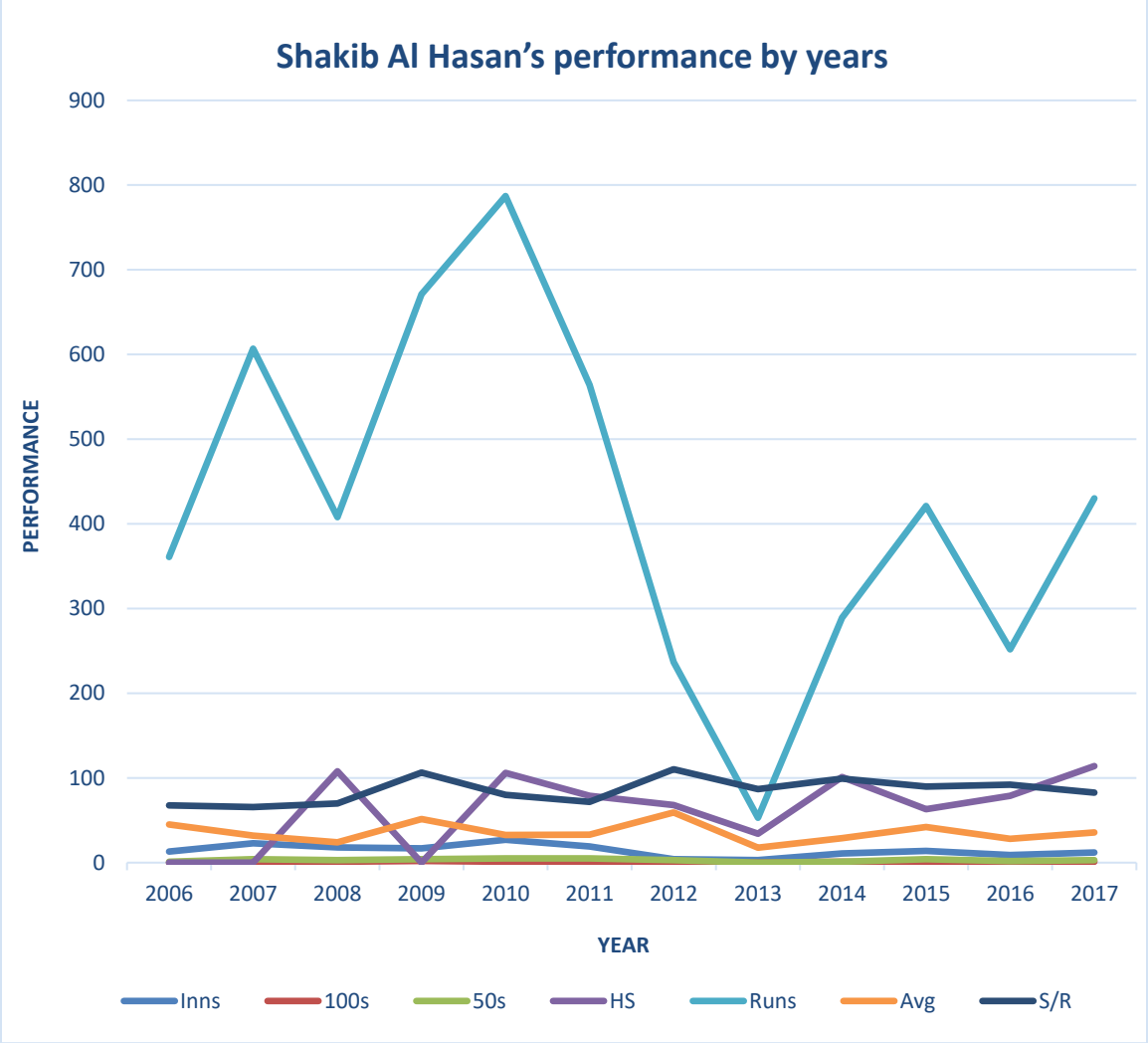

Figure 3: Shakib Al Hasan's performance by years

From the graph we can visualize that he had the most promising time in 2010 and Shakib's performance (runs) fall greatly in 2013. In the year of 2017, Shakib Al Hasan's performance's slope is seeing a positive turn after the sharp fall in 2016. Now let's zoom in a bit and see the performance breakdown of each match in 2017 which is illustrated in Figure 4:

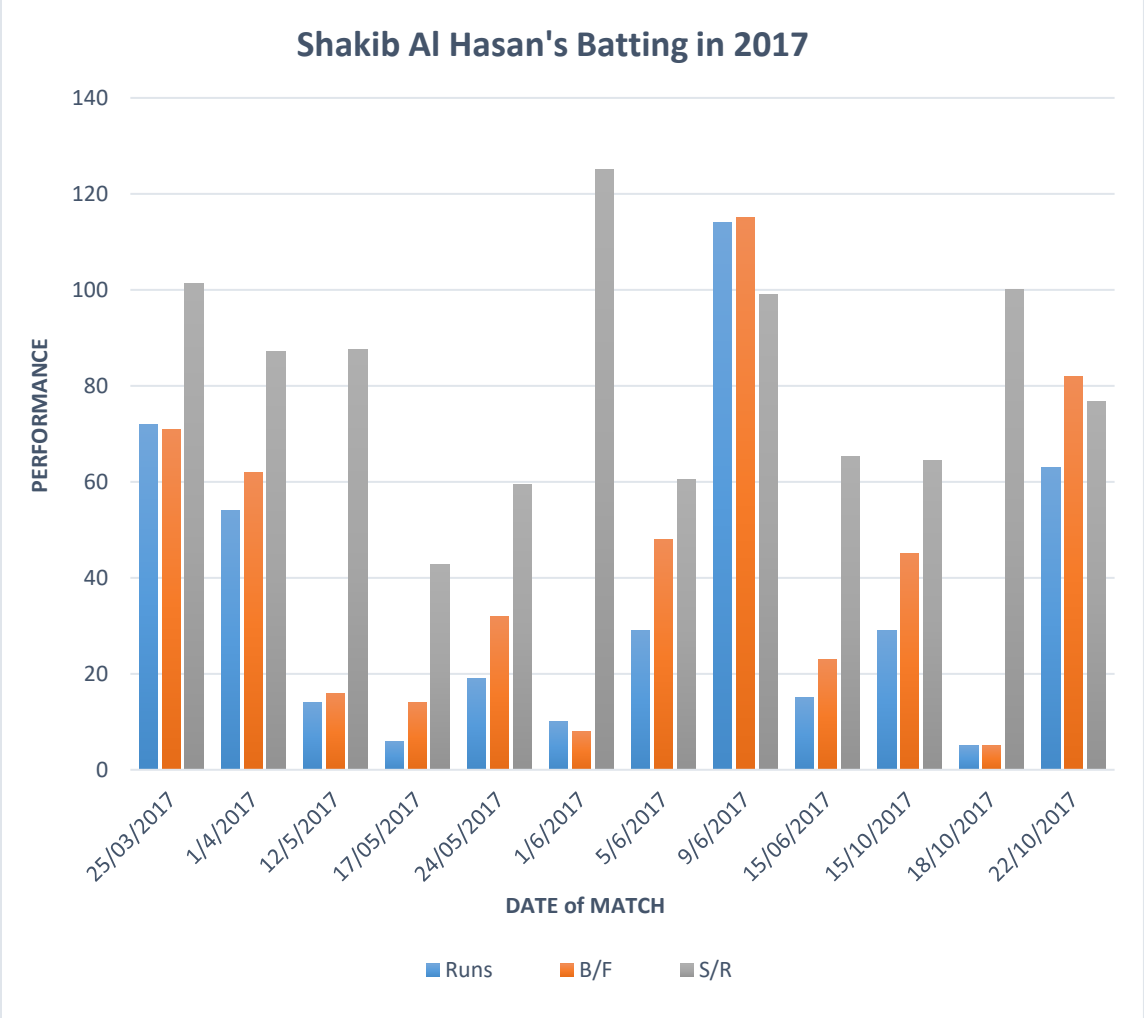

Figure 4: Shakib Al Hasan's batting in 2017

Lastly, ground is one of the main external concern that has a high impact on a player's performance. In Figure 5, we visualized Shakib's performance in different stadium. He scored most runs played most number of matches in the Sher-e-Bangla National Stadium which is the home ground for the Bangladeshi team and the average is very decent here. In the overseas we can see some peaks in Antigua Recreation Ground, Civil Service Cricket Club and Windsor Park.

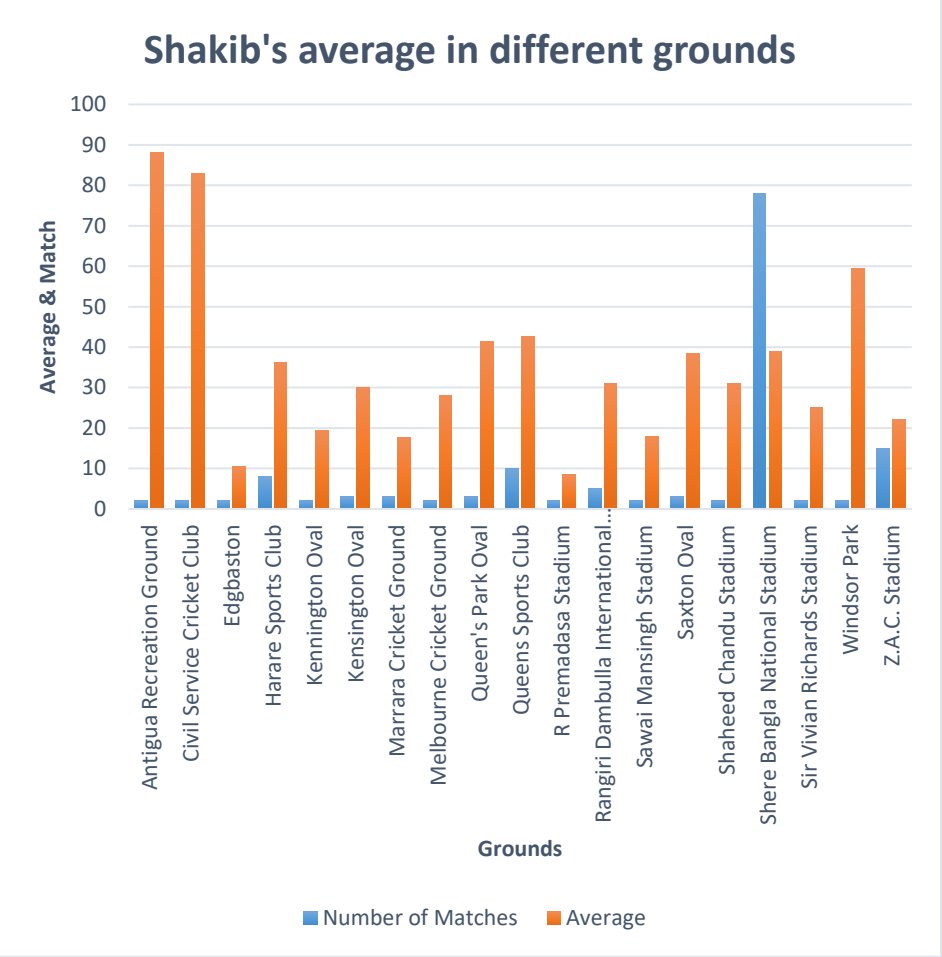

Figure 5: Shakib Al Hasan's average in different grounds

3.1.4 Player Comparison

The visualization in Figure 6, compares performances between Shakib and Tamim. From the graph it can be clearly noted that Tamim Iqbal is well ahead in every. The graph from Figure 6, shows a comparison between Tamim and Shakib's batting performance against different teams. Although against Zimbabwe and New Zealand, Shakib al Hasan is better than Tamim Iqbal but against other teams Tamim has better records of scoring runs. Figure 7, shows the comparison by years.

