PREDICTING ACCURACY OF PLAYERS IN THE CRICKET USING MACHINE LEARNING

Mr. MUJAMIL DAKHANI [1], UMME HABIBA MAGINMANI[2]

[1]*Asst professor, Computer Science and Engineering, Secab Institute of Engineering and Technology ,Karnataka, India*

[2]*M.Tech student, Computer Network and Engineering, Secab Institute of Engineering and Technology ,Karnataka, India*

---***---

Abstract –*The most important and decisive task in any sport is selection of players. Evaluation of individual performance and selection of the players in cricket is most critical job. The player's performances depend on numerous factors such as the location where the match being play, past records, his current form, average rate, strike rate, run scored at a particular venue, number of inning played against the opposition teams etc. The member of selection board, the coach and the captain of the team is conscientious for player selection. They explore special statistic, records and characteristic of the players to select the finest playing11 for each match. Throughout selection process of the players the batsman and bowlers are rated on basis of the batting and bowling average correspondingly. However in the game like cricket it is always important the situation in which it is responsibility of the batsmen to scores at most runs and bowlers to claims wickets. .In this project we attempt to predict the performance of the players. For the prediction model these 2 problem are taken into account as goal as classification problem where the number of 'runs' and number of 'wickets' are classified into dissimilar range using different classifier algorithm. We exploit 'Decision tree', 'Naive Bayes', 'Random Forest' and 'Multiclass SVM' classifiers which produce the useful model to Predict for these 2 problems. Out of these 4 classifier algorithm the random forest classifier produce more accurate result then compared to other classifier and SVM produce least useful result.*

***Key Words*: Cricket, Decision Trees, Naive Bayes, Random Forest, MultiClass SVM, Performance, Accuracy, Prediction Model.**

1.INTRODUCTION

Cricket is game which is being played between two team of eleven players in each team where scoring the 'runs' or taking the 'wickets' is major responsibility of players. This is as a rule done by hitting the ball across the boundary or by taking run by running in between pitch. This pitch is prepared of two set of three wooden post . This pitch is called "wickets". The "pitch" is of 20-metre (22 Yard)in span with the wickets at each side, with each side consist of two bails of 3 stumps. The cricket field is made either circular shaped or oval-shaped grassy ground. The diameter of the ground ranges between 137metre(450feet) to 100metre(500 feet).Team is made up of department of bowler, fielder, batsmen and all rounder and one wicket keeper. The wicket keeper is also a fielder who stand behind the wicket or stump and keep an eyes to focus on batsmen and ready to get catch or stumps the batsmen. Wicket keeper is fielder who is allowable to put on the gloves and external leg guards. The task of the batting department is to contribute toward the wining of match by scoring runs by striking the ball bowled at the wicket using his bat. fielding department responsible to prevent runs and tries to dismiss the batsmen by taking catch or run out and bowling department is to take maximum wickets and to put a ceiling on the other team from scoring runs at same time. All rounder means those players who can performs in the match by batting as well as bowling and they donate towards the team by taking wickets and scoring the runs. After the completion of an inning or after dismissed of all the player of opposite team, the teams switch the roles. Each player donates en route for the overall performance of the team by giving their best performance in each match. The match is evaluated by two umpire in each match and with the aids of 3rd umpire and referee of match in the International matches. The player's performances depend on various factors such as the location where the match being play, past records, his current form, average rate, strike rate, run scored at a particular venue, number of inning played against the opposition teams etc. There are the numerous format of the cricket such as Test match which is played for 5 days with unlimited number of over and length and each team get two inning for batting as well as bowling of unlimited length. In the T20 match format the match is usually played for few hour (3 to 4hour) which each team bat and bowl for single inning of 20overs. One day International (ODI) matches which of 50over.

It is important to select the right players that can contributes their best performance in match in which they get selected . The conscientiousness of the team selection authority, the coach and the captain is to examine each player's ability, characteristics and past record, stats to select the optimum 11 player. They will rate the players according to their past record.

In this project, we attempt to predict the player's performance in One Day International (ODI) matches by analyzing their earlier records using supervised machine learning techniques. For this purpose we predict the performance batsmen's and bowler's disjointedly.

2. RELATED WORK

A few online article produced some useful information related to players performance for the prediction in the game of the cricket.

S.Muthuswamy and S.S.Lam[1] "Bowler Performance Prediction for One-day International Cricket Using Neural Networks," In this paper they predict the performance of Indian bowler against 7 international teams. A neural network approach using Back Propagation Network[BPN] and Radial Basis Function Network(RBFN) used to predict the performance of Indian cricket team bowler. Later Performance of BPN and RBPN model was compared for the prediction and classification.

