**FIFA Player Performance Prediction using Machine Learning Techniques**

**Abhilash Haridasan Premlata (x18135714), Smit Jain (x18135240), Monika Budukiewicz (x18134076), Saptarishi Das (x181127355)**

Advanced Data Mining

MSc Data Analytics

National College of Ireland

***Abstract- Player Performance is a very widely discussed topic when it comes to any sport. People have their own choices in terms of liking and disliking a player. But the fact remains that a player performance is usually dependant on their performance in every match and how well the player utilizes the skills it has. Performance is not a one match fairy-tale, the player has to perform consistently in order to be a top player. In this research we calculate the player performance based on the individual player statistics from last year. Machine learning algorithms like Random Forest Regressor, Decision Tree Regressor, Gradient Boosting Machine, Extreme Gradient Boost and Linear Support Vector Regressor are implemented and it is found out that Gradient Boosting Machine was the best predictor of player performance with an R Squared value of 0.8788 and the least was for Linear Support Vector Regressor with R Squared value of 0.7074. This meant that these models were able to predict player performance with an accuracy of 87.88% and 70.74% respectively. Extreme Gradient Boost had the least mean squared error of 2.39. The research showed that Extreme Gradient Boost was the most efficient algorithm among the other mentioned algorithms.***

***Keywords: FIFA, SVR, XGB, GBM, DTR, RF, Performance Prediction.***

1. INTRODUCTION

Football is the biggest played sport in the world. With millions of people following the sport, there is also a huge fan following of different players. This heavily depends on the performance of the player. It is said that majority of the big and famous players, play for clubs with a lot of money. The main objective of this research is to predict the player performance using different features like the player’s agility, skill and to an extent the age as well. The inclusion of extraction of tweets and data from various internet sources is not done because generally, the common people tend to show more support and have more admiration for the forward line and attacking players. The defensive players and the shot stoppers generally do not tend to get enough support from the fans or the viewers of the game. So, extracting tweets and data from other online sources, all the admiration and votes would be for the forward and attacking players.

A close up of text on a white background

Description automatically generated

**Figure 1 Player rating for FIFA 2019**

There have been various rules and regulations set up by the governing body of football on the money involved in the transfer market. A club is not supposed to do a transfer which violates those rules, if in case a club does so, the club would be banned from signing players in the next transfer market. This in turn helps in calculating the net worth for every player and their net value post their tenure completion with the club. The performance value of the player could also help the clubs in paying the transfer fee for the player, just in case the player wishes to move to some other club before the completion of tenure period. This helps the club owners and training staff in building a winning team by choosing the right candidates. Some players are naturally gifted and skilled, while others perform because of hardcore training. The performance yet depends on how the player performs in a match regardless of the skills the player possesses. The machine learning techniques would do no bias whatsoever, a player is a good performer only if he/she performs in a match. The values of the individual skills of each player is based on how the player has performed last year. With this we can analyse how the player will be performing this year. This would depend on several attributes which include age as well. Machine learning algorithms will be applied on various attributes of these players and their performance value will be predicted.

1. ***Research Question***

*“How accurately can the FIFA player performance be predicted using machine learning techniques?”*

1. ***Objectives***

To predict the player performance values by having a look at the different attributes of the players. This will be done using the exploratory data analysis. A principal Component Analysis (PCA) will be carried out along with the Chi-Square Test which tells the robustness of the model.

To identify the best algorithm for the prediction of player performance. The implemented study will have a look at several data mining techniques and a comparative analysis will be done to predict the player performance.

1. DATASET DESCRIPTION

The dataset has been taken from Kaggle which consists of 18K + records of all the players listed in FIFA 2019. These records are of all the players from forward/attack position player, mid-fielders, defenders and goalkeepers. The goalkeepers have different kind of skillsets and they play only on one fixed position. The other players can play at different positions based on the formation the team is playing with. The players play with different potentials at different positions, which is given in the dataset. The link for the dataset is

<https://www.kaggle.com/karangadiya/fifa19>

The next section will do a critique on the related work done by different researchers on different machine learning algorithms and on the same topic, in a different sport. Section 3 will discuss the methodology used for the research. Section 4 will discuss the different models used and their results. Section 5 will discuss the different evaluation criteria used for evaluating the models and section 6 will conclude the research and also give the future work.

1. LITERATURE REVIEW

After reviewing the machine learning techniques popular in the area of sport analytics, we discovered some papers focus on analysing team performance at games and their results [1]. With the help of FRES approach, the authors [2] have predicted the results of FIFA matches and have used Bayes inference with the rule-based reasoning which is commonly used for future performance predictions. Similar approach have been used by the authors [3], and Bayesian Network model has been used for the prediction of the football match outcomes of the English Premier League considering subjective and objective based information. The result shows that with the help of subjective information, the model performance improved significantly. And in the research [4] Bayesian Hierarchical models has been implemented. The results of the research work carried out shows that the model has performed better that bivariate poisson structure, but the hierarchical model has one limitation that its predictions are all in a single batch.