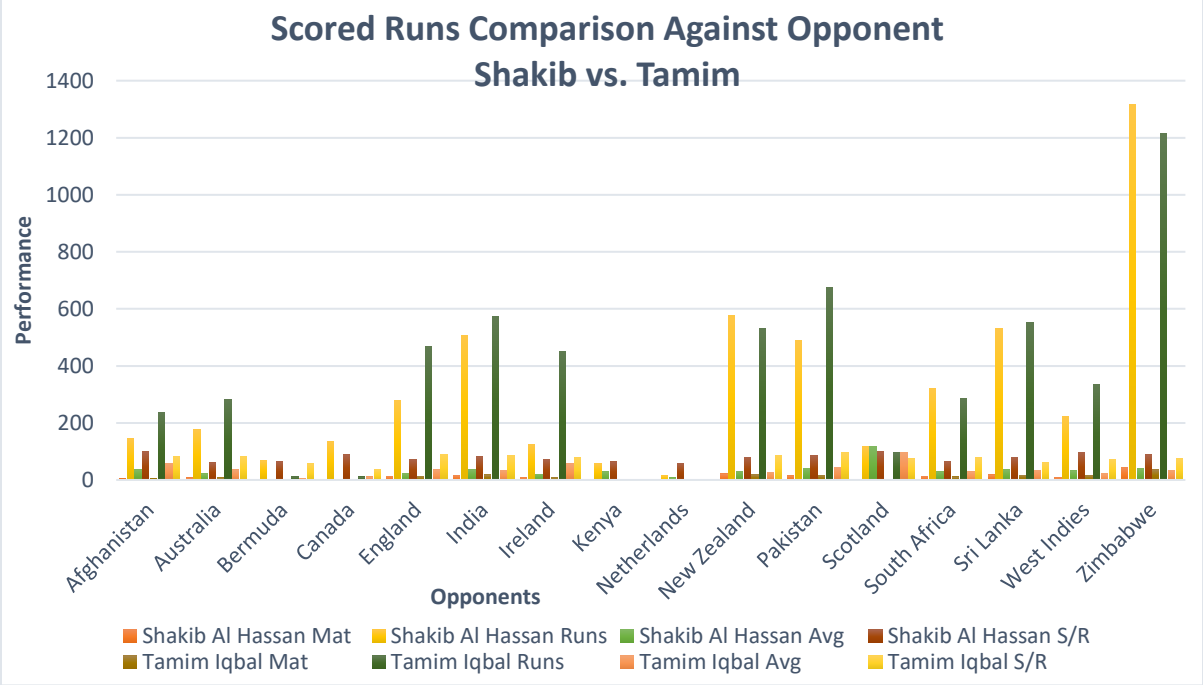

Figure 6: Shakib vs. Tamim comparison against opponent

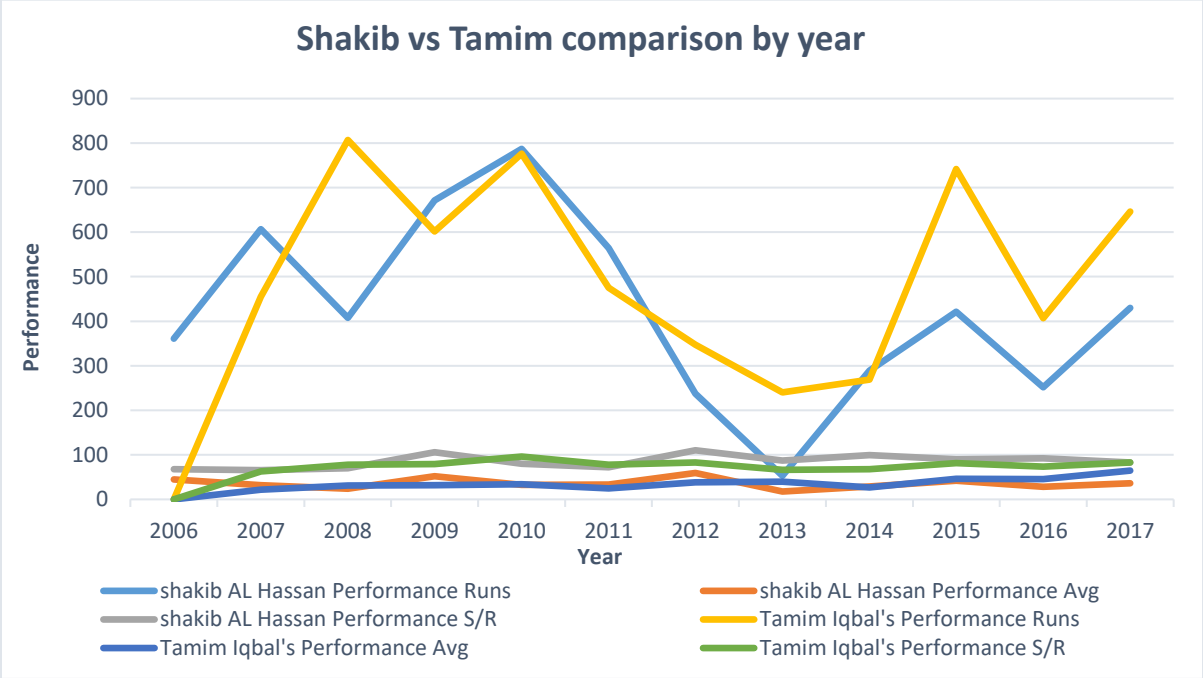

Figure 7: Shakib vs. Tamim comparison by years

The graph from Figure 7 shows, 2010 was almost an equal year for both of them but after that Tamim Iqbal recently is over passing Shakib Al Hasan at a big margin. The last comparison graph illustrates the performance comparison in different venue. Figure 8 illustrate this graph-

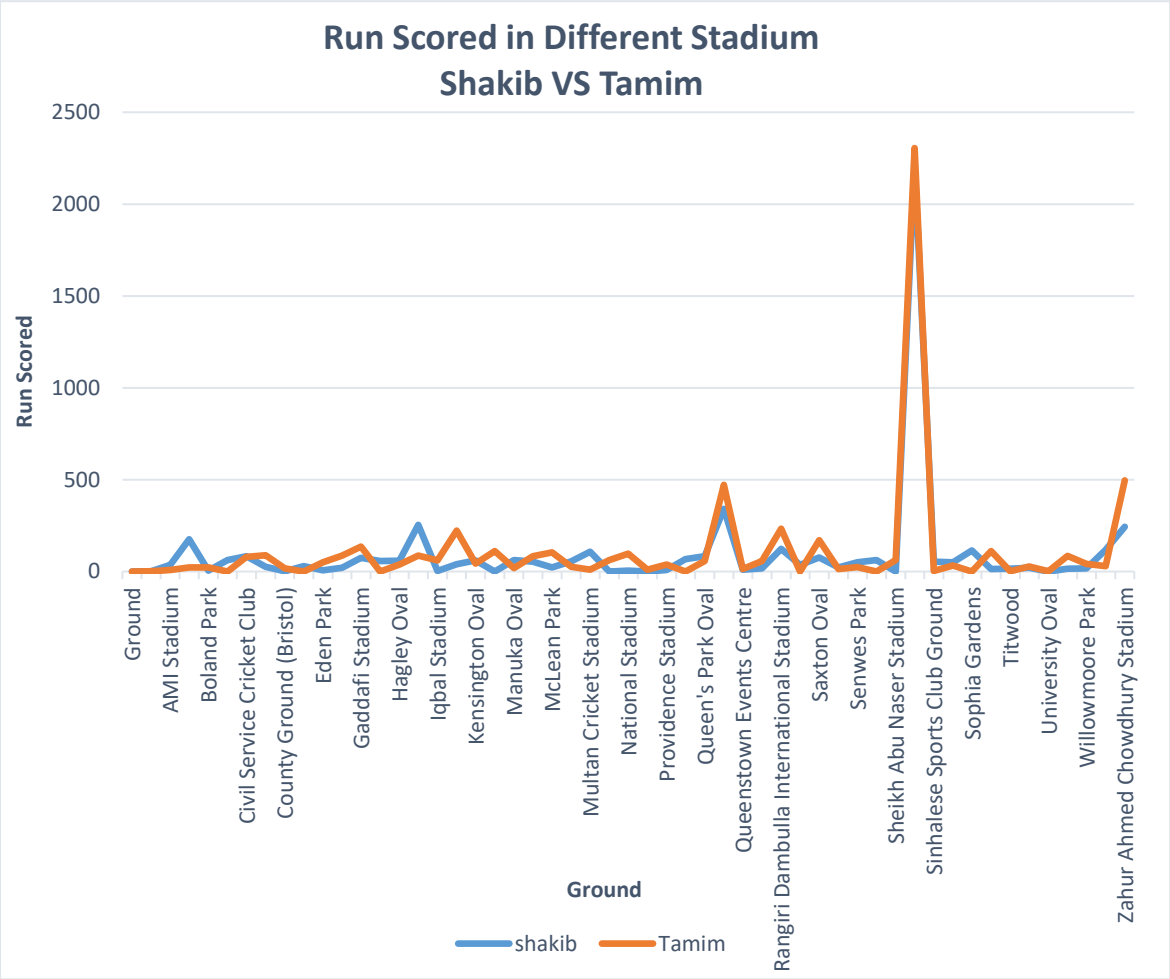

Figure 8: Shakib VS Tamim in Different Stadium

From the graph of Figure 8, we can say that Tamim Iqbal almost dominate Shakib al Hassan in all the stadiums except Hagley Oval. They have scored almost equal runs in the stadium which are situated in home.

