## horizontal line

Rating systems, clustering and classification of FIFA players

Marta Hausman

St. John’s University

Marta.hausman12@stjohns.edu

Zeyad Al Mothafar

St. John’s University

zeyad.al17@my.stjohns.edu



# 

# Abstract

The goal of this paper is to develop a predictive model for the rating and position of soccer players for FIFA, a series of association football video games released annually by Electronic Arts under the EA Sports label. Another objective of this research is to classify the players and predict their position and rating with the use of multi-class neural networks (MLP) and classification algorithms (KNN, SVM, and Naïve Bayes). Clustering of similar players is also performed with k-means, as well as linear regression. The study was developed using a complete FIFA 2017 player dataset containing 17,589 observations

(<https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global>).

# Introduction

Soccer is considered one of the most popular sports in the world. It’s been estimated, there are roughly

265 million players actively engaged in soccer and 13 million active FIFA players around the world. Though simple to watch, soccer is a very complex sport and basic statistics often does not yield significant results. Machine learning tools are necessary to gain more valuable insights about the players, clubs, etc. In this paper, we are going to apply regression as well as clustering and classification models to find interesting trends in the FIFA player dataset.

# Related work

With the increasing popularity of sport simulation video games, it is important to develop a reliable rating system reflecting the actual skills of players. The first video game rating system was developed by the Correspondence Chess League of America in the late 1930s and further advanced in the 1960s by Elo [1]. It ranks players based on the outcome of each game rather than individual skills. That was the biggest shortcoming of this rating system. As it was further proven, skills are more crucial in establishing an accurate rating. In our rating prediction, we therefore focus on individual player’s skills as provided by FIFA.

In his paper, Aleksi Visti developed a rating model to create more accurate rating systems for video game players combining skills and performance using TrueSkill algorithm that calculates the addition or loss of rating for the participating teams [1]. However due to high complexity of the games being played and rated these days, it is very difficult to find a universal system fitting different game structures. It is not applicable to soccer video games.

In their paper, Arnu Pretorius and Douglas A. Parry focused on predicting the outcomes of sporting events. They employed the Random Forest classification algorithm to predict the results of the 2015 Rugby World Cup. The training data consisted of historical team statistics and past games for three consecutive years. The classification model built in this study consisted of two classes (game won or lost). Forest-RI Model was able to correctly predict the outcome of the games with an accuracy of 89.58%. The biggest shortcoming of this model is that it didn't consider a draw to be a valid outcome of a match as it only worked for binary classes [2]. In our study, we used more simplistic Naïve Bayes rather than Random Forest due to its higher speed (less processing time).

Research presented at the 10th International Conference on Informatics and Systems [3] focused on evaluating the performance of football European teams participating in the Europa League or Champions League in the season of 2014/15. It intended to measure efficiency for decision making units (DMUs) with multiple inputs and outputs. It led to the conclusion that teams ranked high within their national leagues, perform better in international competitions, hence have the highest relative efficiency scores [3].

In the research presented by the Journal of Sport Economics [4], econometric frontier model was applied to analyze the performance of soccer clubs in the Premier League. The model used sport and financial variables as input to generate efficiency scores similarly to the study featured at the International Conference on Informatics and Systems [4]. The study concluded that the price of labor, players and stadium are the strongest, thus most predictive variables [4]. In our study, we similarly evaluated independent attributes to identify which ones have the highest explanatory power of ranking.

A 2017 Applied Data Science paper tackled the issue of performance comparison. The unstable performance of Leicester City has motivated a group of researchers to analyze the factors that contribute to the team’s success. The analysis was done using three classifiers: logistic regression, Random Forest, and Multilayer Perceptron. The final predictions were made through averaging the ensemble of three models [5]. The results were visualized via strategy plots. According to the model, Leicester's defense strategy is the key to success.

Predictive model weights can be used to get a data-driven player ranking [6]. In their model, Joel Brooks, Mathew Kerr and John Guttag, relied entirely on the value of passes completed which is derived based on the relationship of pass locations in a possession and shot opportunities generated. Although it’s a powerful model it doesn’t include any of the players features which may lead to a study or a model that is not fully informative and precise and that’s why we decided to include a big number of features that has high correlation with the rankings.

Previous studies and research used many different techniques to try and predict the performance and rating of soccer and basketball athletes; five-factor model of personality is one of the approaches that was used, all subjects completed a bipolar adjective scale designed to measure the five factors, Coaches ratings on several performance dimensions and actual game statistics were also collected [8].However this paper only used the coach’s rating and game statistics in their model, in our model we take a look into features that have really strong correlation with the performance that they didn’t mention like the reaction of the player.

The relationship between salary and the performance of the NBA teams was studied using multiple linear regression model, performance in game was defined in 4 terms: assists per game, rebounds per game, points per game and steals per game, four different regression tests were applied to study this relationship [9].Using the multiple linear regression model tells if the model is powerful or not and looking at this paper we can tell that the salary is not really related with athletes performance so we can just not include it in the model we provide.

Complex networks statistics were used to analyze the NBA database in order to create models to represent the behavior of team in the NBA, the results of the model were compared with box score statistics(scores, assists, rebound per game).Different models were used in order to predict team success based on complex network metrics such as clustering coefficient and node degree, the study shows that combining complex network metrics with box score statistics improves the prediction efficiency [7].The idea of combining two models to enhance the prediction is really solid and robust thing to do and that’s why we are building different models using neural networks, KNN, SVM for classification and KMEANS for clustering.

