Football Player Selection Based on Positions and Skills Using Ensemble Machine Learning and Similarity Measure Techniques

MSc Research Project
Data Analytics

Murugappan Murugappan
Student ID: x19239831

School of Computing
National College of Ireland

Supervisor: Mr. Aaloka Anant

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Murugappan Murugappan		
Student ID:	X19239831		
Programme:	Data Analytics	**Year:**	2022
Module:	MSc Research Project		
Supervisor:	Mr. Aaloka Anant		
Submission Due Date:	31/01/2022		
Project Title:	Football Player Selection Based on Positions and Skills Using Ensemble Machine Learning and Similarity Measure Techniques		
Word Count:	**7447**	**Page Count: 20**	

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Murugappan Murugappan
Date:	31/01/2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Football Player Selection Based on Positions and Skills Using Ensemble Machine Learning and Similarity Measure Techniques

Murugappan Murugappan
x19239831

Abstract

The game of football attracts millions of fans across the world and also the production of players that makes a team effective is really hard when it comes to selecting players based on manual procedures by sports and team analysts. This research has performed various ensemble machine learning algorithms that can automatically predict the position of the players with two novel approaches based on different features and performed statistical feature selection techniques to select the top 30 features in predicting the player's position. Throughout this research, four ensemble machine learning algorithms are performed on which random forest classifier gave the highest accuracy in the approach of predicting only the 4 major positions of the players, and support vector classifier results in better performance in the approach of predicting the 27 different major and minor positions. Also, hyperparameter tuning results in no such huge improvement in both approaches. Also, Used different similarity measure techniques such as cosine similarity and Euclidean distance measures to select the most similar players based on their skills. Finally, cosine similarity performed better compared to Euclidean distance and can be applied in different sports domains where players can be selected based on their skills. The evaluation of all the models has been done with the help of evaluation metrics such as accuracy, precision, recall, F1 score, and cross-validation score. Also, Evaluated both basic and tuned models with the help of the confusion Matrix where the level of truly predicted outcomes and the misclassified outcomes is seen.

1 Introduction

The introduction section in this research "Football Player Selection Based on Positions and Skills Using Ensemble Machine Learning and Similarity Measure Techniques" consists of explaining the background information of this research, the motivation behind this research idea aims and objectives of this research, discussing the research questions, and finally, the project outline and structure of this report are discussed.

1.1 Background

Football is one of the most played games across the world. As per the report supported by the FIFA board in the 21st century, football attracts around 3.5 billion fans across the world and is played in over 180 countries where around 250 million football players are available across the world. The matches played in any football championship become a matter of interest asa prediction of such matches become an important subject to researchers, football experts as well as fans across the world. It was quite easy to determine the probability of winning the particular match based on the form of the key players, skills, combination of players, teamwork as well the strength of the squads, and many other factors. The prediction becomes difficult when matches are either extended or become complex due to different factors. The prediction of football matches also depends on different players of a team where a coach may also, give importance to determining and analyzing the impact of a particular player. The selection of players plays an important role in winning a team and is very crucial to take forward the entire team to win a championship. Various financial issues occur where millions of dollars are lost by particular clubs where the player turns out less expected to the team and the whole team loses based on the selection of players. Different psychology assessments are necessary to evaluate the number of players present in a team and its benefits when it comes to implementing a winning team (Zaini and Salimin, 2020).

1.2 Motivation

Different problems are encountered during team selection which is why the player recruitment process is carried on to select the best players that can take forward a particular team for winning a Championship. This is why different managers, coaches, and other experts are recruited who can choose a better player based on skills and other attributes. Different quantitative and qualitative attributes are necessary to be studied to select the best players and such attributes include the performance, fitness, skills of the players, etc (Haugen, 2017).

There are also other issues encountered after the final selection of the players because if the player moves from a team due to some reason then it becomes difficult to complete the team with another player that matches the skills of the previous player. In such cases, different predictive models are used that can process different attributes to assess the skill of a player in less time.

Different research had been performed to predict suitable players based on skills and other factors. A study offered by (Ozceylan, 2016) performed and implemented the predictive model to select the player with the help of a classification model that can predict the rating of a particular player. Another study is performed by (Qingwen, 2020) where the performance of a player can be predicted based on the previous matches he played. On completion of this research, the attributes taken are the wickets by the bowler or the run which is code by a batsman. Another research (Strnad, Nerat, and Kohek, 2015) performed where a neural network is implemented to predict a player suitable for the teamand achieved a result that is not significant. Some researchers till now have been performed but did not give any significant result that can replace a better player with another player based on a different number of attributes. This research tries to find out the player based on attributes with the help of machine learning models by considering 27 different positions in a novel approach and predicting the closest player based on the attributes based on similarity measures. Also, the results obtained from this research will be evaluated with the help of performance metrics and such metrics include recall, precision, accuracy as well as F1 score measure.

1.3 Aims and Objectives

The project aims to predict the Position of a player with the help of a machine learning model and find the closest players based on such attributes. The objectives of the projects are

- To perform a literature review to assess the approaches used previously in the same topic.
- To explore and visualize the data extracted from the Kaggle to understand the nature of the data and the attributes related to a player.