3.2 Data Processing

3.2.1 Batsman and Bowler dataset processing

Figure 9 represents dataset which is in csv format. This is the screen-shot of a part of the Tamim Iqbal's batting dataset which was collected from the web sites.

Runs	Mins	BF	4s	6s	SR	Pos	Dismissal	Inns	Opposition	Ground	CONDITION
5	10	8	0	0	62.5	1	caught	1	v Zimbabwe	Harare	slow,bouncy
30	40	32	3	2	93.75	1	caught	2	v Zimbabwe	Harare	slow,bouncy
11	0	19	1	0	57.89	1	caught	2	v Bermuda	St John's	fast,bouncy
11	0	29	2	0	37.93	1	run out	1	v Canada	St John's	fast,bouncy
51	55	53	7	2	96.22	1	caught	2	v India	Port of Spain	swinging,spinny
6	16	7	1	0	85.71	1	caught	2	v Sri Lanka	Port of Spain	swinging,spinny
1	0	2	0	0	50	1	caught	2	v Bermuda	Port of Spain	swinging,spinny
3	11	14	0	0	21.42	1	caught	1	v Australia	North Sound	fast,bouncy
29	69	54	4	0	53.7	2	stumped	1	v New Zealand	North Sound	fast,bouncy
38	90	59	6	0	64.4	2	caught	1	v South Africa	Providence	flat,spinny
8	5	7	2	0	114.28	1	caught	1	v England	Bridgetown	flat,spinny
29	96	59	4	0	49.15	1	bowled	2	v Ireland	Bridgetown	flat,spinny
7	27	20	1	0	35	2	run out	2	v West Indies	Bridgetown	flat,spinny
45	81	53	6	0	84.9	2	caught	1	v India	Dhaka	slow,spinny

Figure 9: Tamim Iqbal's initial dataset

We have taken records of all matches that a batsman has played in his career. Here ball faced means number of balls he played, position is the batting position, opposite team is represented as numerical value as the supervised learning algorithm works with numerical value it cannot process text data. So, Teams of same strength are replaced by same number and numbers are assigned according to ICC team ranking. The list of the teams with their assigned number is given below in Figure 10:

```
teams = {
    'v Zimbabwe': 6,
    'v Kenya': 7,
    'v Sri Lanka': 5,
    'v West Indies': 4,
    'v Scotland': 7,
    'v Bermuda': 7,
    'v Canada': 7,
    'v India': 1,
    'v Australia': 1,
    'v New Zealand': 4,
    'v South Africa': 1,
    'v England': 3,
    'v Ireland': 6,
    'v Pakistan': 4,
    'v Afghanistan': 6,
    'v U.A.E.': 7,
    'v Netherlands': 7
}
```

Figure 10: Opponent team representing numerical value

Ground are treated as home or away and pitch are classified in 9 categories like slow and spiny or grassy and bouncy etc. We have collected all the pitch report of different country's venue and classified them into their nature. Then we have assigned number to represent them in our training dataset.

BF	Pos	opposite	ground	pitch	4s	6s	SR
8	1	6	1	8	0	0	62.50
32	1	6	1	8	3	2	93.75
19	1	7	1	4	1	0	57.89
29	1	7	1	4	2	0	37.93
53	1	1	1	3	7	2	96.22
7	1	5	1	3	1	0	85.71
2	1	7	1	3	0	0	50.00
14	1	1	1	4	0	0	21.42
54	2	4	1	4	4	0	53.70
59	2	1	1	9	6	0	64.40
7	1	3	1	9	2	0	114.28
59	1	6	1	9	4	0	49.15

Figure 11: A part of the Batsman dataset after initial processing

After that we have processed the dataset to make it suitable for implementing machine learning algorithm. Figure 11 and 12 represent the processed data which was used in the algorithm for prediction.

overs	opposite	ground	pitch	Econ	Wkts	Pos	Inns
5.3	3	1	8	6.00	4	3	1
9.0	3	1	8	5.77	0	4	2
8.0	2	1	8	2.50	1	4	2
8.0	2	1	8	3.62	2	3	1
3.0	2	1	8	5.66	1	4	1
4.0	4	3	4	7.00	1	6	2
5.0	4	3	4	7.60	0	2	1
8.0	2	1	8	7.37	1	2	2
7.0	3	1	8	6.71	1	2	2
9.0	4	1	8	6.66	1	3	2
8.0	3	1	8	7.37	0	2	2

Figure 12: A part of the Bowler dataset after initial processing

3.2.2 Feature Selection

Feature Selection is an integral part of machine learning project. The result highly depends on the features that are used in the training set to train the model. Good relevant features dynamically improve the output prediction. On the other hand, selecting bad or irrelevant features may cause over fitting and can give a poor prediction.

There are many algorithms available which can select the features automatically, which have the most impact of the output feature. Out of many benefits of feature selection before going for the prediction and few of them are given below-

- One important advantage is that it reduces the over fitting problem.

- Secondly, there will be less misleading data and the accuracy will also improve and it reduces the training time

Automatic feature selection algorithms that we have used in our thesis are discussed below. We have used python Scikit-learn package to implement algorithm on our dataset and find out the relevant features. We selected 5 important variables out of 7 variables of out collected data using 2 different algorithms. [9]

3.2.3 Recursive Feature Elimination

This algorithm recursively removes the variables and keep the best variables for the training set. Here an external estimator assigns a value or weight to the input features as co-efficient. The estimator is trained on initial feature set and the importance of every feature is measured and the less important features are pruned. This procedure continues till the intended number of features are selected. We used Tamim's batting dataset and intent was to find out the 5 most important features from the input features which has greater impact on the output variable [10].

```
['BF', 'Pos', 'opposite', 'ground', 'pitch', '4s', '6s', 'SR']
feature number: 5
selected features: [ True  True False  True False  True  True False]
 Ranking:
[1 1 2 1 4 1 1 3]
```

Figure 13: The result of the RFE algorithm on Batsman dataset

We can see from Figure 13 that, features are ranked and the features ranked 1 are most important features in our dataset which are ball faced, playing position, ground, number of 4s and 6s.

```
['opposite', 'ground', 'pitch', 'Econ', 'Wkts', 'Pos', 'Inns']
feature number: 5
selected features: [ True False  True  True  True False  True]
 Ranking:
[1 3 1 1 1 2 1]
```

Figure 14: The result of the RFE algorithm on Bowler dataset

From Figure 14, we can see that opposition, pitch, economy, wickets and innings are the selected attributes for bowler datasets.

3.2.4 Univariate Selection

We have used another technique to our same dataset to find out the features. Select-K-Best method does different statistical test and from there K number of best features are selected and all other are removed.

The univariate feature selection is based on univariate statistical test. The scoring function we used is the Mutual Information. MI is a non-negative value which represents the dependency between variables. If the MI value is higher that means the variables are closely related and dependent and if the value is 0 then the variables are completely independent.

$$I(X;Y) = \iint p(x,y)\, log(\frac{p(x,y)}{p(x)p(y)})dxdy \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

This is the mutual information function shown in Equation (1), where p(x, y) denotes the joint probability function of x and y, and p(x),p(y) represents the marginal probability density function [11]. The function mutual_info_regression in Scikit learn which is used to calculate the MI is based on the entropy estimation from k-nearest neighbor distances [12] [13] [14].

In the figure 15, we can see the output of the univariate selection process on the batsman dataset. The score of each feature is written under the features and the features with the higher score are selected.

```
['BF', 'Pos', 'opposite', 'ground', 'pitch', '4s', '6s', 'SR']
[ 1.16677  0.05058  0.09229  0.       0.       1.18572  0.28795  0.5993 ]
[[   8.      6.      0.      0.     62.5 ]
 [  32.      6.      3.      2.     93.75]
 [  19.      7.      1.      0.     57.89]
 [  29.      7.      2.      0.     37.93]
 [  53.      1.      7.      2.     96.22]
 [   7.      5.      1.      0.     85.71]
 [   2.      7.      0.      0.     50.  ]
```

Figure 15: Feature selection using univariate selection for Batsman

We can see the attribute values of selected features. Ball faced, opposition, 4s, 6s, Strike rate are important features for batsman datasets.

```
['overs', 'opposite', 'ground', 'pitch', 'Econ', 'Wkts', 'Pos', 'Inns']
[ 0.39981  0.       0.       0.01769  0.63746  0.       0.       0.07428]
[[ 5.3   8.    6.    3.    1.  ]
 [ 9.    8.    5.77  4.    2.  ]
 [ 8.    8.    2.5   4.    2.  ]
 [ 8.    8.    3.62  3.    1.  ]
 [ 3.    8.    5.66  4.    1.  ]]
```

Figure 16: Feature selection using univariate selection for Bowler

From Figure 16, we have found for the bowler dataset- overs, pitch, economy, position and innings of bowling are the selected features.

3.2.5 Final Datasets after feature selection

After using the feature selection methodologies, we have got the selected features for batsman and bowlers. But there were some problems in the selected features for example:

the feature ‘4s’ and ‘6s’ means number of 4’s and 6’s a batsman will hit in the next match. Here this feature is totally uncertain because we cannot say how many 4’s or 6’s a batsman will score in the upcoming match. Though these are better features, we have to drop them. Finally, we have chosen five features for batsman to train and test our machine learning model. The screen-shot of a part of the dataset is given below in Figure 17 and Figure 18:

BF	Pos	opposite	ground	pitch
1	1	1	1	4
105	1	6	1	9
39	1	1	1	4
42	1	4	1	6
5	1	1	2	9
66	1	4	1	6
138	1	6	1	9
11	1	1	2	9
50	1	4	1	4
21	1	1	2	9
20	2	4	1	9
14	1	1	2	9
26	1	3	2	9
78	1	4	1	8

Figure 17: Part of the Batsman dataset after feature selection

Overs	opposite	ground	pitch
10.0	7	3	8
10.0	5	3	4
6.3	5	3	4
8.0	5	3	4
2.0	8	3	4
5.0	8	3	8
10.0	7	3	8
10.0	7	3	8
10.0	7	1	8
10.0	7	1	8

Figure 18: Part of the Bowler dataset after feature selection

CHAPTER 4

Implementation and Result Presentation

4.1 Batsman's run prediction process

Consequently, we dive into the batsman's run prediction. After the data processing the biggest challenge is to find out the Algorithm which is best suited for our dataset. We want to predict the run the batsman will score. So, this is a Regression problem as our output is continuous. At first, we have linear Regression and find out the performance in our dataset.