The performance predictions can be done with the use of Artificial Neural Networks also, as implemented by the authors [5] where the performance of the models applied have been impressive. However, in another research [1] the use of Naive Bayes model has not been impressive and provided lower accuracy levels. The sport events characterised with a score that determines the winning team. The analysis of players’ traits was at second or third position in hierarchy of analysis, while the betting analysis gained most interest among academic authors. However physical traits are easy to measure, each sport characterise with weights to different skills and traits having more value. For example, analysing efficiency of players’ performance in cricket proposed by [6] was achieved by using stochastic multicriteria acceptability analysis, that provides more information to separate the players, while data envelopment analysis was used to measure players’ efficiency at one season. Both these techniques are best for limited-over format of cricket game.

For purpose of players’ evaluation, one cannot underestimate structure of data being analysed. [7] managed to “develop a sports player evaluation model that utilizes both structured quantitative data (game statistics) and un- structured qualitative data (news articles) based on a deep neural network (DNN).” Text based data could lead to sentiment analysis, as presented by [7], machine learning techniques such as Naïve Bayesian classifier (NB) or support vector machine (SVM) have been used to build a sentiment classifier for a given collection of text. For performance analysis multivariate linear regression shows the best results. There have been numerous researches done in this field with the use of sentimental analysis as most of the fans post on social media after the team they support wins or lose. Such a study has been done by [8] where the changes in the emotions have been examined using sentiment analysis whenever a goal is scored. After analysing, it has been established that the most negative sentiments have been posted by the fans when the opponents scored a goal. On the other hand, it has been found that joy and anticipation are posted more often while a match is going on. The limitations of this research is that the tweets have been taken only based on English language and for a matter of fact, there are many nations who are not having English as their first language and the fans post tweets in other languages like Spanish, French, Portuguese, etc.

The same paper authors proved the regression trees [9] as more adaptive to select most interesting variables, were linear regression results suffered due to lack of variance.

The [10] suggest the regression analysis for players’ physical traits will be next after their paper focused on classification using Gaussian Processes for pattern recognition. As per this recommendation, our analysis will focus on analysing over seventy skills and physical traits. In this paper, the socio-economic attributes of the players play a major role in the player's selection in the NBA team. As per the research, the players from low-economic background are less likely to end up securing a position in the NBA team in comparison to the players from rich-economic background. The player's criminal record is also correlated to the player's status. The player's hometown details have also been taken into account.

[11] has discussed about the player performance and the player’s market value. Research has been constrained to players from the La Liga (Spanish Football League) only. The research shows the dependence of the performance of the player to their market value and on what factors does the performance of the player depend on. LASSO regression was used for the analysis purpose. The drawback of the research was that the performance prediction was restricted only to the forward players. The study concluded that the performance of the players and direct relation to their net worth in the market.

[12] did a research on the performance characteristics according to playing position in elite soccer. A computerized match analysis system was used to monitor the top-most players on the ground. The study concluded that the distance covered by the midfield players was the most when compared to players in other positions. The study provided the findings with respect to the demands placed on every player’s shoulder. This help in helping every player grow individually by organizing performance-oriented training programs.

[13] did a research on the prediction of National Basketball Association matches with the help of regression machine learning techniques. The study not only focuses on the match winner, but also the different reasons behind the match winning performance of the team. The player performance will also be classified to see which player contributed the most towards the victory and the position in which the player plays in. The study concluded that the ML techniques were more successful in the analysis that the traditional vote system. The multilayer perceptron did the prediction with an accuracy of 68.44% followed by 67.89% of linear regression and 66.81% of likelihood method.

[14] studied the performance of tile based and task-based techniques in predicting the player performance. A tile-based approach was established which used to trace the path or distance travelled by every player and invented the game analytics. The study concluded that the newly discovered tile-based approach for tracing the size of the grids is equally helpful in the game -analytics as the task-based techniques.

[15] studied the prediction of football league matches and predicting the long-term league winners using smart data. Using smart data, no large datasets are required to make predictions. The stud says that the machine learning techniques provides obvious benefits, those benefits always come at a cost of accuracy. The researcher focuses on smart data and knowledge engineering approach which emphasise on the real-world facts and casual knowledge. The study concluded that even though the study is not as accurate as the machine learning techniques, it provides a long shot bias in bookies odds of match predictions.

[16] provided an analysis of different machine learning techniques used for the result prediction in different sports and did comparative analysis on them to have a look at the advantages and disadvantages of every data mining techniques. Various data mining techniques such as Artificial Neural Network (ANN), decision trees, Bayesian method, logistic regression, Support Vector Machine (SVM), and fuzzy methods were used for predicting the result of different sports matches. The studied did have a future scope as the predictions made were not accurate enough, but out of the lot, Artificial Neural Network (ANN) proved out to be the best performer.

[17] studied the different features and did an exploration of different features in a soccer match. It covers both single and multi-players and also event based analytical views. The most important features are selected from the data and are processed semi automatically, visualized henceforth. The pipeline in the visual analytics also helps in improving the results of the classification. The study concluded that the feature driven approach was a data driven approach and helped in performing analysis on the soccer data.