Another study was on Classification of passes in soccer using spatio temporal data; using the object tracking systems in the stadiums that generates high-resolution and high-frequency spatiotemporal trajectories of the players and the ball. The results of the classifier were compared to those made by human observers and there was an agreement between the classifier and the observer similar to that between to other observers. Assigning a single rating to each pass was the main purpose; however, it is apparent that the overall quality of a pass includes several factors: how well the player executed the pass, the difficulty of executing the pass given the situation, the riskiness of the pass [10].Comparing a model’s results with the human observers results and finding if there is an agreement between them informs that the model is powerful, we will be using this concept to make sure our model is powerful enough to predict accurately.

Multilayer Perceptron (MLP) neural network classifiers are commonly used to solve complex classification problems through minimizing the total sum of errors (TSE) caused by each training pattern [11]. It is especially efficient for multiclass classification with high misclassification cost and class imbalance problems. One of the challenges in our study is class imbalance – the number of instances in one class greatly overweighs the other class. For instance, class 1 in rating contains 18 very strong players, while class 3 has over 2,900 mediocre players. Similarly, with the position class – some positions are more unique – such as the goal keeper, hence the class is automatically smaller. MLP helps overcome the class imbalance problem as proven on the example of credit rating scoring systems [12]. Higher accuracy with MLP can be achieved through manipulating different inputs and numbers of hidden units.

# Methodology

This research is based on the full FIFA 2017 dataset, containing 17,589 player profiles, focused on individual player’s skills, such as: moves, ball control, dribbling, marking, sliding, tackling, aggression, reactions, attack, interception, composure, crossing, long pass, acceleration, speed, stamina, strength, balance, agility, jumping, heading, shot power, curve, free kick, volleys and goal keeper skills, ranked on a scale from 0 to 100. On top of the skill features, this study considers appearance features, such as weight, height as well as age. Descriptive statistics and correlations was used to describe the features above and relationships between them.

Further analysis includes Linear Regression, Classification (SVM, Neural Networks, KNN, Naïve Bayes) and Clustering (k-means).

## **Preprocessing**

The rating of the players was partitioned into 5 classes with the following boundaries:

Class 1: 89->94  
Class 2: 81->89  
Class 3: 73->81  
Class 4: 61->73  
Class 5: 45->61

Categorical features were converted to numbers. The position of the players and the preferred foot was converted into numeric form to be used in our analysis as the following:

Preferred foot:

0: Right foot, 1: Left foot

Positions:

0: SUB 1: GK 2: CB 3: RCB/LCB 4: CDM/RDM/LDM 5:CM/RCM/LCM

6: CAM 7:CF 8: RB/LB 9: RWB/LWB 10:RM/LM/RAM/LAM 11:RW/LW

12: RF/LF 13:RS/LS 14:ST 15: RES.

Feature selection:

Looking at the correlation between each independent feature and the rating of the player combined with the multicollinearity, we eliminated some of the features to get a more powerful and precise model. Feature selection was conducted based on t-statistic and correlation results. Not statistically significant variables were removed.

Looking at the t-statistic values, we removed the following features below the 3.0 cut-off value: Marking and curve. At the end of feature selection, we were left with 25 highlighted variables presented in Table 1.

Table 1. T-statistic results

|  |  |  |
| --- | --- | --- |
|  | coef | t-stat |
| Intercept | 3.6918 | 3.138 |
| Position | -0.0086 | -2.342 |
| Height | 0.0064 | 1.021 |
| Weight | 0.0203 | 4.181 |
| Preffered\_Foot | 0.2572 | 4.89 |
| Age | 0.0625 | 10.303 |
| Weak\_foot | 0.1354 | 3.794 |
| Skill\_Moves | 1.1429 | 23.283 |
| Ball\_Control | 0.1803 | 38.904 |
| Driblling | 0.0124 | 2.951 |
| Marking | 0.0127 | 3.095 |
| Sliding\_Tackle | -0.016 | -3.384 |
| Standing\_Tackle | 0.0457 | 9.273 |
| Aggression | 0.0055 | 2.397 |
| Reactions | 0.2775 | 72.703 |
| Attacking\_Position | -0.0434 | -13.784 |
| Interceptions | 0.0067 | 2.091 |
| Vision | -0.0037 | -1.267 |
| Composure | 0.0463 | 16.848 |
| Crossing | 0.0129 | 4.614 |
| Short\_Pass | 0.0757 | 15.613 |
| Long\_Pass | -0.0168 | -4.561 |
| Acceleration | 0.0311 | 7.227 |
| Speed | 0.0371 | 9.017 |
| Stamina | 0.0103 | 3.907 |
| Strength | 0.0374 | 12.827 |
| Balance | -0.0186 | -5.649 |
| Jumping | 0.0152 | 6.602 |
| Agility | 0.0026 | 0.843 |
| Heading | 0.102 | 34.843 |
| Shot\_Power | 0.0226 | 7.465 |
| Curve | 0.0113 | 3.783 |
| Long\_Shots | -0.0234 | -7.224 |
| Freekick\_Accuracy | -0.002 | -0.737 |
| Penalties | 0.0024 | 0.847 |
| Volleys | -0.0066 | -2.217 |
| Finishing | 0.025 | 7.237 |
| GK\_Positioning | 0.0954 | 15.834 |
| GK\_Diving | 0.0757 | 12.471 |
| GK\_Handling | 0.0968 | 15.955 |
| GK\_Reflexes | 0.0963 | 15.996 |

After this step we looked at the collinearity between the features and removed the ones that have high collinearity with each other.