4.1.1 Batsman's Run prediction with Linear Regression

Linear Regression is very popular and simple machine learning technique. So, there is independent variable which represents the input features and there is dependent variable which is the output class. Linear regression uses the equation which is given below:

$$Linear\ Regression, \qquad Y = B_0 + \sum B_i \times X_i + e \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

B_0 is the intercept and B_i are the co-efficient that are learned during the model fitting time. Linear regression gives output which is continuous. Here X is the input variable and i denotes their number and e is simply the error function [15] [16]. In this method there are five input features such as- opponent, ball faced, ground, pitch and position. The output of this method is run. Therefore, for Tamim Iqbal's dataset we have used this technique and we have got 89.6% accuracy. For our model, the intercept value which represents the expected mean value of Y given all X values are 0 is 7.77498355227 and the co-efficient values are [0.9099559, -9.19083708, -0.71768333, 0.55779413, -0.16113162]. The co-

efficient represents the difference in prediction value, which is the input independent values are changed by per unit if all the other independent variables remain same. For example, if Tamim Iqbal plays one more ball he will score 0.9099559 runs given all the other variable remains same. To make it interpretable we want to expand this result and relate it with the input independent variables.

TABLE I: Calculated Co-efficient of Feature's

Variables	Calculated Co-efficient from the algorithm
BF	0.9099559
Pos	-9.19083708
opposition	-0.71768333
ground	0.55779413
Pitch	- 0.16113162

$$Run = 7.77498355 + 0.90995 \times BF - 9.190837 \times Pos - 0.71768333 \times opposition + 0.55779 \times ground - 0.161131 \times pitch \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

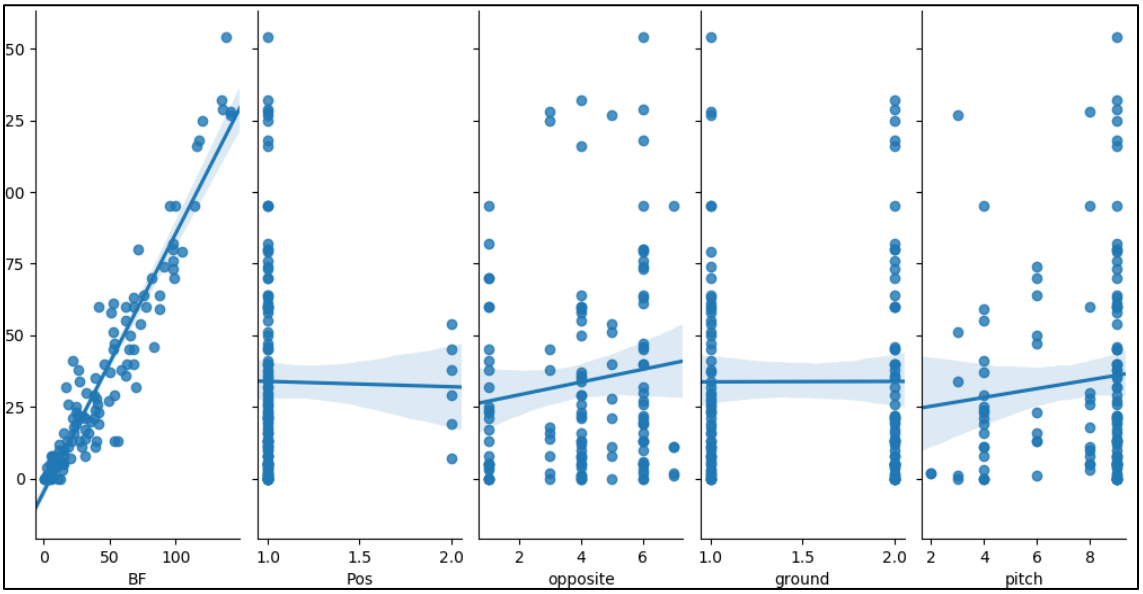

Figure 19: Pairwise best fitted Regression Line for Tamim's batting dataset

Therefore, our outcome accuracy is up to the mark with this model and this also shows that the input data are linearly related with the output variable. In Figure 19 above, we have visualized the data points pairwise with the output variable and found out the best fitted Regression line from that pair.

The next graph shows the picture of the output Runs and Actual runs scored by Tamim Iqbal. Here we can see the comparison between the outputs given by the Linear Regression model (y_pred) and the actual match score (y_test). From the Figure 20, it is easily visible that the predicted output is closely matched throughout all the test cases and maintains the accuracy. The graph is for the Tamim Iqbal's Batting dataset and using the same procedure we have calculated and predicted score for all other Batsmen using the Linear Regression method and the model worked almost equally good for the other Batsmen.

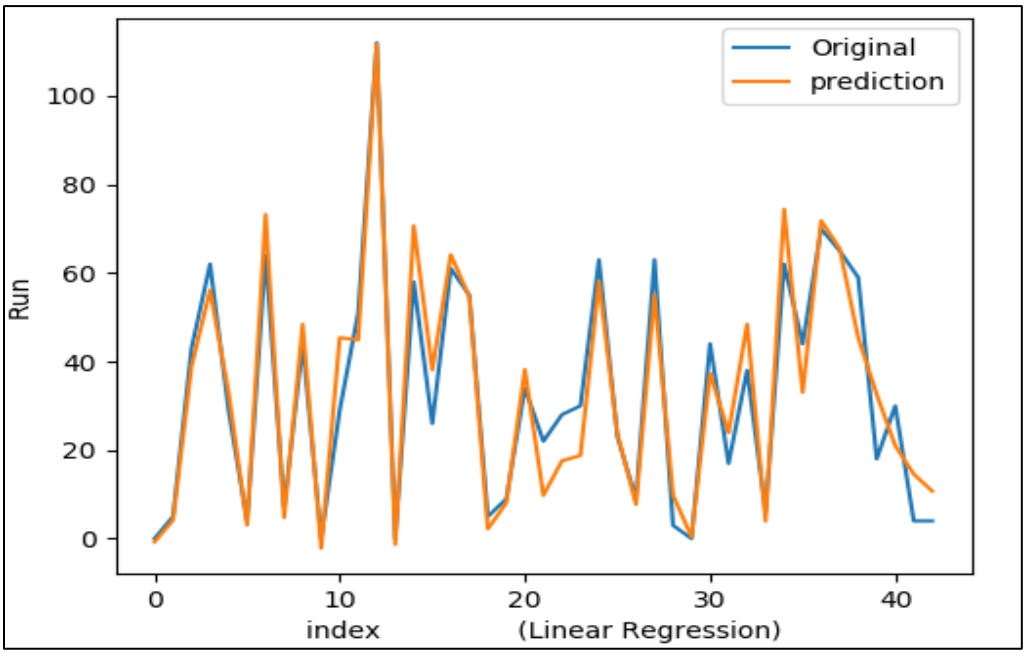

Figure 20: Result comparison with the real data of Tamim Iqbal's batting using Linear Regression

4.1.2 SVM with Linear and Polynomial Kernel

Support Vector Machine is another supervised learning technique which is globally popular for the performance and accuracy on the supervised dataset. It usually gives the prediction for discontinuous outcome and classifies the features into groups. Given a set of rows [[X_i, y_i], [X_j, y_j].....[X_k, y_k]] where the x denotes the feature vectors and y denotes the output class for that particular set of values SVM construct a hyper plane with maximum distance from the data points and classify the dataset into groups. The equation for the hyper plane:

$$Hyper\ Plane, \qquad w^T X + b = 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

Here "b" is the offset and "w" is the vector normal. Then the w is normalized by the w/ ||w|| and for decision making the below mentioned equation is calculated:

$$min \parallel w \parallel, \qquad y_n (w^T X_n + b) = \begin{cases} \geq 1 \\ < 1 \end{cases} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Thus, this equation finds and separate the input features Xi in to discontinuous group [10]. We have used the SVM technique but our task is to find out the Runs scored by the batsman which is not a class output rather it is a continuous output, for that we have used the SVM Regression technique which is offered by the Scikit learn and built over the SVM classification technique**.** We have used the SVM Regression where if we consider [(x_1, z_1) …… (x_l, z_l)] where the x is the input feature and y is the output variable and given parameter c, ξ are greater 0 then the form of support vector regression is:

$$\min_{w,b,\xi,\xi^*} \ \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i + \sum_{i=1}^{l} \xi_i^* \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

The dual problem is:

$$\min_{\alpha,\alpha^*} \frac{1}{2}(\alpha - \alpha^*)^T Q(\alpha - \alpha^*) + \sum_{i=1}^{l} \epsilon(-\alpha_i + \alpha_i{}^*) + \sum_{i=1}^{l} z_i(-\alpha_i - \alpha_i{}^*) \dots\dots\dots (7)$$

After solving the problem, the approximate function is:

$$\sum_{i=1}^{l} (-\alpha_i + \alpha_i{}^*)\, K(x_i, x) + b \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (8)$$

This is the function used by the Scikit learn library and the output is estimated [17] [18].

4.1.3 Tamim Iqbal batting score using Linear Kernel

Scikit learn SVM offers different kernel technique among which we have used the Linear and the polynomial kernel and measure the performance on our datasets. The linear kernel worked better than the linear Regression and overpass the score of the Linear Regression with a score of 92 percent accuracy and the Root Mean Square error is 2.44. The comparison graph between the actual data (y_test) and the predicted data (y_pred) is given below in Figure 21.