- To perform machine learning algorithms to predict the position and to assess the skills of the players such as Support Vector Classifier, Logistic Regression, Random Forest Classifier, and Decision Tree Classifier.
- To perform preprocessing of the features to understand the relationship between the features and select the best features that can predict the similarity of a player
- To find out the similarity between two players with the help of similarity measures suchas Euclidean distance
- To evaluate all the machine learning models with the help of performance metrics such as accuracy, precision, recall, F1 score, etc
- To tune the models with the best combination of parameters that can give the highest accuracy obtained from each model.

1.4 Research Questions

This research focuses on creating different ensemble machine learning models to select a player by predicting the player position in two novel approaches and to find similar players based on skills and other attributes that match with the selected player using similarity measure techniques. The research questions are as follows

- "Can the prediction of the player's position in an advance will helps in selecting a player to play in a particular position in the team?"

- "Can I predict a suitable player based on the skills and features had by another player with the help of machine learning techniques?"

1.5 Project Outline

The remaining section is organized as follows. Section 2 indicates the Related Work, Section 3 indicates the research methodology, Section 4 indicates the design specification, Section 5 indicates the implementation specification, section 6 indicates the evaluation and discussion, section 7 indicates the conclusions and future work, then finally acknowledgments and reference section of this research.

2 Related Work

Researchers had shown deep interest in sports prediction and selection of different players based on particular skills with the help of machine learning and their approaches. Let us look into some of the approaches used by them with the help of Different techniques and compare the results with our approach.

2.1 Sports prediction using machine learning

A Study (F.I. and J.C, 2015) used a machine learning technique and developed a Framework that can predict sports with the help of an Artificial Neural Network. The architecture consists of a crisp DM Technology where the data sources are organized and evaluated with the help of an Artificial Neural Network and a Framework is developed to predict sports that can be applied with the help of neural network techniques.

Another study (Berrar, Lopes, and Dubitzky, 2018) predicted the football competition with the help of public data and different machine learning approaches in this study. They conducted Bayes classifier, multi-layer perceptron as well as principal component analysis to reduce the dimension of the data. With this study, they achieved a 54.7% accurate result and the result was obtained by combining the techniques of naive Bayes, multilayer perceptron, and principal component analysis.

Another technique used technique of big data and machine learning approaches to develop a framework to predict different games. The system developed in the study (Baerg, 2016) was used to collect and store data and perform different analyses to look into the performance ofplayers and select the most important features that can predict the output of the games. To predict the output, the study developed software that can predict 90% accuracywith the help of a Support Vector Machine and ensemble learning algorithm that can select the best features to give the desired outcome.

A concept of the study of Social Network and gradient boosting technique is applied to predict the fate of the team of a particular game. In the experiment (Cho, Yoon, and Lee, 2018), they have conducted a gradient boosting technique and had concluded that it is the best technique compared to the study of social network techniques. So they had evaluated the result with the help of win draw loss accuracy and win-loss prediction accuracy and concluded that wind- draw-loss accuracy is 33% and win-loss-prediction accuracy is 50% which is more compared to the wind-draw-loss accuracy in the experiment.

2.2 Player prediction using machine learning

Player replacement in football was conducted in a study (Abreu and Sannikov, 2014) with the help of machine learning algorithms where players were replaced based on the ratings and different classification models were performed to predict the rating on the players so that they can be replaced with their similar players with similar ratings. They had performed different machine learning algorithms and had concluded in the study that the LDA algorithm gave better performance compared to all other algorithms.

Another study (Jayalath, 2018) had built different classification models to predict the performance of players based on the batsman score and the wickets taken down by a bowler. They predicted both these parameters based on the runs a batsman makes and the wickets a bowler takes. In the study, they predicted the player’s performance based on these two parameters and they had also used other features to predict this parameter such as batting average, strike rate, number of innings, etc. In this study, they applied different machine learning algorithms from which they had concluded that decision tree and random forest performed better compared to different algorithms used in the study.

Another algorithm was used known as a competitive neural network which was performed in a study (Chen, 2019) where the main aim of the research was to select the best opposite playersfor a team by predicting the chances of winning a game with the help of the neural network. The rating of each player was predicted to calculate the winning chance of a team and the feature selection was performed where the neural network predicted the probability for both win and loss. The model has been applied to 11 players that were found to be 54% accurate and another model had conducted and doubled the number of players that gave 60% of accuracy score.

Another player prediction experiment was conducted with the help of a neural network in a study (Guan and Wang, 2021) where they applied the data set that includes the result of the matches and the historical performance of different teams. With the help of different features that also include the player's data, they could predict the chances of winning based on the stakeholders. They also divided the attributes into different categories such as the physical status, technique of the players, speed of the player, and the resistance of a player. With the helpof neural network techniques, they calculated the accuracy and found that it is capable to predict the performance of a player but gives very little accuracy.