G. D. I. Barr and B. S. Kantor [2]"A Criterion for Comparing and Selecting Batsmen in Limited Overs Cricket,". In this paper the author defined major Criteria for comparing of the and selecting of the batsmen in the limited over crickets. This paper shows a 2D graphical representation of strike rate on one axis and probability of getting out i.e. P(out) on the other axis. We develop the selection principle of batsmen based on this 2D framework which coalesce the average and strike rate As an example for this application we apply this principle to batting performance of the 2003 world cup to demonstrate the strong and consistent performance batsmen of the Indian and Australian team. A few record of Indian and Australian batsmen is shown in below table.

PLAYER	MATCH	INNS	RUNS	AVG	SR	4s	6s	Rank
Sachin	11	11	673	61.18	89.26	75	4	1
Sourav	11	11	465	58.12	82.30	30	15	2
Rickey point	11	10	415	51.88	105.43	56	7	3
Adam Gilchrist	11	10	410	40.80	100.79	52	10	4
Mathew hayden	11	10	328	32.8	80	37	5	8

Table -1. Indian and Austrian batsmen 2003 world cup records[18]

S. R. Iyer and R. Sharda[3] "Prediction of athletes performance using neural networks: An application in cricket team selection," In this paper they employ the use of advance non-linear modeling technique called neural network approach to rating of the player. which is later used to predict the performance of the players in the future based upon on their past performance where they classify the batsmen and bowler in three separate categories "performer" ,"moderate" and "failure" with the aids of experts of cricket.

They show how these heuristic rating will be used for selection of the batsmen in world cup team. For the selection of a batsmen should received 1 or 2 "performer" or moderate rating but did not receive " failure" rating . the selection criteria for the bowler is much like similar to batting ,the bowler who is having "moderate" and "performer" rating and not receive any failure rating be selected in the team.

I.P.Wickramasinghe[4] "Predicting the performance of batsmen in test cricket," in this paper defined how they Predict the performance of the batsmen in the test series using longitudinal and hierarchical linear methods. In this paper they collect the sample data of test cricket batsmen who played during period of 2006-2010 from nine international teams and shows that these collected sample data exhibit both the longitudinal and hierarchical structure of three level.

Lemmer, H. H[5] An Analysis of players' performances in the first cricket twenty 20 world cup series which is hosted in south Africa. In this paper they defined how the performance measure of the batsmen and bowler for the One day International match has been adapted for use in the first T20 world cup series. How these measure can then used to ranked the batsmen and bowler.

Lewis, A. J [6] defined towards fairer measure of player performance in the one day cricket. In this paper they show the use of well established methodology called "Duckworth/Lewis" to model an alternative measure to analyze the player performance in the cricket.

H Saikia and D Bhattacharjee, [7] "Is IPL Responsible for Cricketers' Performance in Twenty20 World Cup". In this article they defined how the IPL tournament responsible for T20 world cup. This paper show the comparison of the player performance of both the Indian players and Foreign players who plays the T20 matches. They also shows how the players performance has been changed when they plays in IPL matches and for their national teams.

Brooks, Bussies're , Jennion and Hunt[8]"sinister strategy successes at the cricket world cup". In this paper they shows the significance of the handedness of the player on their performance. According to their studies left handed batsmen played better than right handed batsmen in the world cup 2003.

Trawinski [19] "A fuzzy classification system for prediction of the results of the basketball game." In this paper they describes how the Fuzz classification system is used for the result prediction but this method is used to predict the result of basketball. It also show how the Weka tool has been used for the attribute selection process.

After studying the studies produced in these article and by considering important studies of their work we have to develop a prediction models to predict the performance of the player in a particular match.

The study defined by the S. Muthuswamy and S.S.Lam[1] in their paper they attempt to predict only Indian bowlers performance against 7 international squad and study done by them is limited for only Indian bowler and their study cannot be exercise to predict the foreign bowlers.

Major work done in our project is to build prediction models that can be used to predict the performance and accuracy of any player in a given match using some supervised machine learning algorithm. In this model they rating of players has been done using different attributes of bowling and batting such as consistency, current form, form against opponent and venue the location where the match being player. This rating ranges from 1 to 5 for both batting and batting. We used weka tool for the selection of these listed attributes.