For example,GK\_Positioning and GK\_Handling have a high correlation between each other, so the variable with lower correlation was removed. At the end of feature selection, we were left with 21 variables.

## **Classification**

Four classification models were used: SVM, KNN, Naïve Bayes and Neural Networks. For all these models “train\_test\_split” function and cross validation function used to cross-validate and split the data into testing and training with a test size of 0.33. All these models were used to classify the rating and position of the players and the classification of the position gave less accuracy because there are 16 position classes and 5 classes for the rating.

### 1. SVM

For this model Linear kernel was use; the model takes a training set of n points of the form:

(x1, y1), …, (xn, yn)

Where x is the point and y is the class to which the point belongs to. Each x is a p-dimensional real vector. The model tries to find a “maximum-margin hyperplane” that divides the group of points x for which y= 1, for example from the group of points where y=2, which is defined so that the distance between the hyperplane and the nearest point x from either group is maximized.

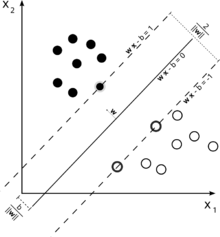


Fig.1 Maximum-margin hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors.

After this step the same concept is applied but, on the testing set where the class values are unknown which is either the position or the rating of the player and then we can get the accuracy, recall, f-1 score and precision and compare the results with other models.

### 2. KNN

Three different values for K-nearest Neighbors are used: 2, 3 and 4. The Euclidean distance is used in this model as a distance metric, there are training and testing sets for this model like the SVM model.

Fig.1: Euclidian Distance

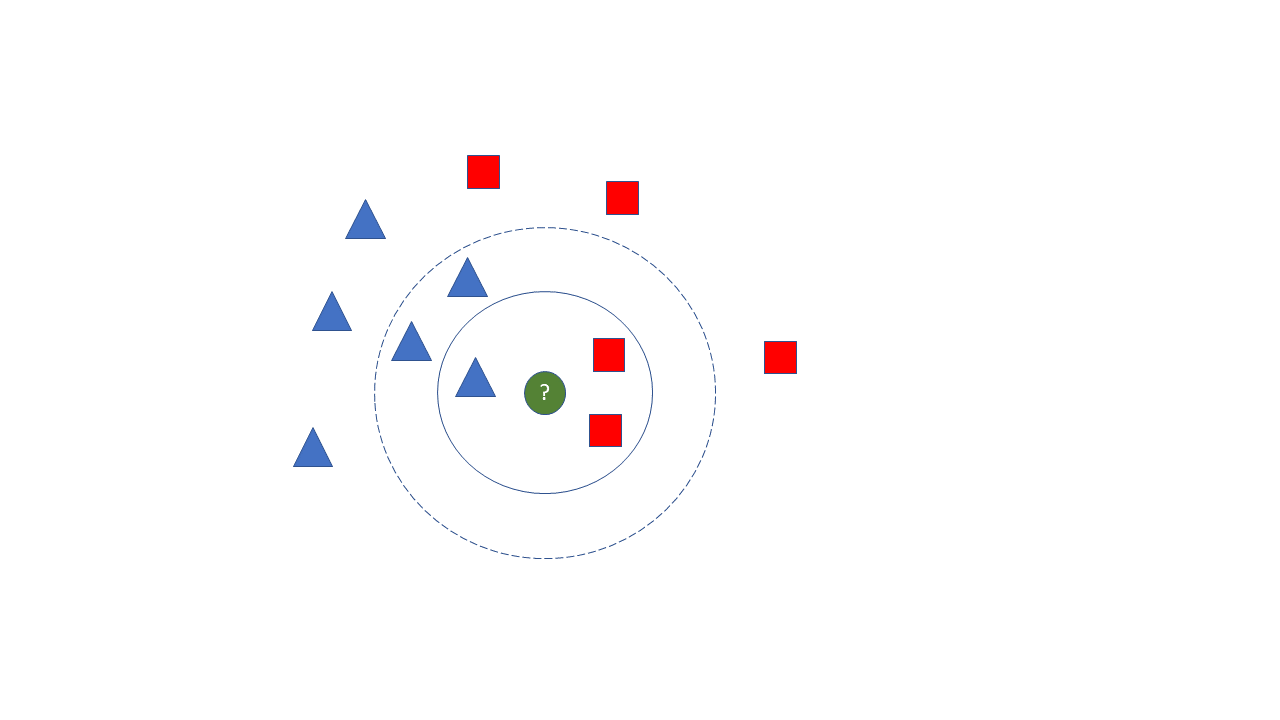
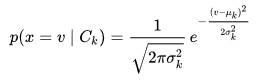


Fig.2 KNN with different varaitions

For example, if we want to classify the green point in our dataset it would be either classified as red or blue which can be rating 1 and rating 2 of the players respectively then this player would have the rating 2 if we use K=3 nearest neighbors but if we use K=5 for example the player would have rating 1 and so on.