2.3 Player prediction using Other Algorithms

Apart from machine learning and AI approaches, some algorithms were used to predict the players based on particular features. A study conducted by (Fan *et al.*, 2019) predicted a player by building a team based on categorizing them by different task dependencies and the partitioning of the task was done among the number of the teams to build a team-building approach.

Another approach (PANAIT and BUCINSCHI, 2018) was used to build a multi-functional Team by using the concept of hierarchy process and Myers–Briggs indicator that can predict the member of a team based on the characteristics and concurrent engineering concept was used in the study.

Another concept known as the axiomatic design principle was used to build and predict a team where a team can be determined based on the skills and also increase maximum utilization of a particular team member (Betasolo, 2016). A skill development approach was implemented in the study to predict the talents of a particular team member.

Another player placement was used (Shajila and Vimala, 2017) with the concept of Fuzzy Logic where bipartite graphs with the help of particular weights were developed to implement a polynomial-time algorithm to replace the player based on particular skills.

3 Research Methodology

Various approaches and techniques will be applied as part of the Customized KDD approach in this research to determine the research outcomes. KDD also known as Knowledge Discovery in Database ispreferred in performing any research as it can be used iteratively and interactively (Kumar, Kumar Sehgal, and Singh Chauhan, 2011). Also, various stages are designed in this approach to perform any research, and some features such as looping as well as iterations through backstages had been made possible through this approach at any point. This KDD is customized according to my research idea and implementation but the high-level process remains the same as shown in Figure 1.

There are different machine learning models and preprocessing steps required to perform similarity to select similar players based on particular features. Also, various steps are necessary to predict the position of the player with the help of machine learning models and this can be explained with the help of the KDD approach in this section.

3.1 Dataset Description

The data set is known as a FIFA21 data set which is collected from Kaggle [1] which is an online data science platform where different machine learning models and deep learning models can be applied and data can be imported and exported. This platform is accessible to the public and does not encourage any social and ethical issues. Also, this data set contains around 18541 rows and 92 columns where the name of players, age of players, nationality of players, and other features related to the players are present in this data. There are missing values that are present in some of the features of the data which should be treated in the prepossessing steps. This data will then be separated for training and testing and different similarity measures will be conducted to select the players based on particular skills (Sharma *et al.*, 2019).

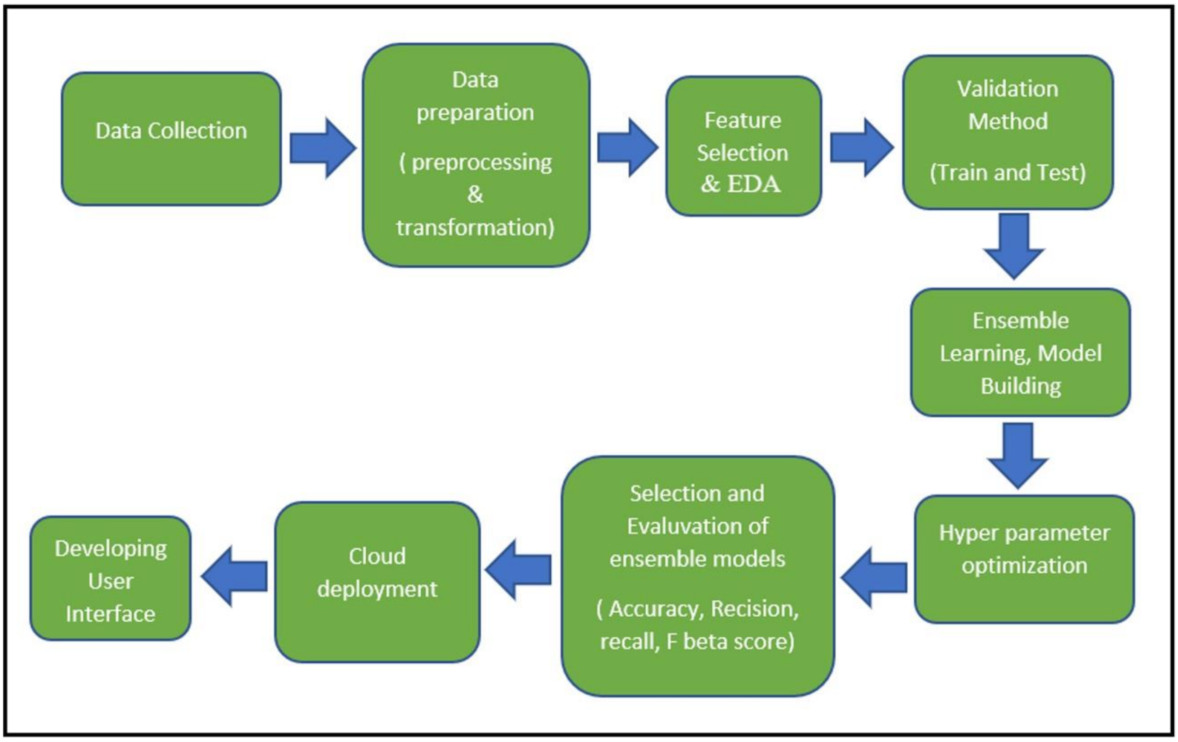

Data Collection
Data preparation
(preprocessing & transformation)
Feature Selection & EDA
Validation Method
(Train and Test)
Ensemble Learning, Model Building
Hyper parameter optimization
Selection and Evaluvation of ensemble models
(Accuracy, Recision, recall, F beta score)
Cloud deployment
Developing User Interface

