Classification of Overall Performance of Fifa Players
Final Project
CS663

Students
Dipak Dulal,Thomas Petrossian
Professor
Dr.Chengcui Zhang
Department of Computer Science
UAB

May 8, 2021

Abstract

There are many ways people determine the overall performance of professional Fifa players. The overall value is used during forming line-ups, trades and even in video games. Due to incorrect evaluation of players' skills, numbers of players don't get the chance to show themselves on the field, or have limited opportunities.We have used data mining and machine learning algorithms to determine the overall performance of players based on their categories. Our model is based on various attributes of professional Fifa players. By using this model players could be categorized fairly and get the opportunity they truly deserve.

1 Introduction

FIFA(Federation Internationale de Football Association) was established on May 21st 1904. There are many professional athletes with different types of skills that play for many soccer leagues such as World Cup, Champions League, La Liga, Europa League etc. Therefore, Fifa holds large amounts of player data. We have used the FIFA data from www.kaggle.com which provides over eighteen thousands soccer player's records. Our goal is to build a prediction model using data mining and machine algorithms that can effectively classify the overall performance according to category. In the upcoming section we will show the preparation and the preprocessing of our data. In section 3 we will visualize our data, and in section 4 we will discuss our classifier models and evaluate them with couple of evaluation metrics. And in the last section we will present our future work and improvements planned for this project.

2 Data Preparation

Our data consists of 18207 rows and 89 columns. We have dropped the majority of the attributes to its low or no contribution to the overall performance of players. For instance, "Jersey Number" and "Contract length" do not have affect on the overall performance of the player. Our next step in data preparation was to make sure that we do not have any missing values in our data. Since we had only 48 rows where some values were missing, the optimal decision was to drop those records instead of doing data augmentation. Additionally, we had an attribute "Work rate" which consisted of two values for instance, "Low/High". Since, this is an ordinal value we took the mean of these two values. Furthermore, we have used the technique of standard scaling to normalize our data. This has been applied to the features height, weight, age, overall, potential, balance and to other numeric attributes. Moreover, we have removed unit measures and converted the values to one unit measure. For instance the "6'2" value for height was converted to 74. For our last step in the data preprocessing we have divided the overall values into three groups according to the standard scaled results, where values higher than 2 are considered "High", lower than -1.5 are considered "Low", and others are "Average".

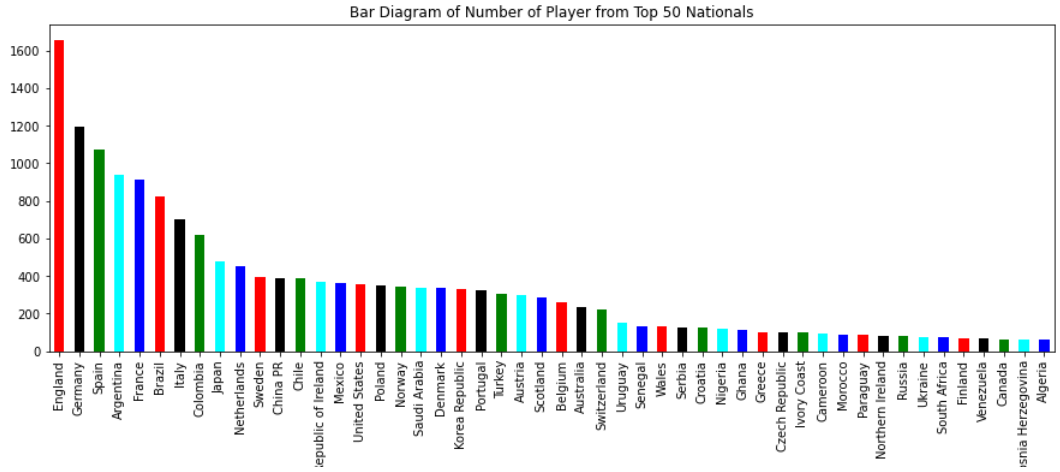
The Data before processing

	Name	Age	Nationality	Height	Weight	Overall	Potential	Balance	Stamina	Strength	Dribbling	ShotPower	Jumping	Acceleration	Work Rate
0	L. Messi	31	Argentina	5'7	159lbs	94	94	95.0	72.0	59.0	97.0	85.0	68.0	91.0	Medium/ Medium
1	Cristiano Ronaldo	33	Portugal	6'2	183lbs	94	94	70.0	88.0	79.0	88.0	95.0	95.0	89.0	High/ Low
2	Neymar Jr	26	Brazil	5'9	150lbs	92	93	84.0	81.0	49.0	96.0	80.0	61.0	94.0	High/ Medium
3	De Gea	27	Spain	6'4	168lbs	91	93	43.0	43.0	64.0	18.0	31.0	67.0	57.0	Medium/ Medium
4	K. De Bruyne	27	Belgium	5'11	154lbs	91	92	77.0	90.0	75.0	86.0	91.0	63.0	78.0	High/ High