3. Naïve Bayes

### Gaussian NB model is used with its default settings. We first segment the data by the class, and then compute the mean and [variance](https://en.wikipedia.org/wiki/Variance#Estimating_the_variance) of x in each class. Let μ k be the mean of the values in x associated with class Ck, and let σ k ^2 be the variance of the values in x associated with class Ck. Suppose we have collected some observation value v. Then, the probability distribution of v given a class Ck, p ( x = v ∣ Ck ) , can be computed by plugging v into the equation for a [normal distribution](https://en.wikipedia.org/wiki/Normal_distribution) parameterized by μ k and σ k 2 . That is,



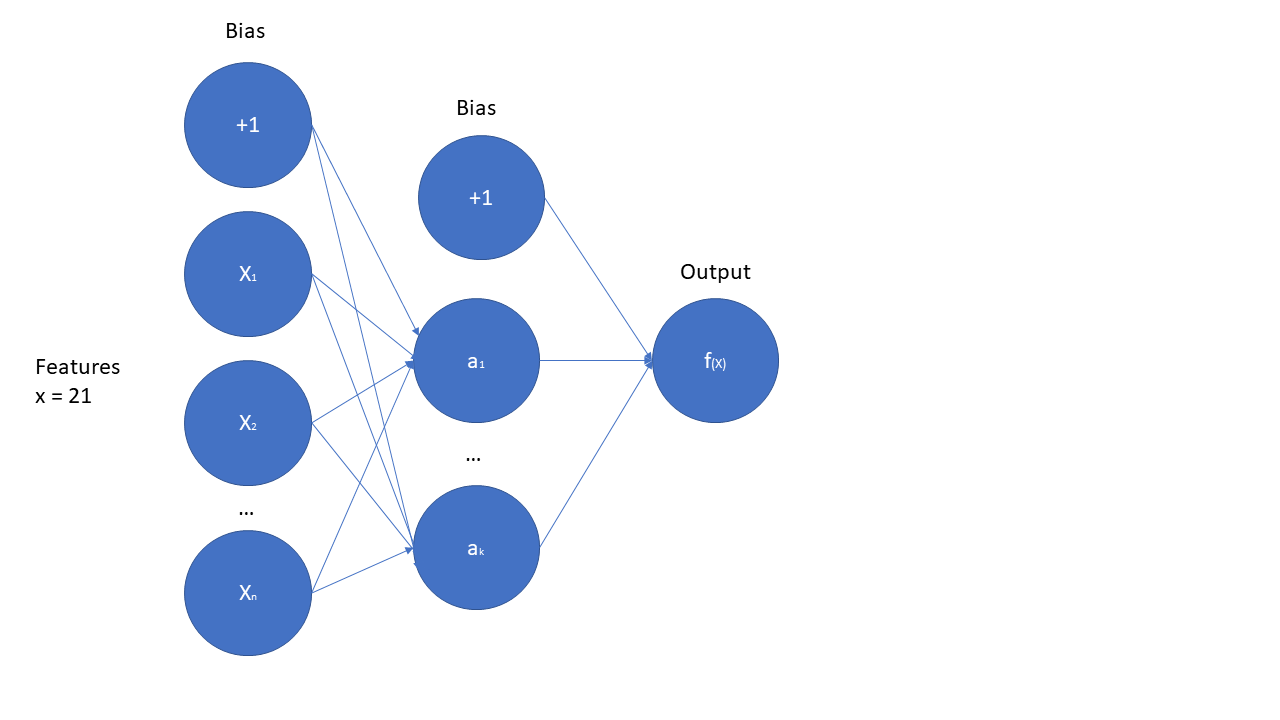
### 4. Neural Networks:

## MLP Classifier is used for the Neural networks model with several different values for the number of hidden layers but settled at the values that give the best accuracy. Relu activation function has been used since it gives better results than other activation functions.

After some experimentation with other solvers, it has been found that the “Adam” solver is the best for the solver argument

The graph below shows one hidden layer MLP, the leftmost layer, consists of a set of neurons representing the input features. Each neuron in the hidden layer transforms the values from the previous layer with a weighted linear summation followed by a non-linear activation function (the relu activation function). The last layer transforms them into output values which are the class values. In this project 3 hidden layers used to give better results.

Figure 2: 1 hidden layer MLP



## Fig3.MLP Model

## **Clustering**

K-means is applied to the dataset with k = 5 in combination with Axes3D library for 3D plotting of the clusters. Clusters are further improved via Principal Component Analysis (PCA).