Figure 1: Process Flow Diagram for the Proposed Methodology

3.2 Data Exploration and preprocessing

Data exploration plays an important role in exploring the nature of the data, presence of missing variables, characteristics of the data, presence of outliers, and the distribution of the data. It also helps to understand how dirty the data is and the preprocessing steps required cleaning and processing the data for various model building and evaluation steps (Iso *et al.*, 2019). In this step, I will be exploring the number of rows and columns of the data, exploring the presence of missing values present in the data. Also, I will perform data visualization to understand and analyze some of the important factors present in the data that can be helpful in this research.

[1] https://www.kaggle.com/umeshkumar017/fifa-21-player-and-formation-analysis

Also, various preprocessing steps are required such as treating the missing values with mean, median, and mode value and encoding the categorical variables present in the data into numeric values. Categorical encoding is done with the help of a Label encoder where each categorical attribute is converted to numeric values that can be fed into the model. Also, the preprocessing steps include cleaning of the data, considering important features, and data normalization that will convert all the values of the data in the range of 0 and 1. The preprocessing steps also include treating outliers, balancing the data in reducing the noise of the data as well as other steps that are considered very important before building any machine learning models in it.

3.3 Feature Selection

Feature selection is a very crucial step in choosing the important number of features when the size of the data is very large and contains lots of features. Also, it contains different typeswhere only important features are selected to reduce the number of input features that are not considered important. It is advantageous in reducing the computation cost and complexity of the system and also improves the accuracy of the model with the help of important features. There are different types of feature selection techniques used to build machine learning models. Some of the important methods of feature selection techniques are wrapper and filter method where wrapper method searches for subjects that are considered well-performing to select the important features and filter method selects those features subset which depends on the relationship with the target variable. Both these methods are used in a supervised approach where a target variable is necessary to find the relationship with the input variable. In the case of unsupervised learning, correlation is used to find the relationship between two input features and remove the redundant variables. There is also another feature selection technique called dimension reduction techniques such as PCA where the input data is projected in lower dimension space with the help of different mathematical operations (Sheikh, 2017).

3.4 Data Balancing

Machine learning models tend to give overfitting results or more error when it gets very imbalanced samples that belong to classes that are not balanced. When the model is trainedon minority samples and later tested in a large data set then it often gives overfitting and underfitting results as it got very less chance to understand the features belonging to minor classes. In such cases, I need to balance the data with the help of different balancing techniques such as oversampling, under sampling, SMOTE technique, ADASYN technique, and many other balancing techniques that can balance the classes with the help of different methods (Shim, Oh and Kweon, 2018).

3.5 Data Evaluation

Different models are trained and are tested on the data, they are evaluated with the help of different performance metrics to understand the level of classification that can be predicted by a particular model. Different performance metrics include confusion Matrix, Precision, recall, accuracy score, F1 score, etc where the level of false positive, true positive, false negative, and true negative rates are considered to understand the number of misclassification that can be given by a model after testing on an unseen data. This evaluation is important before applying the model in real-life applications so that I can confirm that the model is accurate enough to be applied in real-life applications and is fit for public use (Ma and Ladisch, 2019).

3.6 Model Implementation and Tuning

The ensemble models of support vector classifier, logistic regression classifier, random forest classifier, and decision tree classifier are implemented and cosine similarity and Euclidean distance approach are implemented as part of similarity measures approaches. Model tuning is done as an attempt to improve the model after looking into the level of miss classification given by the model with the help of evaluation metrics. Model tuning includes training the model with different combinations of parameters that are included in particular models where the best-fit parameters are considered to train the model and test it on unseen data to improve the model compared to the model that is trained with default parameters. This step is important to confirm if there is a chance of improving the model and if it gives better accuracy compared to the model with default parameters before applying it in real-lifeapplications (Falch and Elster, 2016).

3.7 Similarity Measure

Similarity measures are performed to understand the similarity between particular attributes based on the features connected to them. Different similarity measures are performed to understand the relationship between two attributes and in this research, I had performed similarity measures to predict the number of similar players compared to a particular player based on a particular skill. Cosine similarity measure and euclidean distance are some of the most commonly used similarity measures.

4 Design Specification

This research is designed based upon two-tier architecture. i.e One is the Application Tier and theother is the Client Tier. There is no data tier in this research design since the data has not been stored anywhere in the database to perform any DB-related activities. The application tier andthe client tier is explained below concerning this research

4.1 Application Tier

In the application tier, where all the business logic is incorporated. The logic behind the development of these ensemble machine learning models and similarity measure techniques logics are incorporated in this application tier.