3. PROBLEM DEFINITION

Predicting outcome of the game has recognized some fundamental problem. In the existing method a person need to note down the run scored by batsmen and wickets taken by bowler which is manual process and take lot of times . this project aim at developing an online application for accuracy prediction in the game of cricket using machine learning.

4. NEED OF SYSTEM

To predict the performance of the player which is used as a basis for the selection of the players for the team. prediction model is constructed by consideration of major factor such as wickets claimed by the bowlers and number of runs scored by batsmen to select the best player for each match.

5. PROPOSED SYSTEM

We have proposed a system which overcomes the major weakness of manual work which is time consuming and required man power to manually maintain the records and statistics of each player. In this proposed system user will be provide with an interface for the coach and captain using which they will easily predict the accuracy of batsmen as well as bowler.

5.1 ARCHITECTURE DIAGRAM

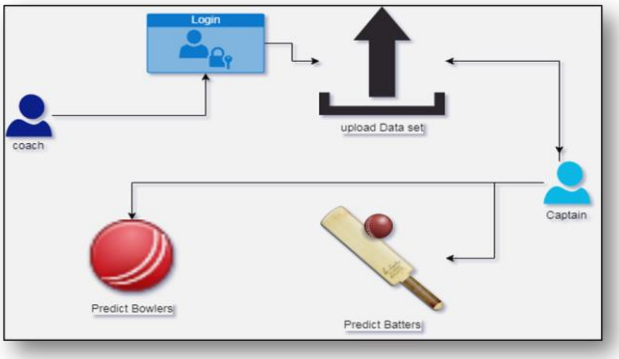


Figure 1. Architecture Diagram

6. DATA AND TOOLS

We accumulate all the information and data with use of the web scraping tools which is freely avaialable online, Parsehub[9] and import.io[10] data scraper which scrap the required data from the website https://www.espncricinfo.com/ or www.cricbuzz.com. We will import these in MySQL tables soon after collection of data and PHP is used to perform operation on these composed data. Further more we used the python 3.6 technology and Pycharm IDE software for the partical implementation of this project. MySQL is "Rational Database Management System"(RDMS) is the keypart of LAMP(Linux,Apache,MySQL,Perl &PHP).The 'Weka' and 'Dataiku' tool having the collection of the machine learning algorithms.Weka is a tool which offer some sort of the preprocessing functioning of the data and mining of the data,visualization ,regression,association of rules and clustering. We used these tools for prediction of performance and accuracy based on collected data. Dataiku is software platform for the use in collaborative data science which provide a way for the scientists, data analyst,and engineer to investigate , construct, distribute and prototype their own product.

7.LEARNING ALGORITHM

We used supervisod machine learning algorithm to generate prediction models. We used Decision Trees, Naive Bayes,Random Forest,Mutlclass Support Vector machine for our research. These algorithm explained in brief as follow.

7.1 NAIVE BAYES

The probabistic algorithm based on bayes theorem is Navie bayes algorithm .Based on prior probability, the probbabilty of observing various data given the

hypothesis and probabilty of observed data Bayes theorem offer a method to compute the probability of the hypothesis [17].

Bayes Theorem: Bayes theorem is keystone of Bayesian learning method because its provide a way to calculate the posterior probability P(h|D) from the prior probability p(h) and P(D|h) tog.ether with P(D)

$$P(h|D)=\frac{P(D|h)P(h)}{P(D)}$$

where

P(h|D) is the posterior probability of h, replicate the confidence that h hold after D has been Observed.

- P(D) is the prior probability of D.
- P(D|h) is the probability of observing D given a world in which the h holds.
- P(h) is the prior probability of Hypothesis h. which reflects any background knowledge about the chance that the h is correct.

7.2 DECISION TREE

Decision tree is technique to estimate the values of Discrete value target function, in which the learned function is pictorial represented as a decision tree. Each node in the tree specifies a test of some attribute of the instances and each branch descending from that node match up to the one of the possible values for that attribute. An instance is classified by starting at the topmost node(root) of the tree, testing the attribute specified by this node then moving down the tree branch corresponding to the value of the attribute in the tree[17]. “ID3” algorithm introduce by “Ross Quinlan 1986”. ID3 algorithm expertise to learning Boolean valued functions. It is greedy approach that cultivate the trees from the top-down and at each node selection of attribute is done which best classifies the local training examples. Its successor called “C4.5” (Quinlan 1993).Unlike ID3, “C4.5” apply for both continuous and discrete attributes values.

$$E(S)=\sum_{i=1}^{c}-p_i\log_2 p_i$$

where; pi is the proportion of S belonging to class i. logarithm is the base 2 because entropy is a measure of the expected encoding length which is measured in bits.if the target attribute is taken on c possible values then the entropy can be as large as log2c. The Information Gain Gain(S,A) of an attribute A relatives to the collection of examples S is defined as

$$Gain(S,A)\equiv Entropy(S)-\sum_{v\in Values(A)}\frac{|S_v|}{|S|}Entropy(S_v)$$

Where Values(A) is the set of all possible values for the attribute A, and Sv is subset of S for which the attribute A has a value v.