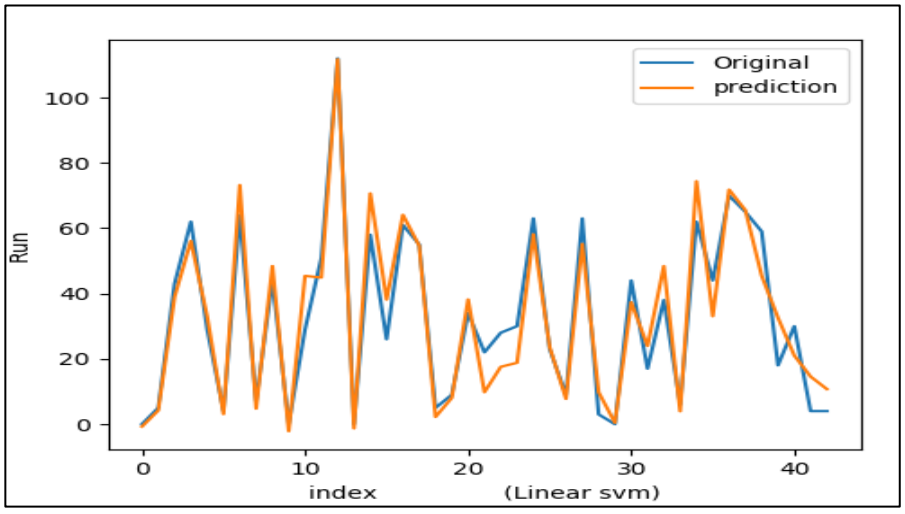

Figure 21: Result comparison with the real data of Tamim Iqbal's batting using Linear Kernel technique of SVM

4.1.4 Tamim Iqbal batting score using polynomial Kernel

The polynomial kernel is given by the function $\gamma\left((x, x') + r\right)^d$. d is the degree and r are the co-efficient. The score using this technique for Tamim Iqbal is 87%. The performance is slightly lesser than the linear kernel because the data are linearly distributed for Tamim Iqbal. Here is the result comparison graph.

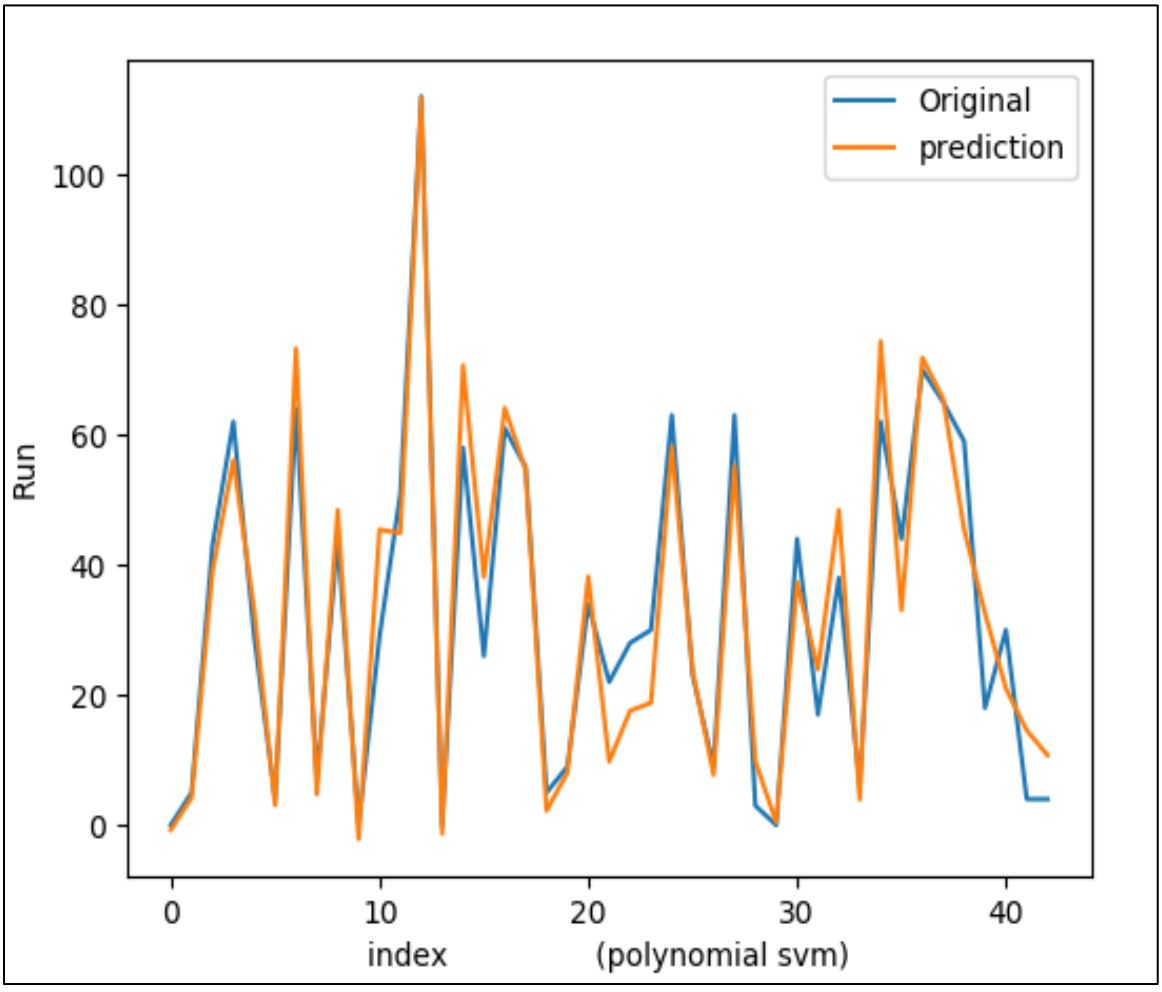

Figure 22: Result comparison with the real data of Tamim Iqbal's batting using Polynomial Kernel technique of SVM

As we have got the output score in which the accuracy is quite high so we used this above-

mentioned technique for all the batsmen and find out the scores and root mean square error for all of them. Here is the result shown as table:

TABLE II: Batsmen Score Using Linear and Polynomial Kernel

Name	Linear	Polynomial
	Score	Score
Tamim	0.9153756	0.8735332
Shakib	0.8488194	0.8731244
Mushfiq	0.8524944	0.7934314
Mahmudullah	0.9004662	0.7378206
Sabbir	0.8581789	0.5913919
Mashrafee	0.5358919	0.5767481
Nasir	0.8520773	0.8464901
Rubel	0.3790782	0.0597999

TABLE III: Batsmen Error Using Linear and Polynomial Kernel

Name	Linear	Polynomial
	Error	Error
Tamim	2.44254090865	2.7050651923
Shakib	2.79033305164	2.65284459936
Mushfiq	2.85812965164	3.01763098569
Mahmudullah	2.33643283855	2.76696691863
Sabbir	2.4168873237	3.172919188
Mashrafee	2.38784477809	2.41834979965
Nasir	2.40743235386	2.96136121369
Rubel	1.35533719848	1.51498749017

4.2 Bowler's performance prediction process

Now moving on to the bowler, the procedure we used for bowler's data representation is almost same as the batsman. For bowlers we have taken the [Over, Opposition, Ground, pitch] as input features. The opponent teams and pitch are both converted to the numerical value like batsman.

4.2.1 Prediction of Runs given by a Bowler in his spell

We will now predict the runs given by a bowler. First, we take Mashrafee's bowling dataset and we have used the SVM Regression with polynomial kernel. But the output accuracy is not up to the mark in this time. Score is 28 percent and error are 2.87 percent.

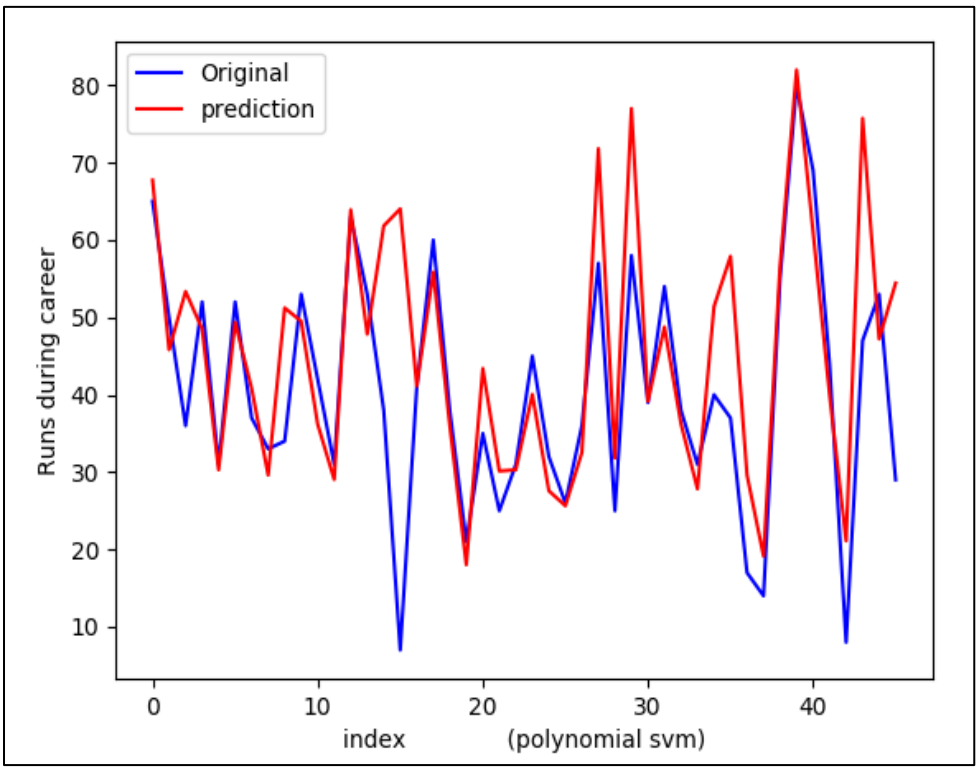

Figure 23: Result comparison with Real data of Mashrafee using polynomial kernel-SVM

Better visualization from the graph of Figure 23 shows, we need to improve our score as the difference between original and predicted data is slightly higher. For that, we have used a train test split technique called K-fold cross validation.

4.2.2 K-fold Cross Validation for Mashrafee's Dataset

The concept of K-fold cross validation is to divide the dataset into K smaller sets. Then using K-1 folds as training data the model is trained and the outcome is evaluated on the remaining part of the data which is regarded as test data. This process will continue for K times and the performance is measured based on the average of the outcome in each iteration [19]. This method provides a slight advantage because here whole dataset is trained and tested by k times of iteration. Every time it iterates, it trains and tests different portion of dataset.