4.2 Client Tier

In the client tier, all the visualization logic is incorporated. The libraries behind the model evaluation visualizations, libraries hosting the web app are incorporated in this client tier.

5 Implementation

5.1 Data Collection

The data set is collected from Kaggle which is also known as FIFA 21. The data set contains around 18541 rows and 92 columns. All the attributes of the data contain information about different players and different matches played between them and also the skills related to particular players. Also, the data contain a different number of missing values in different columns which should be treated during the data preprocessing steps.

5.2 Data Pre-Processing

During the data preprocessing, I have seen the data contains a large number of missing values in different columns. In this step, I have filled the missing values with their mean value and performed data visualization to understand the distribution of players in different countries or other relationship that exists between players and their skills. Also, the data contain different categorical variables which are converted to numeric values with the help of Label Encoder in this step.

5.3 Feature Engineering

This step includes the selection of top features which will be fed into our machine learning algorithms and also balancing the data that contain imbalanced classes. To detect the top features I have performed the SelectKBest algorithm which will select the top features based on the existing statistical analysis inside them. Also, I have balanced the data with the help of the ADASYN algorithm that will create synthetic points and balance the number of classes to achieve better accuracy.

5.4 Building Models

In this step, I have built different machine learning algorithms such as a Support Vector Classifier, Logistic Regression, Random Forest, and Decision Tree Classifier to predict 4 major positions as one approach and predict all 27 different positions as another novel approach in this research and explained below in evaluation section. I have also used different parameters to achieve better performance and performed model training that used the best parameters to give better performance. Also, I had performed a cosine similarity matrix and Euclidean Distance approach to predict the similar players based on particular skills.

5.5 Model Evaluation and Tuning

I evaluated the result with the help of performance metrics such as accuracy, precision, recall, F1 score, and cross-validation score where I trained the model 10 times and determined the accuracy.

Also, the tuned model results are compared with the base model based on the confusion Matrix where I can see the number of misclassifications in both types of models. Also, I have tuned the model with the help of RandomizedSearch CV and selected the best parameters that gave the highest accuracy. All these models are compared with the base model to look into the label on misclassifications that might be improved in the tuned model.

6. Evaluation

6.1 Exploration of Data

In this section, I had explored the data to look into the important features observed in this research.

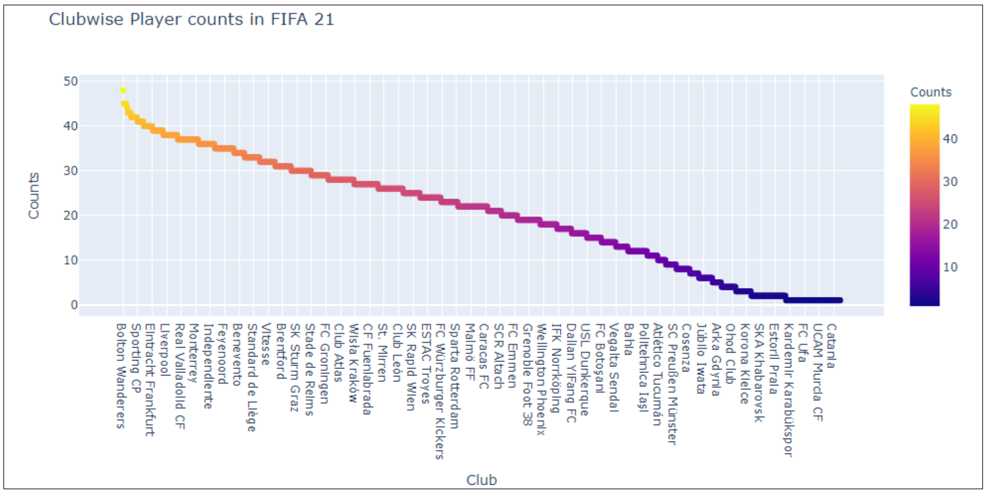


Figure 2: Club wise player analysis

From Figure 2, the club online Bolton wanderers, Sporting CP, Eintrack Frankfurt, Liverpool produces the highest number of players in FIFA 2021, and the clubs like FC UFA, UCAM Murda CF, Catania produce the least number of players representing the FIFA 2021 game.