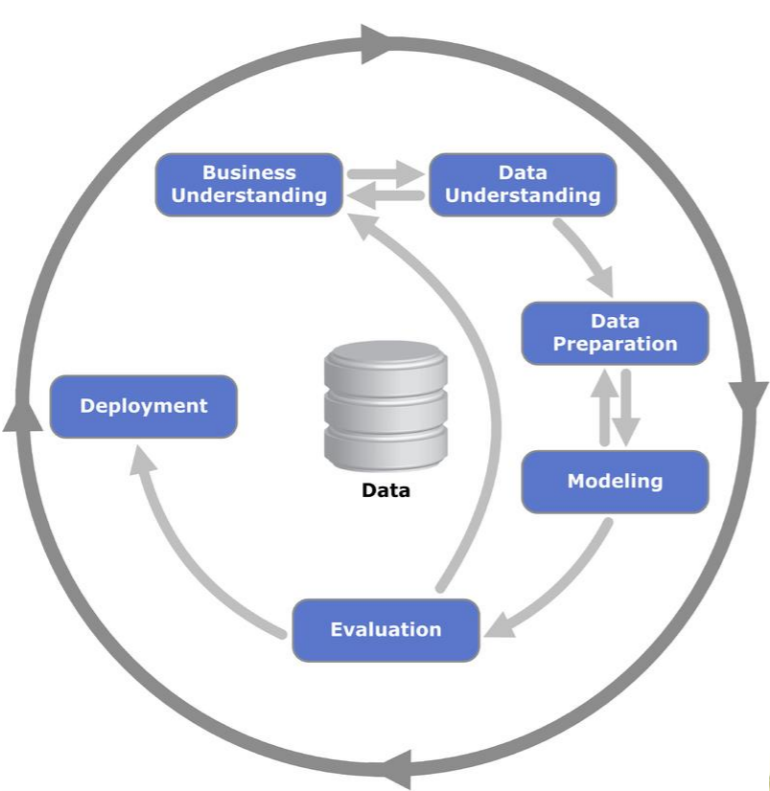
[18] studied the efficiency of machine learning techniques in rightly predicting the outcome of a soccer game with the help of easily and most freely available match data. The basic aim of the study was to to develop a predictive machine learning model to predict the accuracy of results of matches commencing from 31st March 2017. The study concluded results with the prediction of the teams in rank-wise order.

[19] studied the effect of heart rate variability to predict the game performance in Australian football league. Advanced machine learning regression techniques were used to do the prediction. The technique of variable selection with a wrapper subset gave the best results when compared with principal component analysis. The study also found that the there is a possibility of a training programme to be adjusted specifically and also the temperature which is likely to be on the day of the match.

[20] studied the semi automation of game play analysis using machine learning. The research tends to help ease the work of designers for the analysis of computer gameplay. Semi-Automated Gameplay Analysis by machine learning was used in a game by EA sports named FIFA 99 in which active machine learning was used. Similarly, this active machine learning concept was also introduced to FIFA 04 and turned out to be very helpful in designing and testing purpose.

1. METHODOLOGY

For carrying out the research work smoothly and in a planned way, we need to have a methodology to follow. There are 3 majorly used methodologies – CRISP-DM, KDD and SEMMA. These are most widely used when it comes to implementing data mining research. Based on the results and recommendations provided in previous researches [21] in data mining related fields, CRISP-DM has been selected to go ahead with. The steps followed in this methodology have been explained in brief below.



**Figure 2 CRISP-DM**

**Business Understanding**

The first and the most important step in this methodology is the business understanding of the research and how it can be of use to any field and what it actually contributes in. As explained earlier in the introduction section, The player performance is highly important to do club transfer business and also to understand if the player’s demands for increased wage is needed or not. The need of research in this field can also be explained through the literature review of the previous researches done in this field and since there is no such work done in this field, this brings novelty in the piece of work being done in this research.

**Data Understanding:**

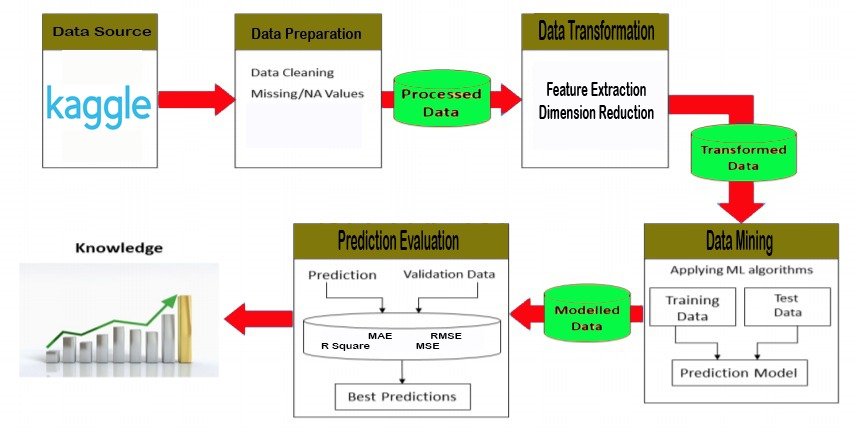
There are various factors which affect the performance of any player on the field, like, speed, dribbling, passing, shooting and the list goes long. To predict the performance, we need to have a good understanding on the factors that impact the performance, which is done through the literature review.