Therefore, we have taken Mashrafee's dataset again to improve the score and we have done the 5-fold Cross Validation on the support vector machine Linear kernel. This technique works well in our dataset and now the output score is 47 percent which is the average of each fold cross validation outcome.

```
[ 0.48788853  0.23232976  0.4615619   0.46850342  0.71846701]
Accuracy: 0.47 (+/- 0.31)
```

Figure 24: Outcome of K-fold cross validation on Mashrafee's bowling

We can see the outcome of each iteration in the Figure 24. It is also noticeable that in the 5th fold iteration, our model worked way better by giving 71 percent accuracy and boost up the average score of the model.

4.2.3 Processing text data for Bowler's Dataset

Text data processing is one of the major field in machine learning. Various methods and systems are available for this text conversion which are described in [20].

To make our dataset more generalized and randomized we have done the text data processing. Input feature ground and team are textual data. Initially we assigned a number based on their importance but to improve the accuracy now we want to process the text data and build a generalized and bigger dataset. To represent the text data into features Scikit learn offers Bag of word representation utilities. Tokenizing is one of the utilities where each string is tokenized and an integer id is given for each string. Here white-spaces and punctuation are used as splitting factor.

- Each individual token is treated as feature.
- The occurrence of that tokens in each feature are treated as feature value.

Now we want to evaluate this technique on the player who is comparatively new and have small dataset. We have chosen Taskin Ahmed's bowling dataset. The opponent column is in text form, so we have taken the column and extract the features out of this. The total number of features from the opponent column for Taskin Ahmed is 13 and they are:

['afghanistan', 'africa', 'england', 'india', 'indies', 'lanka', 'new', 'pakistan', 'scotland', 'south', 'sri', 'west', 'zealand']

After extracting the feature, we have to fill out the whole column with the respective team information against whom Taskin Ahmed played. For that we have transformed the whole column of opponent into the [13: column length] matrix.

[[0 0 0 1 0 0 0 0 0 0 0 0 0]

[0 0 0 1 0 0 0 0 0 0 0 0 0]

[0 0 0 0 1 0 0 0 0 0 0 1 0]

[1 0 0 0 0 0 0 0 0 0 0 0 0]

[0 0 0 0 0 1 0 0 0 0 1 0 0]

[0 0 0 0 0 0 0 0 1 0 0 0 0]

[0 0 1 0 0 0 0 0 0 0 0 0 0]

[0 0 0 0 0 0 1 0 0 0 0 0 1] …….

Figure 25: Matrix representation of the “opponent” column after feature extraction and transformation.

Following the same way, the ground list is also processed and then 19 features were extracted. The features are: ['adelaide', 'birmingham', 'canberra', 'cardiff', 'chittagong', 'christchurch', 'colombo', 'dambulla', 'dhaka', 'east', 'george', 'hamilton', 'kimberley', 'london', 'melbourne', 'nelson', 'paarl', 'ssc', 'st']

After feature extraction and matrix transformation we have converted them to the Pandas data frame again and then concatenated this data frames with the previously available “Econ” column and find out the final data frame which has the size of [31,33] where 31 is the number of rows and 33 is the number of column or features which is partly shown in Figure 26.

Now we have used SVM with linear kernel and find out the score. The score now is 68 percent for Taskin Ahmed Dataset and the error is 2.75. Figure 27 shows the output visualization graph shown from Taskin Ahmed’s bowling dataset.

Econ	afghanistan	africa	england	india	indies	lanka	new	pakistan	\
4.42	1	0	0	0	0	0	0	0	
3.50	0	0	0	1	0	0	0	0	
5.37	0	0	0	0	0	0	1	0	
5.32	0	0	0	0	0	1	0	0	
3.50	0	0	0	1	0	0	0	0	
5.25	0	0	0	0	0	0	0	1	
1.87	0	0	0	1	0	0	0	0	
7.77	0	0	0	0	0	0	1	0	

scotland	...	east	george	hamilton	kimberley	london	melbourne	\
0	...	0	0	0	0	0	0	
0	...	0	0	0	0	0	0	
0	...	0	0	0	0	0	0	
0	...	0	0	0	0	0	0	
0	...	0	0	0	0	0	0	
0	...	0	0	0	0	0	0	
0	...	0	0	0	0	0	0	
0	...	0	0	0	0	0	0	

nelson	paarl	ssc	st
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

Figure 26:Taskin Ahmed dataset after final text data processing

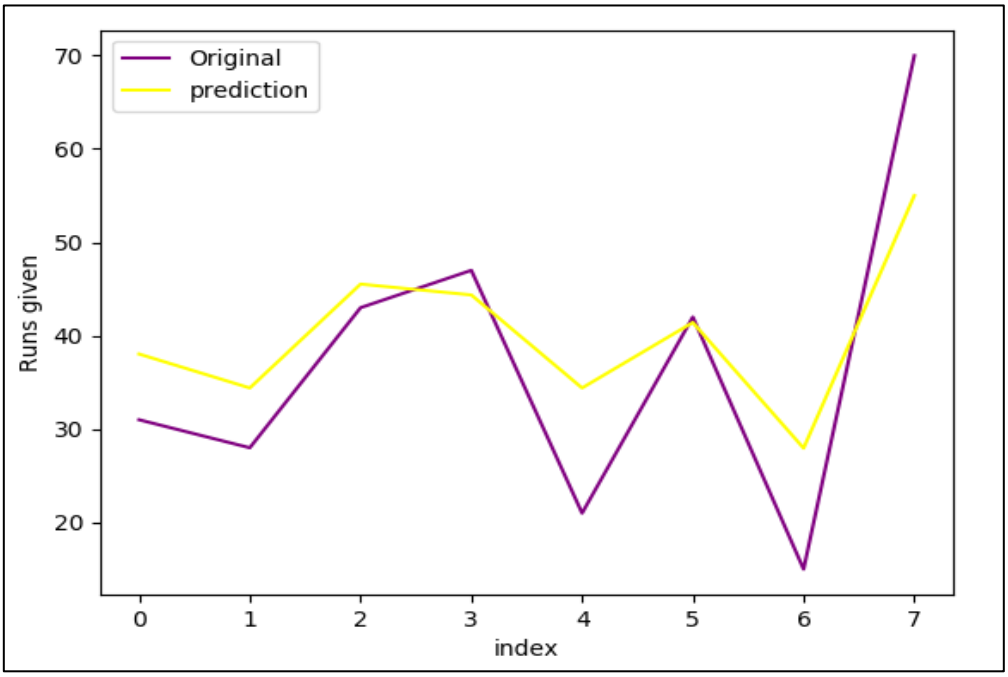

Figure 27: Comparison of Taskin's predicted runs and actual runs (Text processing approach)

This approach gives a better result than the previous approaches. But this approach is computationally expensive because it converts dataset into a huge size.

Now we want to test this approach on a bigger dataset, therefore we have taken Shakib Al Hasan bowling data and used the same approach of text processing and this time from we have got a data set of 177 rows and 79 columns and the output score is now 44 percent.

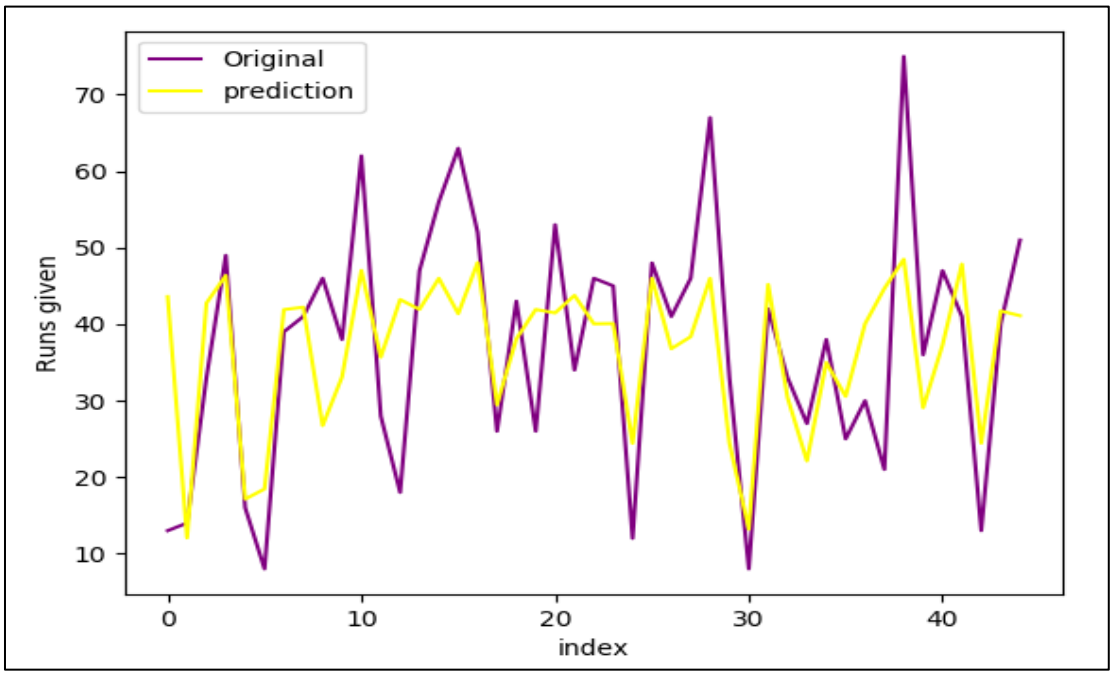

Figure 28: Comparison of Shakib's predicted runs and actual runs (Text processing approach)

But this approach degrades the accuracy of our model in the case of Shakib Al Hasan because previously without text processing where the input data had 177 rows and 4 columns, the output accuracy was 52.5 percent with a 2.95 root mean square error. The reason is here the number of feature has increased 79 but the number of data point or row is only 177 which did not increase. Thus, from the Figure 28, we can come to the decision that text processing approach may work better in small dataset but the performance of the

model degrade as the dataset size increases. Therefore, we can come to the decision that our previous way of manipulating data where opponent teams were replaced via their ICC Rankings and the Ground were replaced via their importance number, was the better approach for bowler dataset and this is the optimal score we have found using the supervised machine learning algorithms. So, for more performance improvement we now want to apply neural network in our bowler's dataset.