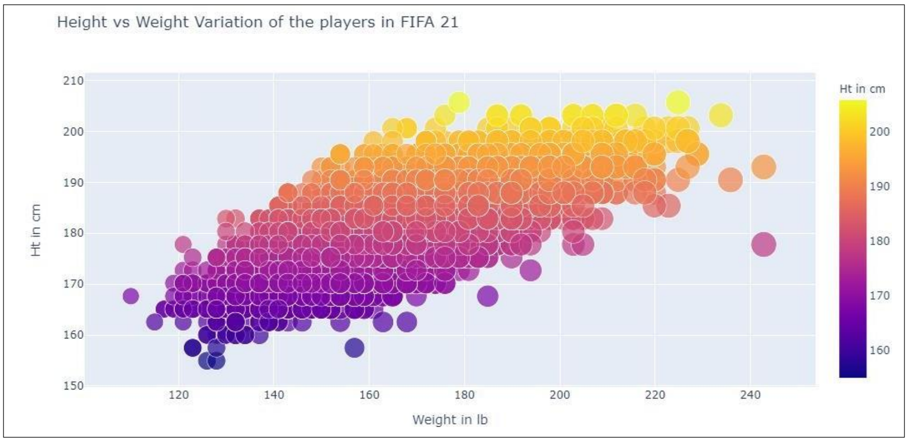


Figure 3: Height VS Weight Analysis

The above correlation represents the relationship between the height and weight of a player where I can see that there is a positive correlation between the height and weight of aplayer because when the weight increase, the height also increases and the correlation is strong that represents that the high weighted players are also longer in height is observed from the Figure 3.

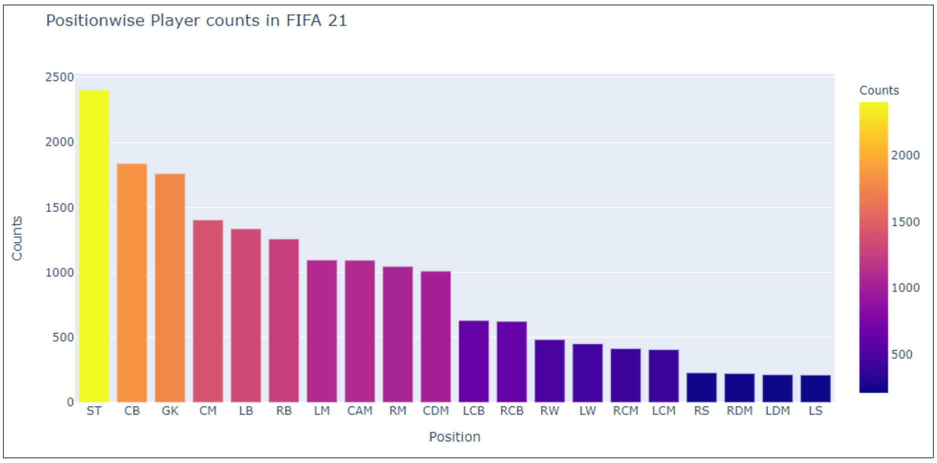


Figure 4: Position wise player

Here, ST represents the striker position, CB represents the center-back position and GK represents the goalkeeper position. So, I can say that these three positions have the highest number of players in the FIFA 2021 games.

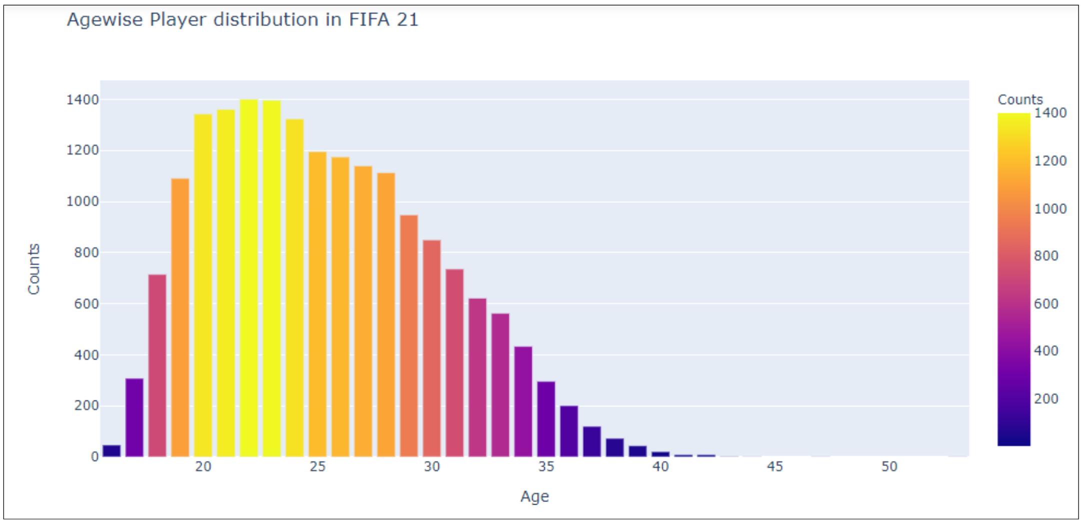


Figure 5: Age-wise player distribution

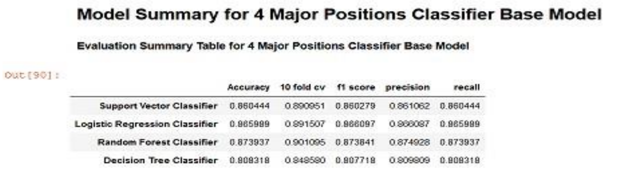
As the age of a player plays an important role in the football game, I can see that the age between 20 to 25 years has the highest number of players in FIFA 2021 where the old players who are above 35 years of age are very less representing the FIFA game. Figure 5 tells that only young players are fit to play in the FIFA game whose age ranges between 20 to 25 years.