**Data preparation:**

The dataset chosen for this research is having details of above 18K players from FIFA. This includes goalkeepers and all the other position players. In this way we can predict the performance of the players better. Before moving on to the modelling stage, we still need to convert all the factors into integers so that the models can be applied to the data set and clean the data as there can be some missing values in the data set.

**Modelling**

For prediction of the player’s overall performance, we can use different models. In this project, we have used different modelling techniques which are – Linear Regression, Decision Tree, Random Forest, Linear SVR, Gradient Boosting and XG Boosting. The results obtained after running the models are then compared and evaluated which is given in the section – implementation methods in brief.



**Figure 3 Implementation Architecture**

**Evaluation**

The evaluation of the results of the models implemented will be carried out using different techniques available. The most commonly used techniques are - Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) and R squared value. A more detailed explanation is briefed in the implementation section.

**Deployment**

The last stage of this methodology is the deployment where the reporting and the documentation of the research work, carried out in the earlier stages, is done. This is important as the final outcome of the research is displayed here which is the proof of the research results as well.

**Preliminary level of cleaning – White noise cleaning**

The dataset has been first divided into two parts based on player positions – goalkeepers and others. Next the columns which are not required at all have been removed one by one. There are details like name, nationality, club name, flag, etc which are not contributing anything towards player’s performance, that is why, these columns are being removed.

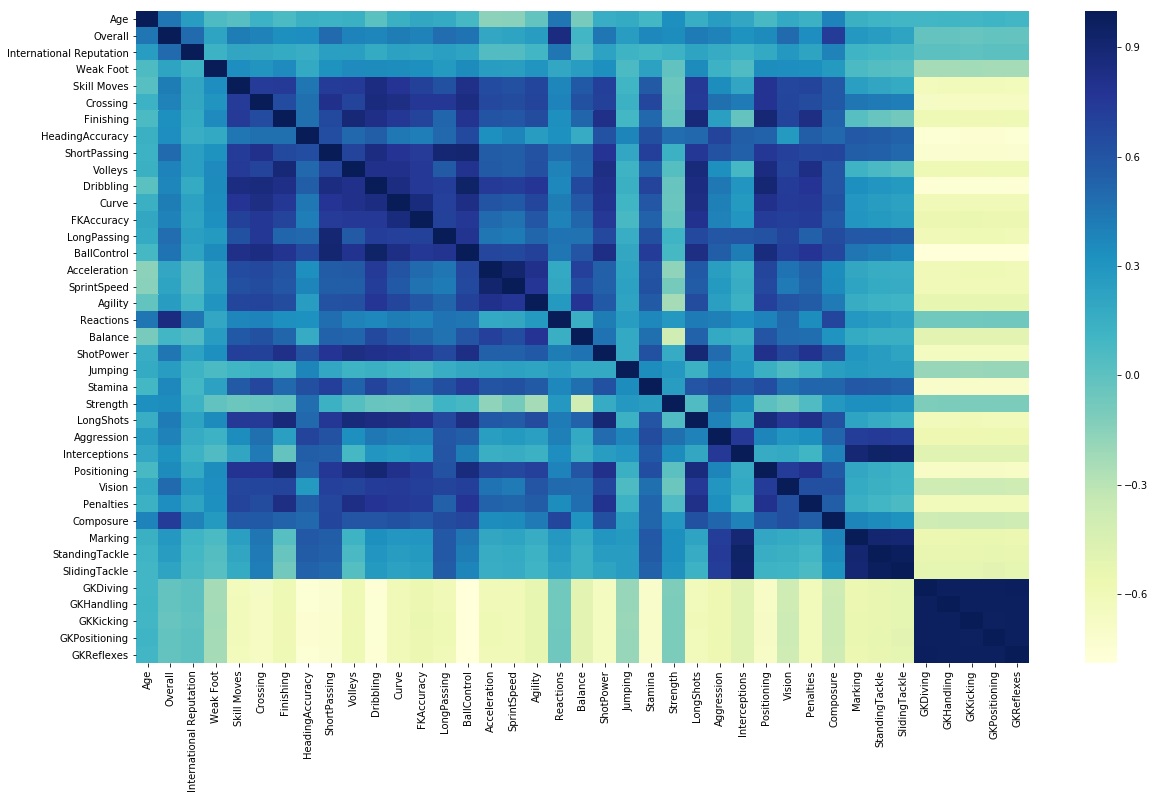
**Transformation**

The columns with the player value, wage and release clause are having values in millions and thousands represented by letters ‘M’ and ‘K’. For applying models, we need to convert these values into numerical numbers. Hence, these columns are also cleaned. After running a check, null values are found which are also removed with the help of R tool. These steps for cleaning the data are applied to both, the goalkeeper’s dataset and the other player’s dataset.