## **Regression**

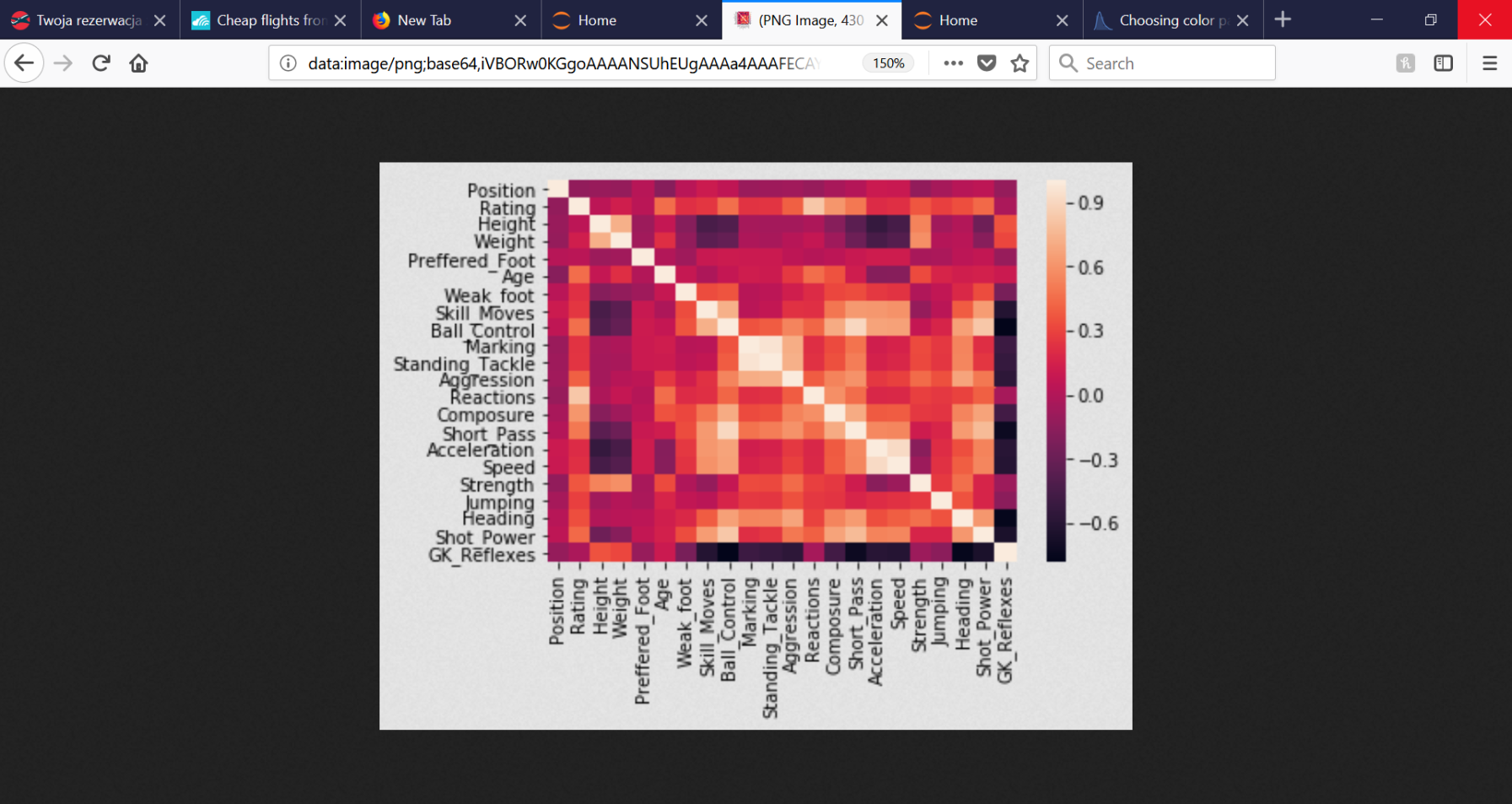
Logistic regression has been conducted with the goal of predicting the players’ rating. Linear List Squares (OLS) regression yielded a (0.84) adjusted R-squared, with an F-statistic: 3530, and Durbin-Watson: 1.539. The model has a strong explanatory power. 84% of variation in the rating is attributed to change in the independent variables. Linear Regression from the sklearn library is also used but only shows the R squared value which is 90%, so in comparison; the OLS model is more informative.

**Results**

The average rating among players is 66 points (min. 62, max. 94). The average age of the players is 25 years old (min. 17, max. 47).

Correlation suggested a strong and moderately strong relationship between rating and the following variables: reactions (0.82), composure (0.61), pass (0.49), age (0.45), shot power (0.44), and ball control (0.42).

Fig 4. Heatmap (Correlation between the features)



Different comparing techniques were applied to the classification models which includes: Accuracy, precision, recall and f-score.

Classification models have been built to classify and predict the rating class and the position of players. The KNN model achieved the highest numbers with 2 nearest neighbors on the test set. In general, the SVM model, compared to the KNN model with its different modifications, provides better results basically because of the advantage: Once a boundary is established, most of the training data is redundant. All it needs is a core set of points which can help identify and set the boundary.

Neural networks outperformed the KNN model and have close performance to SVM model (Table 2). MLP requires less computation time than SVM and therefore is more desirable.

Table 2. Classification results (Rating/Position)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Target/model | SVM | KNN | Naïve Bayes | MLP |
| Accuracy | 100%  /100% | 92%/  77% | 100%  /100% | 99.6%/  97.2% |
| Precision | 100%/  100% | 93%/  78% | 100%/  100% | 99%/  98% |
| Recall | 100%/  100% | 92%/  77% | 100%/  100% | 100%/  98% |
| F-1 score | 100%/  100% | 92%/  76% | 100%/  100% | 99%/  97% |

The Naïve Bayes generative model gives 100% accuracy for classifying both position and rating. It outperforms other classifiers since the dataset contains many records and relatively few features that allows the Naïve Bayes model to train well. It also requires short computational time for training and testing (less than 5 seconds).