4.2.4 Neural Network for bowler dataset

As we are not able to get more accuracy using supervised learning algorithm. Therefore, we have used neural network for a small part of the dataset. In Figure 29, we compared the predicted runs and actual runs of Shakib Al Hasan by implementing neural network.

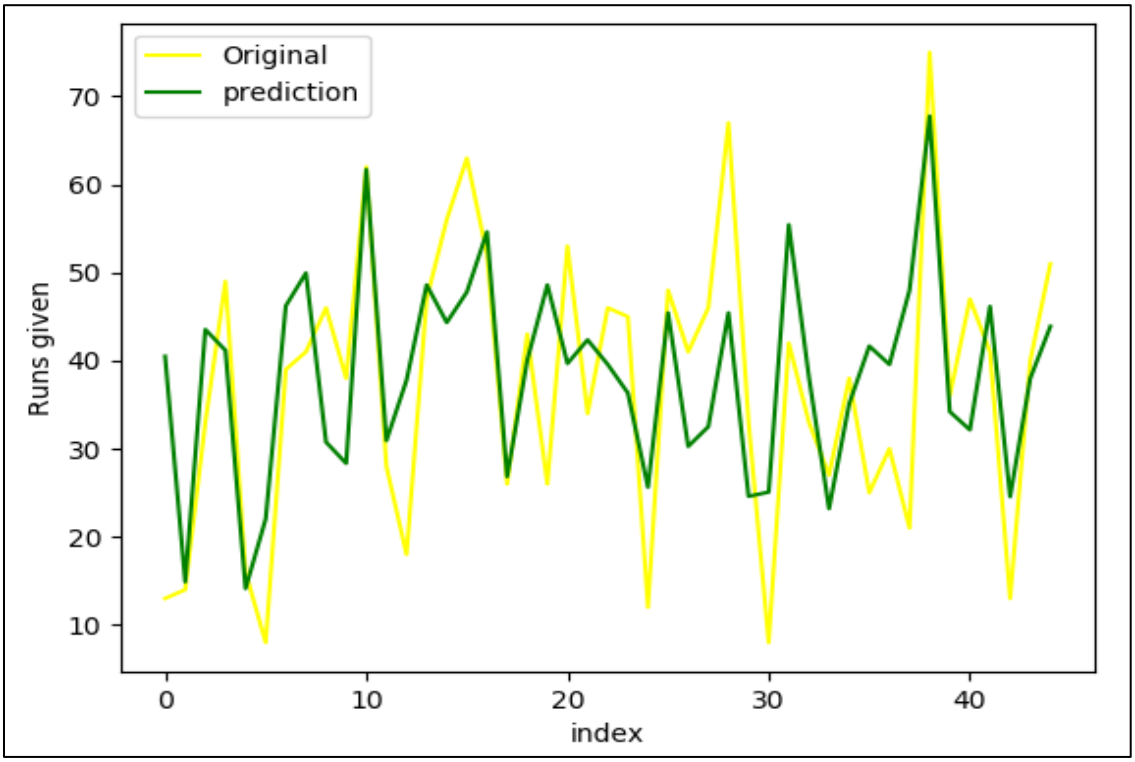

Figure 29: Shakib's Run prediction and actual runs comparison using neural network

To implement the neural network, we used “keras” sequential model which is basically the stack of layers. We have used the activation function “Relu” because it offers the range between [0 to infinity] and this is appropriate for our dataset because our output variable is also continuous which can be any numerical value starting from 0. We have used one input layer, one output layer and two hidden layers. The output root mean squared error is 3.29 which cannot over pass the supervised learning algorithm performance.

4.3 Team Combination

A team combination is a tricky and complex process because there are so many different approaches to follow while selecting any player for the final team. The process of selecting any player vastly depend on the opposition, playing role of the player, weather condition, home or away match, etc. We are going with five batsmen with the most predicted runs, four bowlers who would be picked on the basis of the combination of least given runs and most taken wickets, a wicket-keeper with the highest predicted run and an all-rounder with the combination of predicted scored runs while batting, given runs while bowling as well as taken wickets.

For selecting the best eleven players for the team- firstly, we split up the pool of players in which we categorize the player from the pool into “Batsman”, “Bowler”, “All-rounder”, “Keeper” sections respectively.

```
['Tamim', 'Soumya', 'Imrul', 'Sabbir', 'Muminul']
['Mashrafee', 'Mustafiz', 'Rubel', 'Taskin', 'Shafiul', 'Taijul', 'Al-Amin', 'Sunny']
['Shakib', 'Mahmudullah', 'Mosaddek', 'Nasir', 'Saifuddin']
['Mushfiq', 'Liton', 'Anamul']
```

Figure 30: Player categorization from the player pool

Figure 30 shows the categorization of the players according to their skills. Here, first category is for pure batsmen, second row is for pure bowler, third represents the all-rounders and the last row is for keepers.

4.3.1 Batsmen Selection Process

In second step, the algorithm will be conducted for all the batsmen, all-rounders and keepers for each position from position "one" to position "five". We listed the predicted runs for each player. Figure 31 shows the player name, player category, scored run (predicted) and number of innings batted.

```
['Tamim', 'batsman', 42, 171]
['Soumya', 'batsman', 37, 31]
['Imrul', 'batsman', 32, 70]
['Sabbir', 'batsman', 42, 41]
['Muminul', 'batsman', 0, 24]
['Mushfiq', 'keeper', 40, 164]
['Liton', 'keeper', 18, 12]
['Anamul', 'keeper', 34, 31]
['Shakib', 'all-rounder', 34, 170]
['Mahmudullah', 'all-rounder', 26, 123]
['Mosaddek', 'all-rounder', 18, 12]
['Nasir', 'all-rounder', 6, 47]
['Saifuddin', 'all-rounder', 0, 4]
```

Figure 31: Player role, predicted run and match number of players for position one

Among them; the best suited player for a particular position would be selected. However, the experience of players would be taken also into account as the tiebreaker among the players with the similar number of predicted runs.

$$Batsmen\ Score = (Predicted\ Run \div 5) + (Total\ Match\ Played \div 100) \dots\dots\dots (9)$$

To measure "Batsman Score", we added tiny 1 percent experience with 20 percent of their predicted run to select the right batsman for the team. In the game of cricket, experience

matters to handle any critical situation. Though we are adding experience of the player to give him some advantage in team selection process, we took 20 percent predicted run to calculate batsman score to select the player who can score more runs (minimum 10-15 runs) than an experienced player. “Tamim” and “Sabbir” scored equal runs for the position number “one”. But experience of “Tamim” lead him the way to be selected in the team first. “Tamim” is selected ahead of “Sabbir” in the team by beating him in a low margin of score, which is (10.11- 8.81) = 1.3 point. If “Sabbir” scored 10 runs (10/2= 2 points) more than “Tamim” then he would have been selected in Tamim’s place by beating him by 0.7 point. 10 runs are very crucial in cricket specially in the shorter form of the game. We tried to minimize the uncertainty of the selection process between two batsmen if they score same number of runs by incorporating experience to calculate their scores which eventually helped us to select the right player for the team.

Figure 32 shows that, the score of “Tamim” is higher than other players for position one; which got him selected for that position.

```
('Saifuddin', 0.04)
('Nasir', 1.67)
('Mosaddek', 3.72)
('Mahmudullah', 6.43)
('Shakib', 8.5)
('Anamul', 7.109999999999999)
('Liton', 3.72)
('Mushfiq', 9.64)
('Muminul', 0.24)
('Sabbir', 8.81)
('Imrul', 7.1000000000000005)
('Soumya', 7.71)
('Tamim', 10.11)
```

Figure 32: Score calculated from predicted run and number of match played for pos-1

After selecting “Tamim” for position one, we removed him from the player pool and “Batsman” category as he is already got selected. This process is continued till position “Five” for selecting five best batsmen for each position in the team. Overall selection process of Tamim Iqbal for the position-1 is shown below in Figure 33:

```
['Tamim', 'batsman', 42, 171]
['Soumya', 'batsman', 37, 31]
['Imrul', 'batsman', 32, 70]
['Sabbir', 'batsman', 42, 41]
['Muminul', 'batsman', 0, 24]
['Mushfiq', 'keeper', 40, 164]
['Liton', 'keeper', 18, 12]
['Anamul', 'keeper', 34, 31]
['Shakib', 'all-rounder', 34, 170]
['Mahmudullah', 'all-rounder', 26, 123]
['Mosaddek', 'all-rounder', 18, 12]
['Nasir', 'all-rounder', 6, 47]
['Saifuddin', 'all-rounder', 0, 4]
==========
('Saifuddin', 0.04)
('Nasir', 1.67)
('Mosaddek', 3.72)
('Mahmudullah', 6.43)
('Shakib', 8.5)
('Anamul', 7.109999999999999)
('Liton', 3.72)
('Mushfiq', 9.64)
('Muminul', 0.24)
('Sabbir', 8.81)
('Imrul', 7.1000000000000005)
('Soumya', 7.71)
('Tamim', 10.11)
==========
Selected Player:
['Tamim', 'batsman', 42, 171]
=========================
10.11
=========================
```

Figure 33: Overall process of the player selection for position one

4.3.2 All-rounder Selection Process

Then for the number six position, by using the algorithm we predicted scored runs, given runs in bowling and number of wickets taken for all the all-rounders who have been listed and find out the best fitted player for this position. Here, predicted batting performance is given higher importance than their predicted bowling performance. We evaluated an all-rounder as a better batsman than a bowler. $\text{All} - \text{Rounder Score}$,

$$(Predicted\ Run \times 0.65) - \left[\left\{\frac{Runs\ Given}{10} - Wickets\ Taken\right\}^3 \times 0.35\right] \dots\dots\dots (10)$$

To select an all-rounder, we calculated his bowling performance and extracted some points from the performance. For extracting points from bowling performance, we divided the runs he given by 10 and then we subtracted the number of wicket taken by him. In this calculation, the subtracted value should be smaller (may be a negative value). After that, we cubed the value to exaggerate the value even bigger. Here, the batting and the bowling performances were given importance of 65 percent and 35 percent respectively. We subtracted bowling points from the batting performance points to calculate overall point. Therefore, we achieved our desired all-rounder.

```
['Mahmudullah', 'all-rounder', 21, 123, 46, 1]
['Mosaddek', 'all-rounder', 21, 12, 58, 0]
['Nasir', 'all-rounder', 18, 47, 38, 2]
['Saifuddin', 'all-rounder', 0, 4, 112, 5]
Selected Player:
['Nasir', 'all-rounder', 18, 47, 38, 2]
=========================
-83.4148
=========================
```

Figure 34:Player selection for position six

From this process, our algorithm selected “Nasir” instead of “Mahmudullah”. Though “Mahmudullah” might score 21 runs in the match but he might get 1 wickets and costs 46 runs from his 10 over spell which is crucial factor in the match. In contrast, “Nasir” scored 18 runs, got 2 wickets by costing 38 runs; which is better than “Mahmudullah”.