The Data after processing

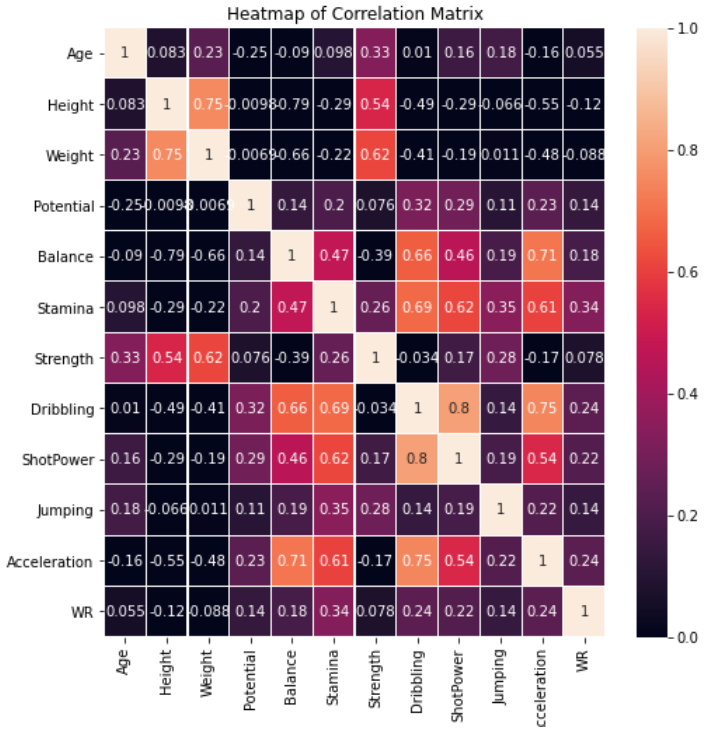
Name	Age	Nationality	Height	Weight	Overall	Potential	Balance	Stamina	Strength	Dribbling	ShotPower	Jumping	Acceleration	WR
L. Messi	1.258441	Argentina	-1.646010	-0.447583	High	3.697415	2.195382	0.552403	-0.502679	2.201445	1.713704	0.246247	1.767621	-0.440246
Cristiano Ronaldo	1.686666	Portugal	0.995907	1.091577	High	3.697415	0.426820	1.559053	1.090102	1.725503	2.293836	2.530565	1.633639	-0.440246
Neymar Jr	0.187878	Brazil	-0.891177	-1.024769	High	3.534396	1.417214	1.118643	-1.299069	2.148563	1.423639	-0.345984	1.968594	1.006448
De Gea	0.401990	Spain	1.750741	0.129602	High	3.534396	-1.483228	-1.272151	-0.104484	-1.976272	-1.419003	0.161642	-0.510075	-0.440246
K. De Bruyne	0.401990	Belgium	-0.136343	-0.768242	High	3.371377	0.922017	1.684884	0.771546	1.619738	2.061783	-0.176775	0.896737	2.453143

3 Data Visualization



By observing the bar-plot of number of players from top 50 countries, we found that England consists

the highest number of players(more than 1600). The majority of the players are from Europe, South America and Asia.



From the above heat map we found the interesting correlations between different features. First, there are high positive correlation between dribbling and shot power(0.8), similarly between jumping and dribbling, stamina and jumping. Acceleration and dribbling", etc. are highly correlated in positive manner, whereas, balance and weight have high negative correlation. In addition, weight and potential , height and age, height and potential, balance and age are almost not correlated.

4 Classification and Evaluation

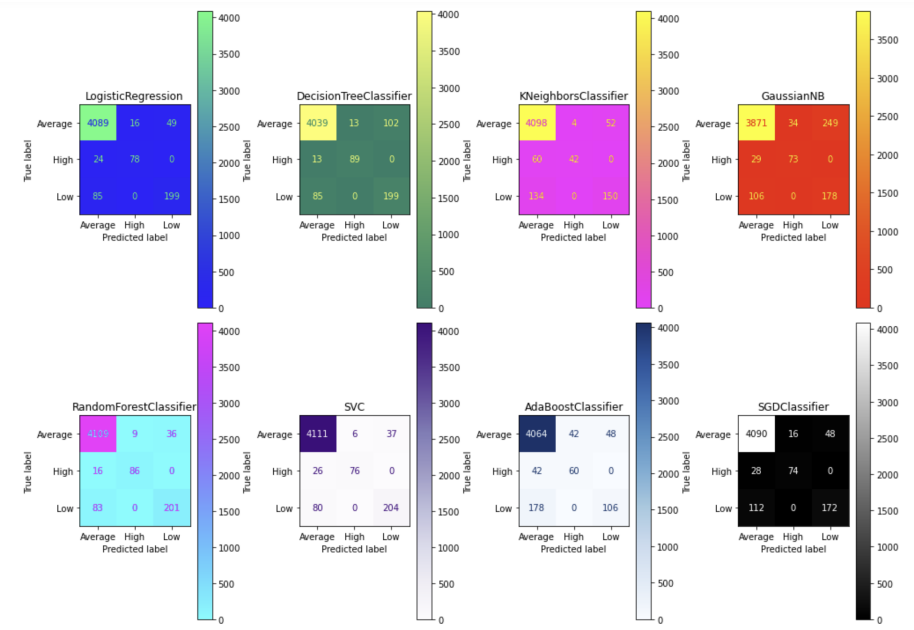
After preprocessing we have applied eight different machine learning models to classify the overall performance. We have used the classifiers Random Forest, Logistic Regression, Decision Tree, Support Vector Machines, K-Neighborhood, Naive Bayesian, ADA Boost and Stochastic Gradient Descent. The best result was achieved with the Random Forest classification with an accuracy of 96.92% and F-Measure of 88.03%. Since the algorithm has an unbiased approach, we could achieve the highest result with it. In contrast, the Naive Bayesian classification which "assumes" that the attributes are independent from each other gave the worst accuracy. The other models Support Vector Machine and Logistic Regression also gave good results. Considering the fact that the majority of the data was in the "Average" category,the overall accuracy was high.