Table 3 shows the results for the classification models (average accuracy) using cross validation with 10 folds for the rating class and 4 folds only for the position class because one of the position class values has only 4 tuples.

Table 3. Classification accuracy with CV

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class average accuracy/Model | SVM | KNN | Naïve Bayes | MLP |
| Rating | 99.9% | 78.6% | 100% | 99.9% |
| Position | 99.3% | 50.7% | 100% | 95.7% |

KNN, being one of the simplest machine learning algorithms, is prone to over fitting and that may be the reason for a lower accuracy in comparison with other models. KNN is also more effective on smaller datasets; hence it is not ideal for our study.

Based on the Euclidean distance measures, we have been able to compare players and identify the most similar player. For instance, the most similar player to Cristiano Ronaldo, based on selected distance columns, is Ivan Perišić.

K-means algorithm is used to identify 5 clusters (Fig. 5). The k-means cluster centroids suggest that the highly ranked players are on average 26 years old and score particularly high on features such as attacking, speed and reactions. Interestingly, highly rated players score lower on aggression. Cluster with the rating centroid of 64.46 is the goal keeper cluster, as the goal keeper skill centroid has a particularly high value and the players are on average taller and heavier. Cluster with the lowest rated players (bottom left) consists of young (under 22) and inexperienced players.

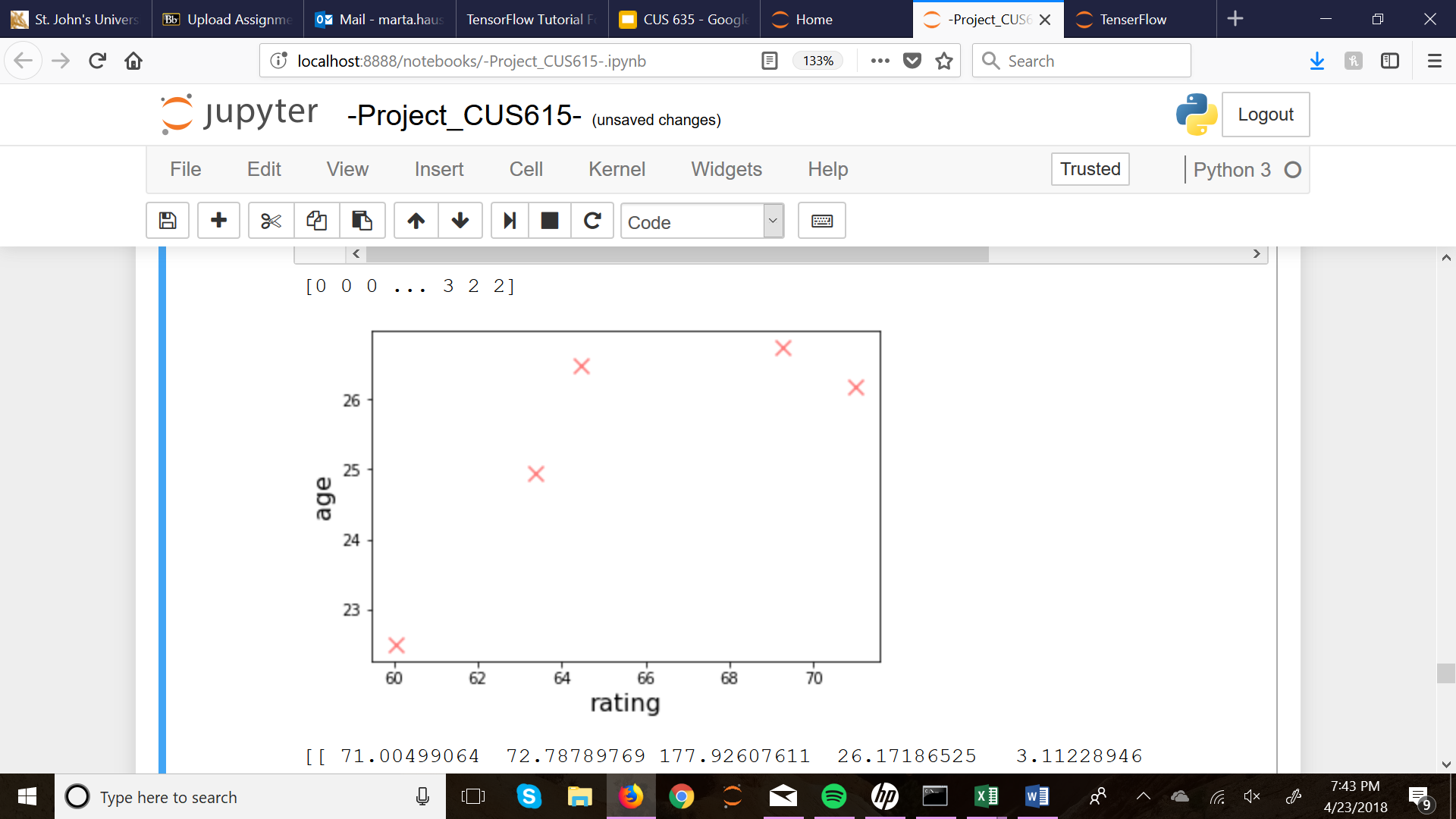
Fig.5 K-means cluster centroids

Figure 6 displays the K-means clusters with age, rating and position on x, y and z axis.