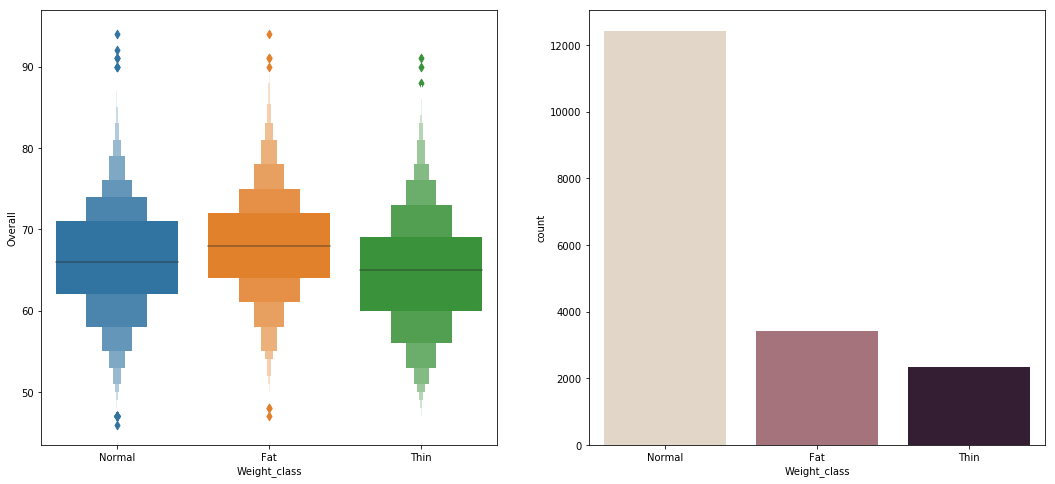
**Exploratory Data Analysis**

The initial level of data analysis is required for the better understanding of the dataset. In this project we have carried out a few analyses like correlation matrix, Principal Component Analysis and have categorized the number of players as per the weight and heights. The correlation matrix is given below:

The total players available in the dataset have been divided into three parts based on their weights. The players having weight lesser than 67 have been categorized as thin, the players with weight between 67 and 81 are categorized as normal and players with weight more than 81 are categorized as fat. Next we plot the categorized data on a graph and get to know the total number of players who are fat.



**Figure 4 Correlation Matrix**

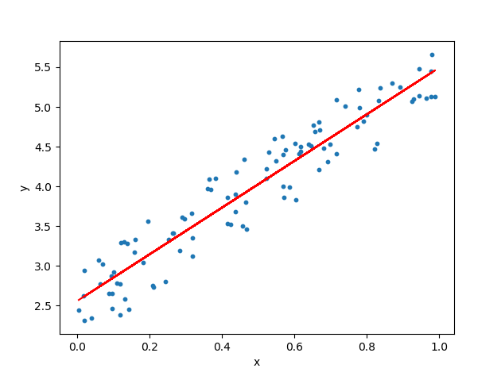


**Figure 5 categorizing the players by weight**

1. Models Implementation

**Linear Regression**

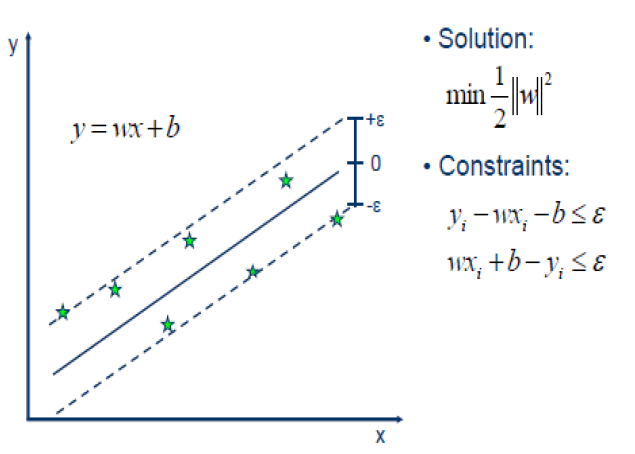
Regression in general is a model which is used to predict a value, so regression generally means prediction. It is used to predict a future value based on a set of values from the past. It tries to establish a relationship between the values of both dependent and independent variables to predict a value for the dependent variable. This regression can be of two types, a simple linear regression and a multiple linear regression which involve one and more than one independent variable respectively.



**Figure 6 Linear Regression**

**Support Vector Regressor**

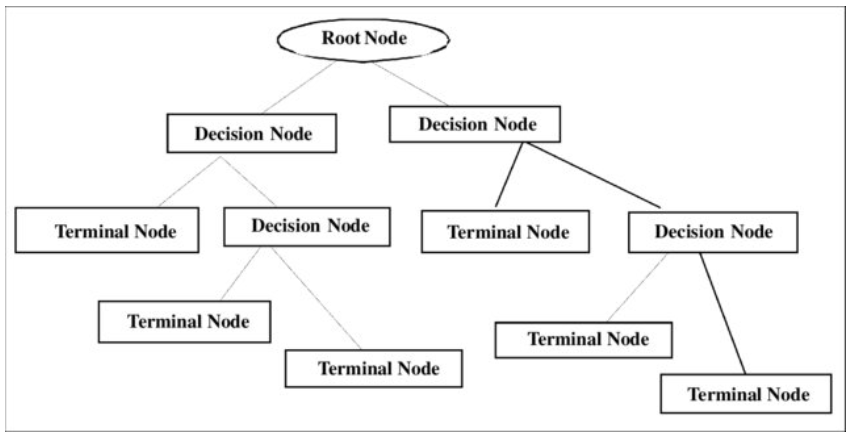
It is the regression model for support vector machine. It has a lot of possibilities, and hence output is a real number, a margin of tolerance is set and the main aim is to find a value whose tolerance deviation is the most, which means the value of the predicted vales is not close to the actual values. It forms a hyperplane in which the independent variables are all close to the line.