4.3.3 Keeper or extra batsman Selection Process

After that, for the number seven position, the algorithm looks if any keepers have already been selected between one-to-five positions. If so, then the algorithm will run on all the batsmen, keepers and all-rounders categories remaining in the list who were not selected earlier in the process and select the player with most predicted run.

```
['Imrul', 'batsman', 0, 70]
['Sabbir', 'batsman', 12, 41]
['Muminul', 'batsman', 0, 24]
['Liton', 'keeper', 12, 12]
['Mahmudullah', 'all-rounder', 10, 123]
['Mosaddek', 'all-rounder', 12, 12]
['Saifuddin', 'all-rounder', 0, 4]
('Saifuddin', 0.04)
('Mosaddek', 2.52)
('Mahmudullah', 3.23)
('Liton', 2.52)
('Muminul', 0.24)
('Sabbir', 2.81)
('Imrul', 0.7)
Selected Player:
['Mahmudullah', 'all-rounder', 10, 123]
=========================
3.23
=========================
```

Figure 35:Player selection for position seven

On the contrary, if any wicket-keeper is not selected previously then only the list of wicket-keepers will be taken into account here and predict the best performer possible. In this case,

two keepers- “Mushfiq” and “Anamul” have already got selected for the batsman category. Therefore, according to the algorithm- remaining players from the batsmen, keepers and all-rounders categories will be evaluated and the player will be selected who might score more among those remaining players. Here, “Mahmudullah” got selected as he scored almost similar like “Sabbir” but got more experience than “Sabbir” or other players.

4.3.4 Bowler Selection Process

Finally, for the four genuine bowlers, the algorithm will be conducted on all the bowlers and the remaining all-rounders and in the process select the players with the balanced combination of given runs and taking wickets. We evaluated the score of the bowlers by dividing the given runs by 10 then subtracting the number of wicket taken by him. This gives us a value which should be small to indicate good bowling performance of the bowler.

$$Bowler\ Score = \left(\frac{Runs\ Given}{10} - Number\ of\ Wickets\ Taken\right) \dots\dots\dots\dots\dots\dots (11)$$

```
['Mashrafee', 'bowler', 36, 2]
['Mustafiz', 'bowler', 36, 5]
['Rubel', 'bowler', 50, 2]
['Taskin', 'bowler', 50, 3]
['Shafiul', 'bowler', 43, 2]
['Taijul', 'bowler', 34, 0]
['Al-Amin', 'bowler', 51, 2]
['Sunny', 'bowler', 34, 2]
['Mosaddek', 'all-rounder', 58, 0]
['Nasir', 'all-rounder', 38, 2]
['Saifuddin', 'all-rounder', 112, 5]
-2
Selected Player:
['Mustafiz', 'bowler', 36, 5]
==========================
-2
==========================
```

Figure 36:Player selection for bowling

Here, “Mustafiz” considered 36 runs in his bowling but he was able to pick up 5 wickets which gets him the negative value in his bowling score. In this case, negative value represents that “Mustafiz” might cost 36 runs but by picking up 5 wickets, he saved more runs than he costs. Finally, we have all the players for the team which is mostly balanced in the scale of different sectors of expertise. There are 5 genuine batsmen, 3 all-rounders and 3 genuine bowlers in the team.

```
==========================
['Tamim', 'batsman', 42, 171]
==========================
==========================
['Shakib', 'all-rounder', 27, 170, 35, 2]
==========================
==========================
['Soumya', 'batsman', 42, 31]
==========================
==========================
['Mushfiq', 'keeper', 22, 164]
==========================
==========================
['Anamul', 'keeper', 24, 31]
==========================
==========================
['Nasir', 'all-rounder', 18, 47, 38, 2]
==========================
==========================
['Mahmudullah', 'all-rounder', 10, 123, 46, 1]
==========================
==========================
['Mustafiz', 'bowler', 0, 13, 36, 5]
==========================
==========================
['Taskin', 'bowler', 1, 16, 50, 3]
==========================
==========================
['Mashrafee', 'bowler', 5, 131, 36, 2]
==========================
==========================
['Sunny', 'bowler', 37, 11, 34, 2]
==========================
```

Figure 37: Final team selection

We can calculate the runs scored by the batsmen and then we can compare it with the runs given by the bowlers to predict the match result as well. If the run scored by the batsmen are higher than the run considered by the bowlers then we can say that this team can win the match. We can also show the margin of winning by comparing the run scored by the batsmen and run given by the bowlers. There is certain drawback in this process as we did not calculate the number of extra runs given which can be change the match result. But a fair assumption of winning condition can easily be made by the comparison.

TABLE IV: Scored Runs of Batsmen and Runs Given by Bowler

Name	Run Scored	Run Considered
Tamim	42	-
Shakib	27	35
Soumya	42	-
Mushfiq	22	-
Anamul	24	-
Nasir	18	38
Mahmudullah	10	46
Mustafiz	0	36
Taskin	1	50
Mashrafee	5	36
Sunny	37	34
Total	228	194(from 5 bowlers)

CHAPTER 5

Conclusion, Limitation and Future Scope

5.1 Conclusion

In our country cricket is now the most popular game. Day by Day the game of cricket is introducing newer technology and technique to make the game smooth and reliable to spectators. But still player selection process and team combination process remain manual. Therefore, our main motivation was to make the team combination selection process automated. With this view in mind we worked to gather all the previous data of the players who are in standby list and find out the runs scored by a batsman and runs given by a bowler using Machine Learning Algorithms. The accuracy of batsman model was up to the mark but we faced problem of low accuracy with bowler's model. Thus, we have used different techniques for example: K-fold Cross Validation, Bag of word to increase the accuracy. Moreover, we have also deployed Neural Network (fully connected dense layer) on the bowler dataset and find out the performance comparison. Consequently, after getting the individual runs for each bowler and batsman we have reused the previous code and tests the batsman in different playing position for example: batting first or second or any other position which batsman scores more runs. After selected the pure batsman we also selected keeper and an all-rounder. Last but not the least we have selected bowlers using the same technique as batsman. During the whole process we have tried to find out the winning team combination and that is the reason why we have used the above-mentioned methodology.

5.2 Limitation

Our models and methods work for only ODI matches. It does not work in T20 or Test matches because we did not train our models on T20 or test data. Moreover, some of the bowlers predicted scores are not up to the mark because of the lack of data. The domestic match's data can be a way here which we did not consider here but this could be a future scope of this work.

5.3 Future Scope

Our method is currently focused on Bangladesh team but there are still other international teams with huge dataset of their players. We will work on domestic level which will eventually improve the accuracy of the model because the more dataset it gets, the better the machine can predict. Moreover, analyzing the opponent team or the opponent best batsman or the best bowler can facilitate the home team. Therefore, there is a big scope in this area to work with using these methodologies.

REFERENCES

[1] A. Kusiak, J. A. Kern, K. H. Kernstine, and B. T. Tseng. 2000. *Autonomous decision-making: a data mining approach*. Information Technology in Biomedicine, IEEE Transactions on, vol. 4, no. 4, pp. 274–284, 20007.

[2] "*Mickey Arthur Cites Racial Discrimination for Cricket Australia"*. 2013.[online]. Via: http://www.smh.com.au/news/cricket/pietersen-tells-of-racialdiscrimination/2006/08/30/1156816967716.html.

[3] R. L. Holder, A. M. Nevill. 2009. *Modelling performance at international tennis and golf tournaments: is there a home advantage?* Journal of the Royal Statistical Society: Series D (The Statistician), vol. 46, no. 4, pp. 551–559, 2009.

[4] H. H. Lemmer. 2010. *A measure for the batting performance of cricket players*. South African Journal for Research in Sport, Physical Education and Recreation, vol. 26, no. 1, pp. p–55, 2010

[5] M. Sipko. 2015. *Machine Learning for the Prediction of Professional Tennis Matches*. Imperial College London, 2015.

[6] R. A. Torres. 2013. *Prediction of NBA games based on Machine Learning Methods*. University of Wisconsin Madison, 2013.

[7] "HowSTAT! The Cricket Statisticians - Home Page", Howstat.com, 2018. [Online]. Available: http://www.howstat.com/. [Accessed: 10- Mar- 2018].

[8] "ESPNcricinfo - Cricket Teams, Scores, Stats, News, Fixtures, Results, Tables", ESPNcricinfo, 2018. [Online]. Available: http://www.espncricinfo.com/. [Accessed: 10- Mar- 2018].

[9] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion. 2011. Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825-283.

[10] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 201

[11] Cover, T. M. and Thomas, J. A. (2001) Entropy, Relative Entropy and Mutual Information, in Elements of Information Theory, John Wiley & Sons, Inc., New York, USA.

[12] A. Kraskov, H. Stogbauer and P. Grassberger, "Estimating mutual information". Phys. Rev. E 69, 2004.

[13] B. C. Ross "Mutual Information between Discrete and Continuous Data Sets". PLoS ONE 9(2), 2014.

[14] L. F. Kozachenko, N. N. Leonenko, "Sample Estimate of the Entropy of a Random Vector", Probl. Peredachi Inf., 23:2 (1987), 9-16.

[15] Weisberg, Sanford. *Applied linear regression*. Vol. 528. John Wiley & Sons, 2005.

[16] Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to linear regression analysis*. Vol. 821. John Wiley & Sons, 2012.

[17] LIBSVM: A Library for Support Vector Machines Chih-Chung Chang and Chih-Jen Lin Department of Computer Science National Taiwan University, Taipei, Taiwan Email: cjlin@csie.ntu.edu.tw Initial version: 2001 last updated: March 4, 2013.

[18] "Support-vector networks", C. Cortes, V. Vapnik - Machine Learning, 20, 273-297 (1995).

[19] R. Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, Intl. Jnt. Conf. AI.

[20] Sebastiani, Fabrizio. "Machine learning in automated text categorization." *ACM computing surveys (CSUR)* 34.1 (2002): 1-47.