Fig. 6: 3D Plot of the 5 clusters (k=5)

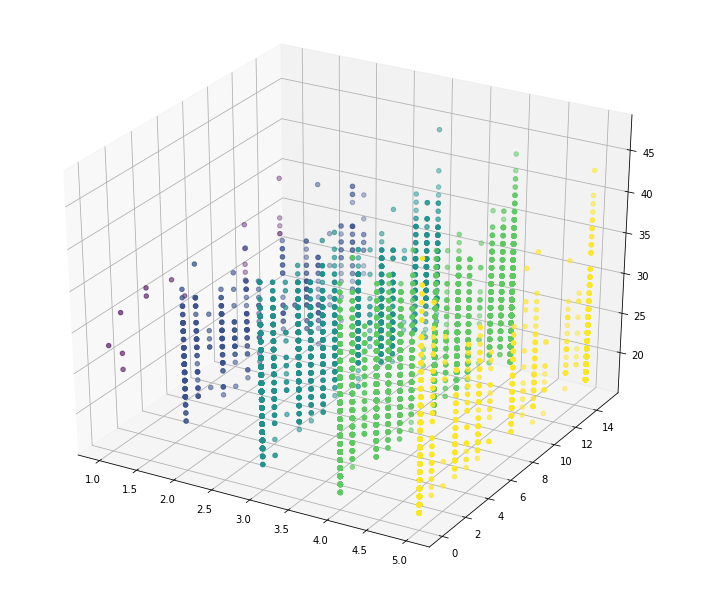
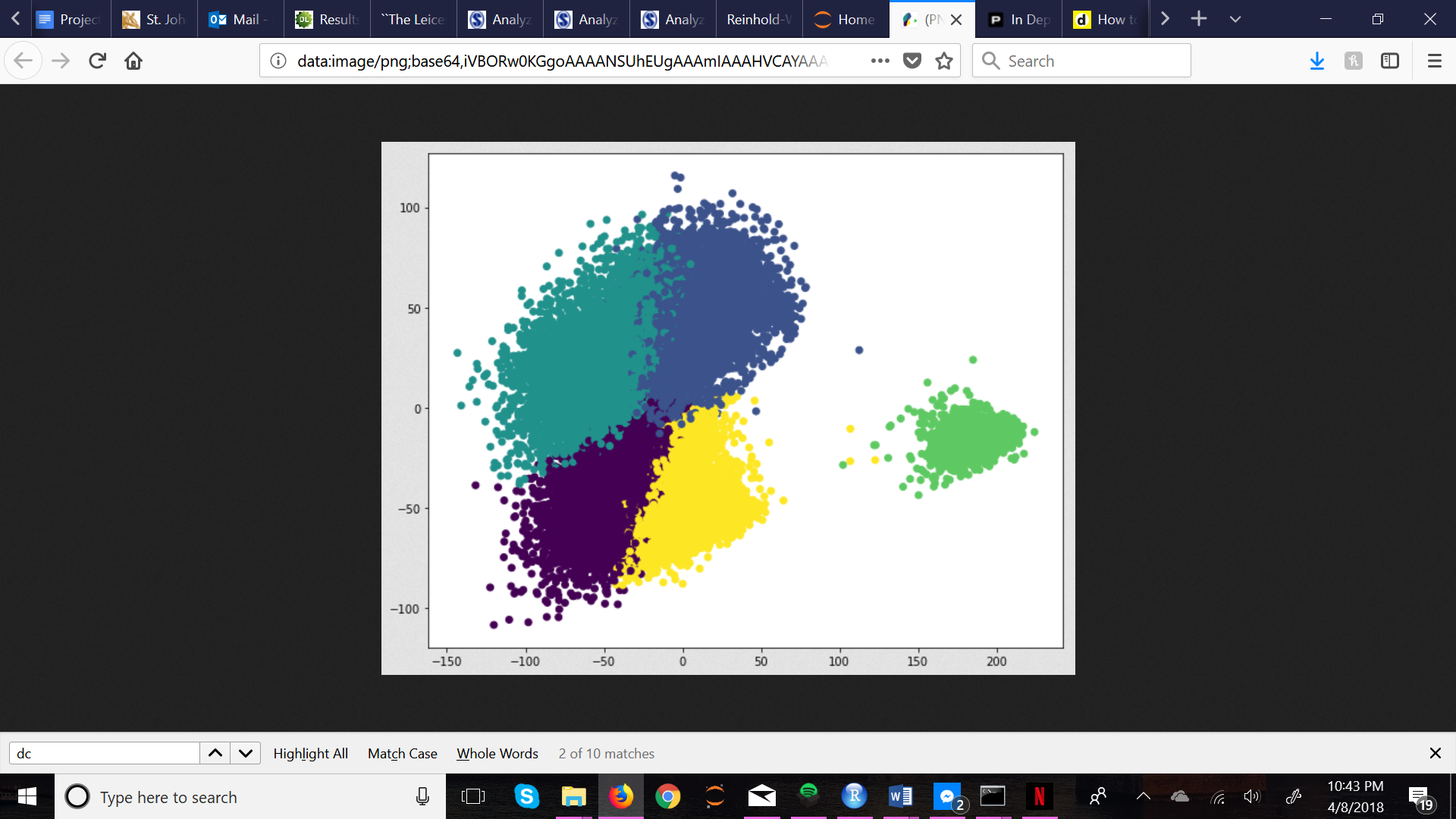


Fig.7 PCA Clusters



To compress the data features and reduce noise, the Principal Component Analysis has been applied. Figure 7 displays the resultant clusters. The first principle component (PCA1, x axis) explains 55% variation in the data, while the second component (PCA2, y axis) explains 19%.