7.3 RANDOM FOREST

Random Forest is a variety of supervised machine learning algorithm based as one of the learning method which is used for both the regression and classification Task. Ensemble learning is the sort of learning where the same or different variety of algorithm are unite as a structure of powerful prediction model. The random forest algorithm coalesce various algorithm together of the same type that is the multiple decision trees which outcome in a forest of trees hence it is called “Random Forest”. Usually random forest supply useful and more accurate outcome as compared to other classifier algorithm.

Definition-A random forest is a classifier which consists of a set of tree structured as classifiers {h(**x**,Θk), k=1, ...} where the {Θk} are independent identical random vectors and each tree play a part to cast its vote for the most admired class at input **x** .

As defined in the above definition Random Forest is collection of decision trees. This algorithm produces certain amount of decision trees by creating a forest. An example of advance the accuracy by integrating the power of multiple classifier[15] all classifier is a decision tree and this combined classifier is called a “decision forest”.

T.K Ho[11] defined in his paper how he started a method using random subspace for the production of random forests.

Leo Breiman [12] used this method to broaden the algorithm afterward this method was certified as Random Forests. The technique which is apply to cultivate the trees of the random forest which is label as “Classification And Regression Trees (CART)”.

7.4 SUPPORT VECTOR MACHINE

Support vector is supervised machine learning algorithm exercise for both the classification, regression and further learning challenge. The concept of support vector machine is proposed by Vapnik , Vladimir Isabell Guyon and Bernhard boser [14]. LIBSVM [16] is a library used for Support Vector Machines (SVMs). Use of this library involves two major steps

- To acquire the model it will train the Datasets.
- It predicts the information of the testing dataset using this acquired model.

SVMs can be used to choose for both classification and numerical prediction. SVM use a non-linear mapping approach to replace the new data into an advanced dimension. SVM detach the class of tuples from one another after that it will look for a linear optimum hyper-plane in this novel dimension. This algorithm exploit the use of support vectors to finds hyper-plane and their boundary. Compact explanations about the learned prediction model is provided by support vector which is found by this SVM algorithm. A hyper -plane can be written in form of

$$W \cdot X + b = 0$$

where 'W' is a set weight vector represented by W = {w1, w2. , wn}, n represent the integer of attributes and 'b' is a scalar also called "bias".

8. IMPLEMENTATION

Consideration of the collected data from literature survey the problem is taken into account and practically implemented as shown in below screen.

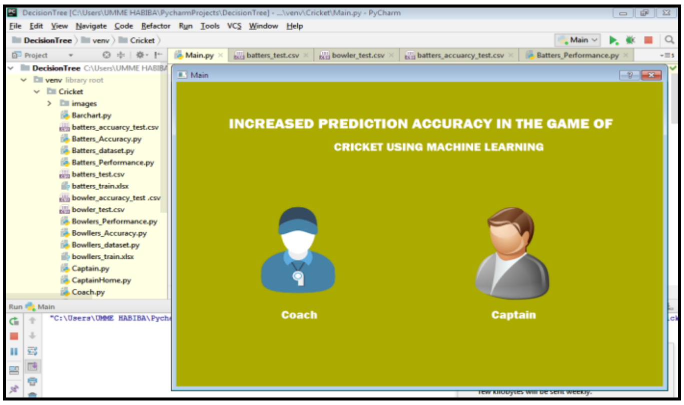
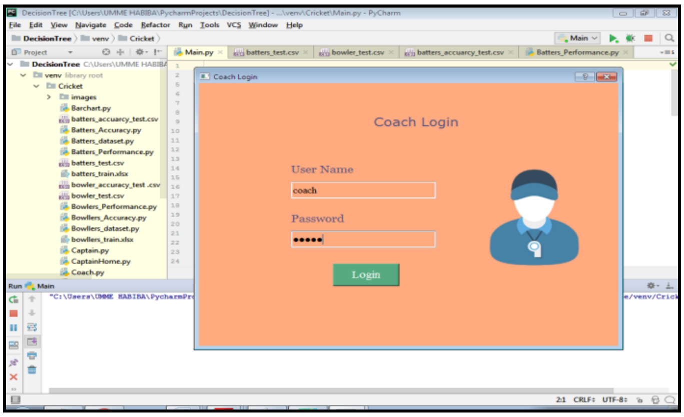