**Figure 7 Support Vector Regressor**

**Decision Tree Regressor**

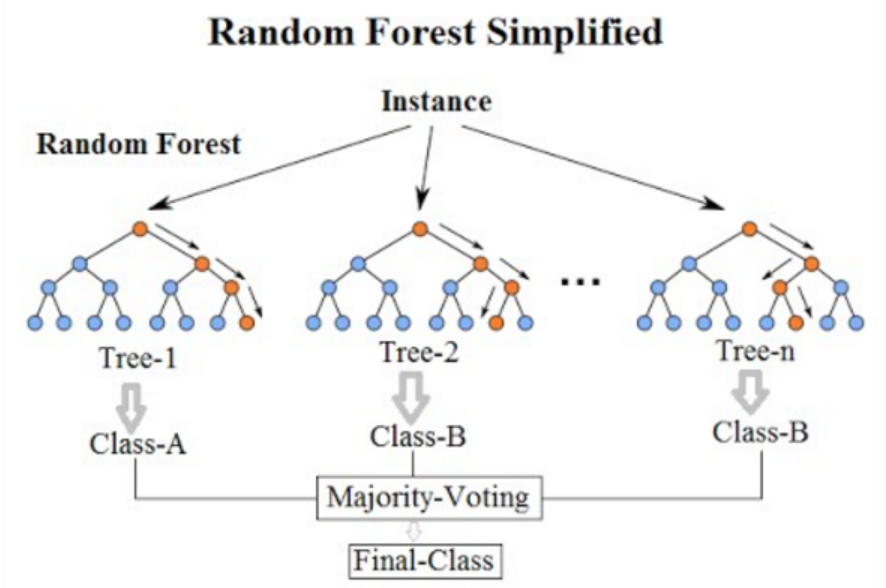
The Decision tree model can be used as both a regression and a classification model. The model which we intend to use is the decision tree regressor. The decision tree is segmented and broken down into smaller and smaller units, basically subsets and parallelly, a simplified decision tree is developed. The decision tree has two parts, a decision node and leaf node respectively. A leaf node is the final node whereas a decision node has multiple branches coming off it. A decision node is basically like an if else statement with multiple decisions. The final outcome comes off a root node.



**Figure 8 Decision Tree model**

**Random Forest**

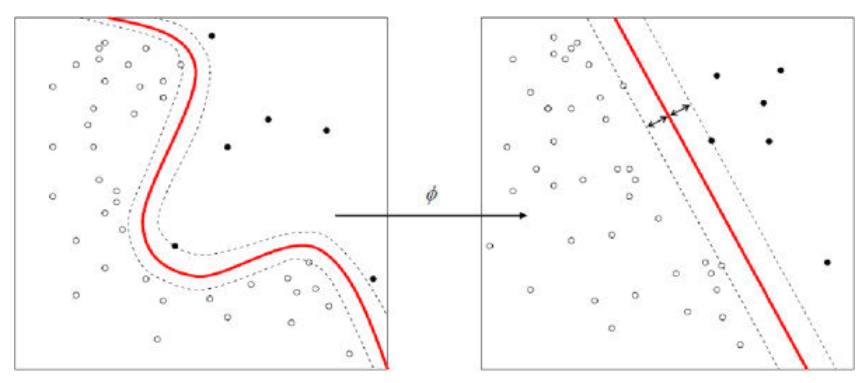
Random forest is an ensemble of multiple decision tress. It can be used as both a classification and a regression model. In this project we talk of random forest as a regressor. The output of multiple decision trees is combined, and the mean value is taken. This results in a better accuracy. The only disadvantage of a random forest regressor is that it results in overfitting, but it deals well with the missing data and also has a fast runtime.



**Figure 9 Random Forest**

**Gradient Boosting Machine**

Gradient boosting methods are optimised machine learning algorithms used in both regression and classification to get high accuracies. Advanced boosting algorithms use simple algorithms as their input and provide better results. One advantage of boosting algorithms is that they reduce the bias and variance in data thereby controlling overfitting which is a drawback of Random Forest.