**Conclusion**

This paper has displayed an implementation of classification and clustering algorithms on a soccer player dataset. The study tried to predict the classification accuracies of position and rating class based on training examples. The most accurate and efficient model for our particular dataset is Naïve Bayes, given its high accuracy and relatively short computational time. In predicting the players’ rating, ‘reactions’ is the most important feature; hence it is the most attributed to variations in rating.

**References**

[1] Aleksi Visti, Tapani N. Joelsson, and Jouni Smed. 2017. Beyond skill-based rating systems: analyzing and evaluating player performance. In Proceedings of the 21st International Academic Mindtrek Conference (AcademicMindtrek '17). ACM, New York, NY, USA, 8-16. DOI: <https://doi.org/10.1145/3131085.3131096>

[2] Arnu Pretorius and Douglas A. Parry. 2016. Human Decision Making and Artificial Intelligence: A Comparison in the Domain of Sports Prediction. In Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists (SAICSIT '16), Frans F. Blauw, Marijke Coetzee, Duncan A. Coulter, Elize M. Ehlers, Wai Sze Leung, Carl Marnewick, and Dustin T. van der Haar (Eds.). ACM, New York, NY, USA, , Article 32 , 10 pages. DOI: <http://dx.doi.org/10.1145/2987491.2987493>

[3] Basma E. El-Demerdash, Ihab A. El-Khodary, Assem A. Tharwat, and Eslam R. Shaban. 2016. Performance Evaluation of European Football Teams Using Data Envelopment Analysis. In Proceedings of the 10th International Conference on Informatics and Systems (INFOS '16). ACM, New York, NY, USA, 325-326. DOI:<https://doi.org/10.1145/2908446.2908512>

[4] Carlos Pestana Barros and Stephanie Leach. 2006. Analyzing the Performance of the English F.A. Premier League With an Econometric Frontier Model. Journal of Sports Economics7,4(2006),391–407.

DOI: http://dx.doi.org/10.1177/1527002505276715

[5] Hector Ruiz, Paul Power, Xinyu Wei, and Patrick Lucey. 2017. "The Leicester City Fairytale?": Utilizing New Soccer Analytics Tools to Compare Performance in the 15/16 & 16/17 EPL Seasons. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17). ACM, New York, NY, USA, 1991-2000.

DOI: https://doi.org/10.1145/3097983.3098121

[6] Joel Brooks, Matthew Kerr, and John Guttag. 2016. Developing a Data-Driven Player Ranking in Soccer Using Predictive Model Weights. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). ACM, New York,

NY, USA, 49-55.

DOI: https://doi.org/10.1145/2939672.2939695

[7] Pedro O.S. Vaz de Melo, Virgilio A.F. Almeida, and Antonio A.F. Loureiro. 2008. Can complex network metrics predict the behavior of NBA teams?. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08). ACM, NewYork,NY,USA,695-703.

DOI: <https://doi.org/10.1145/1401890.1401974>

[8] Ralph L Piedmont, David C Hill, Susana Blanco, Predicting athletic performance using the five-factormodel of personality, Personality and Individual Differences, Volume 27, Issue 4, 1999, Pages 769-777,ISSN 0191-8869,DOI:<https://doi.org/10.1016/S0191-8869(98)00280-3>.

[9] Rao Xiong, Mac Greene, Vahe Tanielian, and Jeffrey Ulibarri. 2017. Research on the Relationship between Salary and Performance of Professional Basketball Team (NBA). In Proceedings of the 8th International Conference on E-business, Management and Economics (ICEME 2017). ACM, New York, NY, USA, 55-61. DOI:<https://doi.org/10.1145/3157754.3157765>

[10]Sanjay Chawla,Joël Estephan, Joachim Gudmundsson, and Michael Horton. 2017. Classification of Passes in Football Matches Using Spatiotemporal Data. ACM Trans. SpatialAlgorithmsSyst.3,2,Article6 (August2017),30pages.

DOI:<https://doi.org/10.1145/3105576>

[11] H. Trieu Pham, Yonggwan Won, and Jungja Kim. 2012. Controlling multi-class error rates for MLP classifier by bias adjustment based on penalty matrix. In Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication (ICUIMC '12). ACM, New York, NY, USA, , Article 24 , 9 pages. DOI: <http://dx.doi.org/10.1145/2184751.2184781>

[12] Zongyuan Zhao, Shuxiang Xu, and Yunling Liu. 2014. Investigation of Multilayer Perceptron and Class Imbalance Problems for Credit Rating. International Journal of Computer and Information Technology 03, 04 <https://www.ijcit.com/archives/volume3/issue4/Paper030418.pdf>