INCREASED PREDICTION ACCURACY IN THE GAME OF
CRICKET USING MACHINE LEARNING
Coach
Captain

Figure 2. Home Page

As shown in above screen the captain and coach provided with an interface to login with their own credential. As shown in below screen Coach uses his username and password to login.



User Name: coach
Password: •••••

Figure 3. Coach Login Page

After login the coach is provided with an interface where he can upload and view the batter and bowler Dataset.

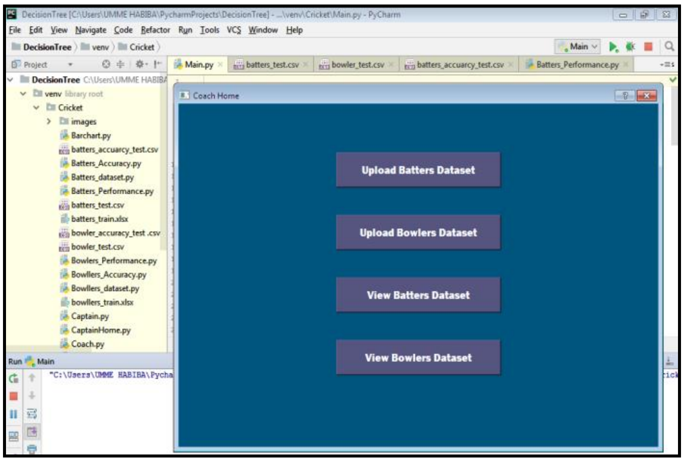
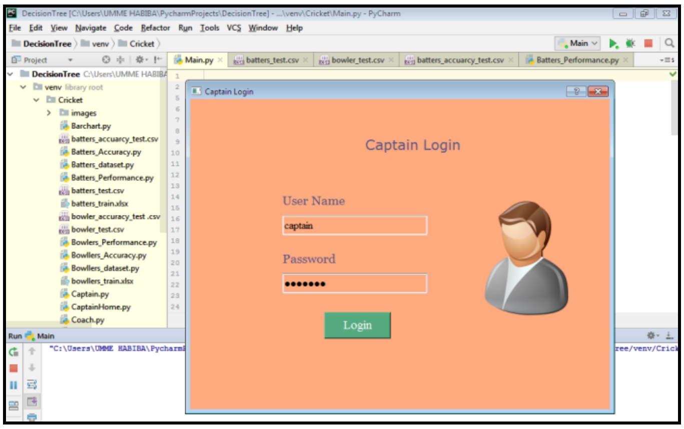


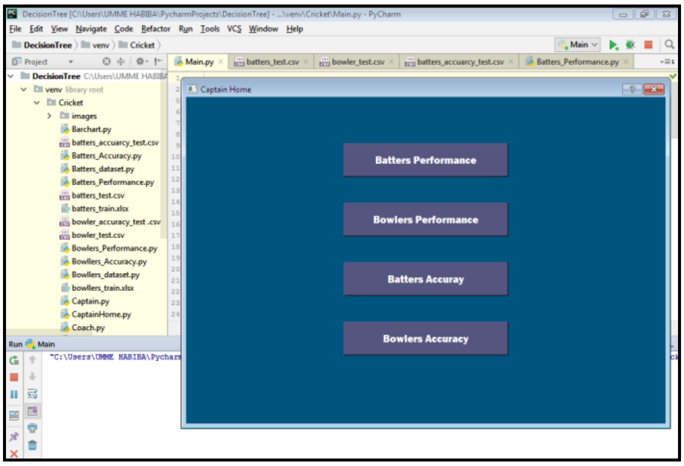
Figure 4. Coach home Page



User Name: captain
Password: •••••••

Figure 5. Coach Login Page

As shown in above screen the coach login with his own login credential by entering username and password. After login the captain is provided with an interface where he can predict the performance and accuracy of the batsmen and bowler.



```
"C:\Users\UMME HABIBA\Pycharm
```