The following table shows the cross validation results of our models. Overall the best result was achieved by the ADA boost algorithm, which showed great results for every group of cross-validation and has the highest mean cross-validation score.

K-Fold Cross Validation Scores(k=4)

Algorithm	cross validation score(in%)			
SVC	78.85	96.37	97.03	69.55
Random Forest	86.05	96.37	97.20	62.79
Decision Tree	86.08	97.42	98.06	58.96
K-Neighborhood	87.04	93.39	93.92	79.95
Logistic Regression	84.14	96.71	97.22	68.01
Naive Bayesian	83.23	93.15	92.20	76.76
ADA Boost	87.92	92.64	95.92	92.92
Stochastic Gradient Descent	91.14	94.27	95.81	72.77

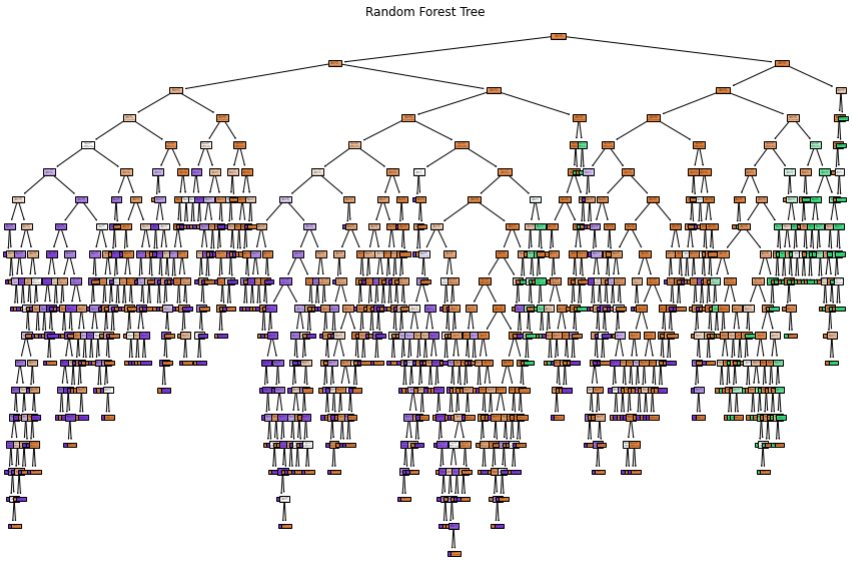
Confusion Matrices for all Models



Evaluation Scores

Algorithm	**Accuracy(%)**	**f1-score(%)**
SVC	96.72	86.18
Random Forest	96.92	88.03
Decision Tree	95.64	79.77
K-Neighborhood	94.1	64.24
Logistic Regression	96.17	84.11
Naive Bayesian	90.80	71.60
ADA Boost	93.17	67.85
Stochastic Gradient Descent	95.37	78.13

Visualization of our best model Random Forest Classification,which can be seen in the following Decision Tree:



5 Future Work and Improvements

The classification of the players' overall performance has room to improve. We hope to gather more data, which includes more records and features, which will be in the "Low","High" categories. Furthermore, we plan to collect more features that will describe the professional performance of the player such as goals, assists that will lead our model to a be a competent classifier.

References

[1] Jiawei Han, Micheline Kamber,Jian Pei, Data Mining Concepes and Techniques

[2] Lecture Slides of Data Mining CS663/763,Prof. Dr. Chengcui Jhang

[3] FIFA 21 complete player dataset
https://www.kaggle.com/stefanoleone992/fifa-21-complete-player-dataset

[4] Classification in Machine Learning — Machine Learning Tutorial — Python Training — Edureka
https://www.youtube.com/watch?v=pXdum128xww

[5] Wikipedia;https://en.wikipedia.org/wiki/

[6] Python Code for this project
`https://github.com/TomPetrossian/CS663_Data_Mining_Final_Project`