6.2 Model Evaluation

The Ensemble machine learning models of Support Vector Classifier, Logistic Regression Classifier, Random Forest Classifier, and Decision Tree Classifier is implemented to predict the player positions in two different approaches using only 4 major positions and another using 27 different major and minor positions to select that player to play in a particular position and implemented similarity measure techniques like cosine similarity and Euclidean distance approach to finding the similar player for replacement as discussed. These implemented models has to be evaluated with certain evaluation metrics like accuracy, Precision, F1 score, recall, and cross-validation score to ensure that the model is performing well and comparisons can be made between the models are explained in this section as follows.

6.3 Comparision of the Base and Tuned Models in predicting the major 4 positions

Random Forest Classifier performs better in terms of Both Base and Tuned models with the evaluation metrics when compared to all other 3 models in predicting the major 4 positions of the player with higher accuracy of around 87% is shown in Figure 6 including both the evaluation summary table and graphical representations of the evaluation metrics and the performance not greatly improved after the hyperparameter optimization.

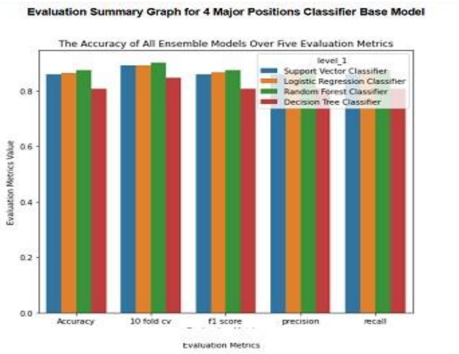


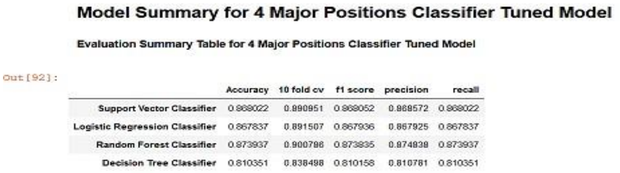
Model Summary for 4 Major Positions Classifier Base Model

Evaluation Summary Table for 4 Major Positions Classifier Base Model

Out[90]:

	Accuracy	10 fold cv	f1 score	precision	recall
Support Vector Classifier	0.860444	0.890951	0.860279	0.861062	0.860444
Logistic Regression Classifier	0.865989	0.891507	0.866097	0.866087	0.865989
Random Forest Classifier	0.873937	0.901095	0.873841	0.874928	0.873937
Decision Tree Classifier	0.808318	0.848580	0.807718	0.809809	0.808318





Model Summary for 4 Major Positions Classifier Tuned Model

Evaluation Summary Table for 4 Major Positions Classifier Tuned Model

Out[92]:

	Accuracy	10 fold cv	f1 score	precision	recall
Support Vector Classifier	0.868022	0.890951	0.868052	0.868572	0.868022
Logistic Regression Classifier	0.867837	0.891507	0.867936	0.867925	0.867837
Random Forest Classifier	0.873937	0.900786	0.873835	0.874838	0.873937
Decision Tree Classifier	0.810351	0.838498	0.810158	0.810781	0.810351

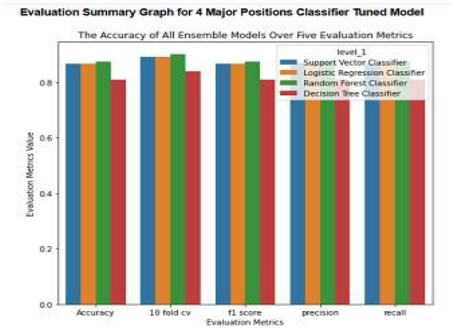
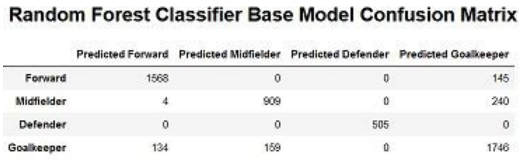


Figure 6: Summary Table and Graphical Representation of all ensemble models predicting 4 major positions

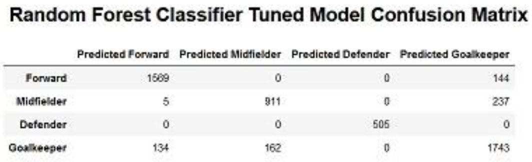
6.4 Comparison of the Best Random Forest Classifier Base and Tuned Models in predicting the 4 Major Positions

Figure 7 below shows the comparison of the random forest base classifier and the random forest tuned classifier in predicting the 4 major positions by discussing the confusion matrix and the heat map. Here, there is not much difference between the base model and the tuned model after selecting the best parameters using hyperparameter techniques. The accuracy remains the same at around 87% for both the base and the tuned model. The confusion matrix in Figure 7 helps to understand how many player positions are truly classified and how many are misclassified and can able to track the misclassification exactly in particular positions.