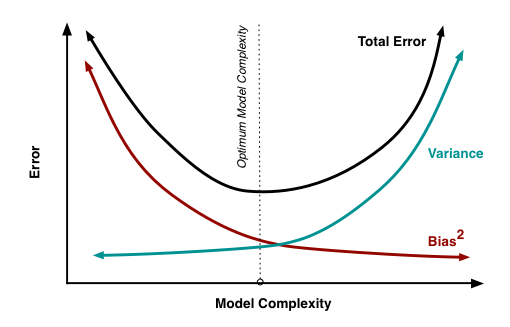
**Figure 10 GBM regression graph**

**Extreme Gradient Boost**

Extreme Gradient Boost is an advanced boosting method capable of predicting better results with less computation time. It is an ensemble of weak decision trees with parameters such as number of learners and learning rate. These parameters can be tuned to get even higher accuracies. The primary reason to go for Extreme Gradient Boost is that it has outperformed most regression models in recent researches.

**Least Operator Shrinker and Selection Operator**

Lasso otherwise known as least operator shrinker and selection operator is used to perform the selection of variables and the regularization. This is basically done to improve the accuracy of the prediction and also for the better interpretation of the statistical model which it makes. The basic aim is to minimize the prediction error.



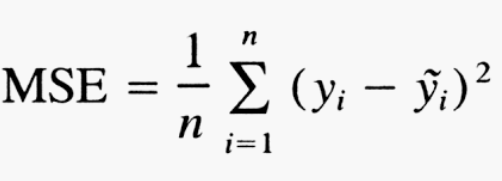
**Figure 11 LASSO model**

1. Evaluation Techniques

This is the final stage where the results will be evaluated for the forecasting models developed in the earlier stages. Most of the prediction models are evaluated based on the 4 basic criteria which are- Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and R-squared Value. A comparative analysis will be done for the models developed in the earlier stages and results will be formulated.

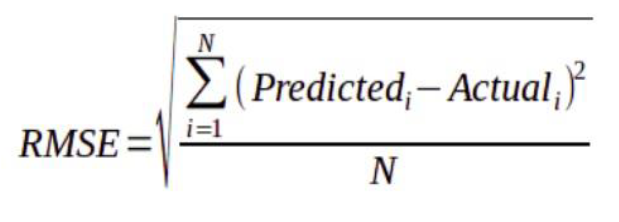
**Mean Squared Error**

The MSE is a quantity of measure which estimates the quality of the predictor. The MSE is always positive and the closer its value is to zero, the better is the performance. It calculates the variance of the value that is estimated from the sample of the data set. This way we can check the performance of the model evaluated.



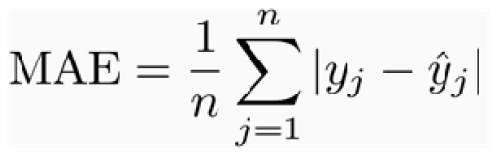
**Root Mean Squared Error**

The standard deviation of prediction errors (residuals) from regression line is known as RMSE. It represents the residual spread. It also represents the density of the data points near the line of best fit.

It is widely used in forecasting for the verification of the results of experiments performed.

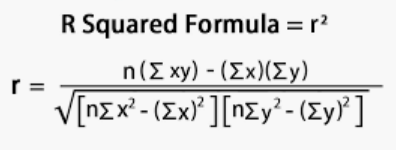
**Mean Absolute Error**

The MAE is quite similar to RMSE. The only difference is that in MAE, absolute value of the difference is taken into consideration instead of square of the difference between predicted value and actual value.



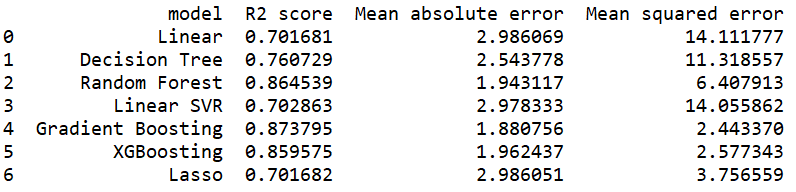
**R-squared Value**

R-squared value is the measure of fit which describes the variance proportion between the dependent variable and the independent variable. It is considered to be opposite of correlation. The measure of variance of a variable from another variable is explained by R-squared value. It ranges between 0 to 1. The higher the value of R-squared, the better the variance of the variables are explained.



1. Results and Discussions

The dataset has been taken from Kaggle. The dataset has more than 18,000 records of the FIFA 2019 edition and is having more than 70 player attributes based on the player’s performance in the year 2018/19. The models proposed in the above sections were implemented and their results are discussed below:



**Figure 12 Model performance comparison**

The above figure clearly tells us that the Gradient boosting model has performed the best with the least error and highest R-squared value. The Random forest model and Xtreme Gradient Boosting models are marginally behind the gradient boosting. Although, the results produced by Gradient boosting are the best, but the training time taken by it is the maximum and Xtreme Gradient boost and Random forest have taken lesser time for the prediction. The models that have performed the least are the linear regression and the Linear SVR model. The MSE for both the linear models are very high and the mean absolute error is also significantly high.