Figure 6. Cpatain home Page

9. CONCLUSION AND FUTURE WORK

Players selection posses a vital role in the team's triumph. The selection committee board member ,coach and captain of team is responsible for selection of the best players for team for each match. The player's performances depend on various factors such as the location where the match being play, past records, his current form, average rate, strike rate, run scored at a particular venue, number of inning played against the opposition teams etc. Taking into consideration these information they employ an accurate prediction model which predict the accuracy of the batsmen and bowlers. In this project we modeled datasets based on players earlier record. Decision Tree ,Naïve Bayes, Random Forest and support Vector Machine supervising machine learning algorithm were evaluated and used. Random forest algorithm found to be produce more accurate and useful outcome among the other classifier algorithm. Whereas the SVM produce unexpected and less useful result.

This model work well with further format of cricket i.e. "T20 matches" and "Test series matches" and equivalent procedure can be applied these 2 format of game. But while considering the format T20 match here the match is limited for only 20 for each team so here the main job of the batsmen is score maximum run in less number of ball and bowler must have advanced wicket skill by yielding less run. And we can apply the same procedure in test matches where the batsmen need to have longer staying power as well as capable for playing longer innings and bowler need to have persuasive wicket taking skill in test matches.

ACKNOWLEDGEMENT

We would like to thank our college Principal, *Dr. Syed Zakir Ali,* Head of the Department, *Dr. Syed Naimatullah Hussain*, our co-ordinator Asst. Prof. Shruti Jalapur and my guide Asst. Prof *Mujamil Dakhani* for their valuable advice and technical assistance.

REFERENCES

[1] S. Muthuswamy and S. S. Lam, "Bowler Performance Prediction for One-day International Cricket Using Neural Networks," in Industrial Engineering Research Conference, 2008.

[2] G. D. I. Barr and B. S. Kantor, "A Criterion for Comparing and Selecting Batsmen in Limited Overs Cricket," Operational Research Society, vol. 55, no. 12, pp. 1266-1274, December 2004.

[3] S.. R. Iyer and R. Sharda, "Prediction of athletes performance using neural networks: An application in cricket team selection," Expert Systems with Applications, vol. 36, pp. 5510-5522, April 2009.

[4] I.P.Wickramasinghe, "Predicting the performance of batsmen in test cricket," Journal of Human Sport & Excercise, vol. 9, no. 4, pp. 744-751, May 2014.

[5] Lemmer, H. H. (2008). Analysis of players' performances in the first cricket twenty 20 world cup series. South African Journal for Research in Sport, 30(2), pp.71-77.

[6]Lewis, A. J. (2005). Towards fairer measures of player performance in one-day cricket. Journal of the Operational Research Society, 56, pp.804-815.

[7] Saikia , H., Bhattacharjee, D., & Bhattacharjee, A. (2012). Is IPL Responsible for Cricketers Performance in Twenty20 World Cup? International Journal of Sports Science and Engineering, *6*(2), pp.96-110.

[8]Brooks, R. , Bussie`re, L. F., Jennions, M. D., & Hunt, J. (2003). Sinister strategies succeed at the cricket World Cup. Proceedings of the Royal Society.

[9] "Free Download web scraping tool -web scraper | ParseHub," parsehub,[Online].Available: https://www.parsehub.com.

[10] "Data extracted from the web," Import.io, [Online].Available : https://www.import.io.

[11] Tim Kam Ho, "The Random Subspace Method for Constructing Decision Forests," IEEE transactions on pattern analysis and machine intelligence, vol. 20, no. 8, pp. 832-844, August 1998.

[12] Leo. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[13] Leo Breiman, J. Friedman, C. J. Stone and R. A. Olshen, Classification and regression trees, CRC Press, 1984.

[14] B.E. Boser, I. M. Guyon and V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," in Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, 1992.

[15]T.K. Ho, J.J. Hull, and S.N. Srihari, "Decision Combination in Multiple Classifier Systems," *IEEE* Trans. Pattern Analysis and Machine Intelligence, vol. 16, no. 1, pp. 66-75, Jan. 1994.

[16]Chang, C.-C. and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 3, Article 27 (April 2011),

[17] Tom M. Mitchell - Machine Learning-McGraw-Hill (1997).

[18] ICC world cup 2003 statistics: https://www.cricbuzz.com/cricket-series/796/icc-world-cup-2003/stats

[19] Trawinski Krzysztof . "A fuzzy classification system for prediction of the results of the basketball game." Fuzzy system(FUZZ) 2010. IEEE International conference on IEEE 2010s.