1. Future Work

As part of future work, Machine learning can be implemented in the different football leagues where the awards are given mostly to the players who are top goal scorers and the other players who perform exceptionally well round the year, on other positions are shadowed, even if their efforts helped the team win the games. The voting system can be eradicated and instead, based on individual performances, the awards can be distributed. The Ballon d’or which is the award for the best European player will be given on the basis of individual statistics, which is how it is meant to be and not to recognized star name of a particular team. Also, the award would be given to a player regardless of the player’s playing position.

**Acknowledgement**

We would like to thank Dr Anu Sahni for her extended support throughout the semester. She encouraged and was always ready to help us in the completion of the project. We would also like to take an opportunity to thank our classmates and all the lab assistants in helping us complete our Advanced Data Mining Project.

**References**

[1] D. Miljkovi, L. Gaji, A. Kova, and Z. Konjovi, “The Use of Data Mining for Basketball Matches Outcomes Prediction,” pp. 309–312, 2010.

[2] B. Min, J. Kim, C. Choe, H. Eom, R. Ian, and B. Mckay, “A Compound Framework for Sports Prediction : The Case Study of Football,” no. February, 2007.

[3] A. C. Constantinou, N. E. Fenton, and M. Neil, “Knowledge-Based Systems pi-football : A Bayesian network model for forecasting Association Football match outcomes,” *Knowledge-Based Syst.*, vol. 36, pp. 322–339, 2012.

[4] G. Baio and M. Blangiardo, “Bayesian hierarchical model,” vol. 37, no. 2, pp. 253–264, 2010.

[5] A. Mccabe and J. Trevathan, “Artificial Intelligence in Sports Prediction,” pp. 1195–1198, 2008.

[6] T. Gweshe, “An analysis of the efficiency of player performance at the 2011 Cricket World Cup,” vol. 29, no. 2, pp. 137–153, 2013.

[7] Y. J. Park, H. S. Kim, D. Kim, H. Lee, and S. B. Kim, “Knowle dge-Base d Systems A deep learning-based sports player evaluation model based on game statistics and news articles,” *Knowledge-Based Syst.*, vol. 138, pp. 15–26, 2017.

[8] Y. Yu and X. Wang, “Computers in Human Behavior World Cup 2014 in the Twitter World : A big data analysis of sentiments in U . S . sports fans ’ tweets,” *Comput. Human Behav.*, vol. 48, pp. 392–400, 2015.

[9] P. Steenkiste, “Finding the Optimal Fantasy Football Team,” 2015.

[10] L. P. Lejon and F. Berntsson, “Regression Analysis on NBA Players Background and Performance using Gaussian Processes.”

[11] M. He, R. Cachucho, and A. Knobbe, “Football player ’ s performance and market value Football Player ’ s Performance and Market Value,” no. December, 2017.

[12] H. Tschan, “Performance Characteristics According to Playing Position in Elite Soccer,” no. March, 2007.

[13] I. Y. H. Hu, “Prediction of NBA games based on Machine Learning Methods,” 2013.

[14] C. Sebastian, Y. Sheng, and I. Li, “Computers in Human Behavior Predicting expert – novice performance as serious games analytics with objective-oriented and navigational action sequences,” *Comput. Human Behav.*, vol. 49, pp. 147–155, 2015.

[15] A. Constantinou and N. Fenton, “Knowle dge-Base d Systems Towards smart-data : Improving predictive accuracy in long-term football team performance,” *Knowledge-Based Syst.*, vol. 124, pp. 93–104, 2017.

[16] M. Haghighat, H. Rastegari, and N. Nourafza, “A Review of Data Mining Techniques for Result Prediction in Sports,” vol. 2, no. 5, pp. 7–12, 2013.

[17] H. Janetzko, D. Sacha, M. Stein, T. Schreck, D. A. Keim, and O. Deussen, “Feature-Driven Visual Analytics of Soccer Data,” *2014 IEEE Conf. Vis. Anal. Sci. Technol.*, pp. 13–22.

[18] W. Dubitzky, P. Lopes, J. Davis, and D. Berrar, “The Open International Soccer Database for machine learning,” *Mach. Learn.*, vol. 108, no. 1, pp. 9–28, 2019.

[19] K. Nesbitt and H. F. Jelinek, “Prediction of game performance in Australian football using heart rate variability measures Prediction of game performance in Australian football using heart rate variability measures David Cornforth \*, Piers Campbell and,” no. May, 2015.

[20] F. Southey, R. Holte, G. Xiao, M. Trommelen, and J. Buchanan, “Machine Learning for Semi-Automated Gameplay Analysis.”

[21] R. Wirth, “CRISP-DM : Towards a Standard Process Model for Data Mining,” no. 24959.

**Appendix**

The contribution of each team member is provided in the matrix below:

|  |  |  |
| --- | --- | --- |
| **Member** | **Tasks Assigned** | **Contribution** |
| Abhilash | Introduction | 26% |
| Research Question & Objectives |
| Literature Review |
| Dataset description |
| Model descriptions |
| Smit | Literature Review | 26% |
| Methodology |
| Data cleaning and pre-processing |
| Data trasnformation |
| Models Implementation |
| Evaluation techniques |
| Saptarishi | Implementation Architecture | 25% |
| Video presentation |
| Monica | Literature Review | 23% |