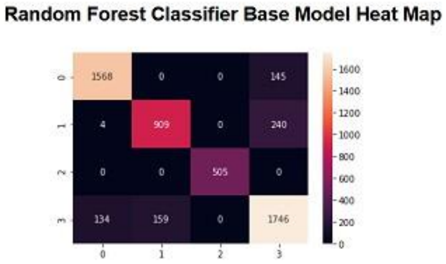
Random Forest Classifier Base Model Confusion Matrix

	Predicted Forward	Predicted Midfielder	Predicted Defender	Predicted Goalkeeper
Forward	1568	0	0	145
Midfielder	4	909	0	240
Defender	0	0	505	0
Goalkeeper	134	159	0	1746

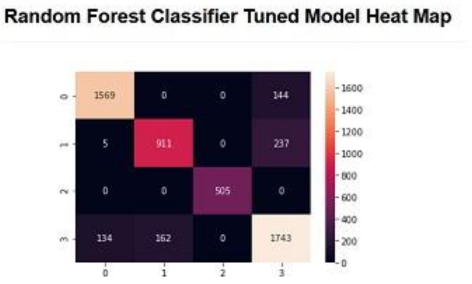


Random Forest Classifier Tuned Model Confusion Matrix

	Predicted Forward	Predicted Midfielder	Predicted Defender	Predicted Goalkeeper
Forward	1569	0	0	144
Midfielder	5	911	0	237
Defender	0	0	505	0
Goalkeeper	134	162	0	1743



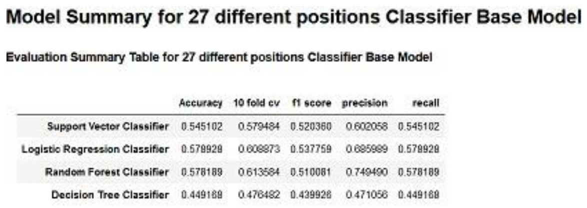
Random Forest Classifier Base Model Heat Map



Random Forest Classifier Tuned Model Heat Map

Figure 7: Confusion matrix and Heat Map for Best Random Forest Classifier predicting 4 major positions

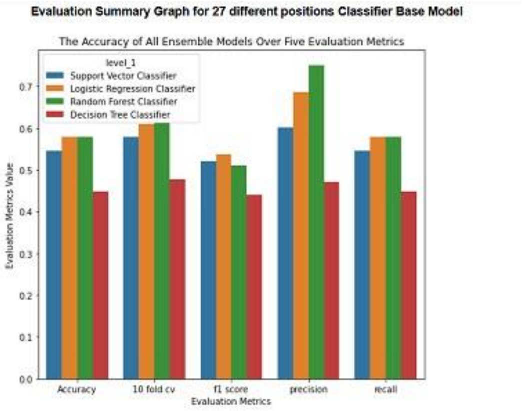
6.5 Comparision of the Base and Tuned Models in predicting the 27 Different Major and Minor positions



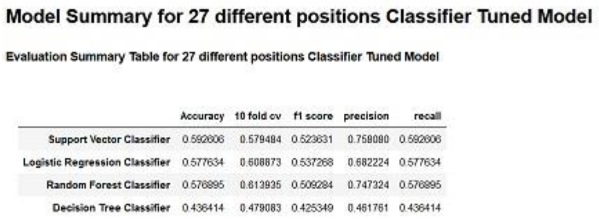
Model Summary for 27 different positions Classifier Base Model

Evaluation Summary Table for 27 different positions Classifier Base Model

	Accuracy	10 fold cv	f1 score	precision	recall
Support Vector Classifier	0.545102	0.579484	0.520360	0.602058	0.545102
Logistic Regression Classifier	0.578928	0.608873	0.537759	0.685989	0.578928
Random Forest Classifier	0.578189	0.613584	0.510081	0.749490	0.578189
Decision Tree Classifier	0.449168	0.476482	0.439926	0.471056	0.449168



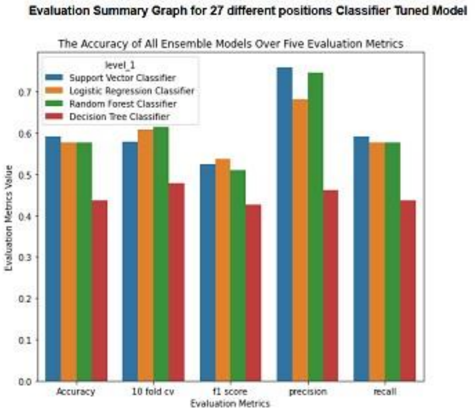
Evaluation Summary Graph for 27 different positions Classifier Base Model



Model Summary for 27 different positions Classifier Tuned Model

Evaluation Summary Table for 27 different positions Classifier Tuned Model

	Accuracy	10 fold cv	f1 score	precision	recall
Support Vector Classifier	0.592606	0.579484	0.523631	0.758080	0.592606
Logistic Regression Classifier	0.577634	0.608873	0.537268	0.682224	0.577634
Random Forest Classifier	0.576895	0.613935	0.509284	0.747324	0.576895
Decision Tree Classifier	0.436414	0.479083	0.425349	0.461761	0.436414

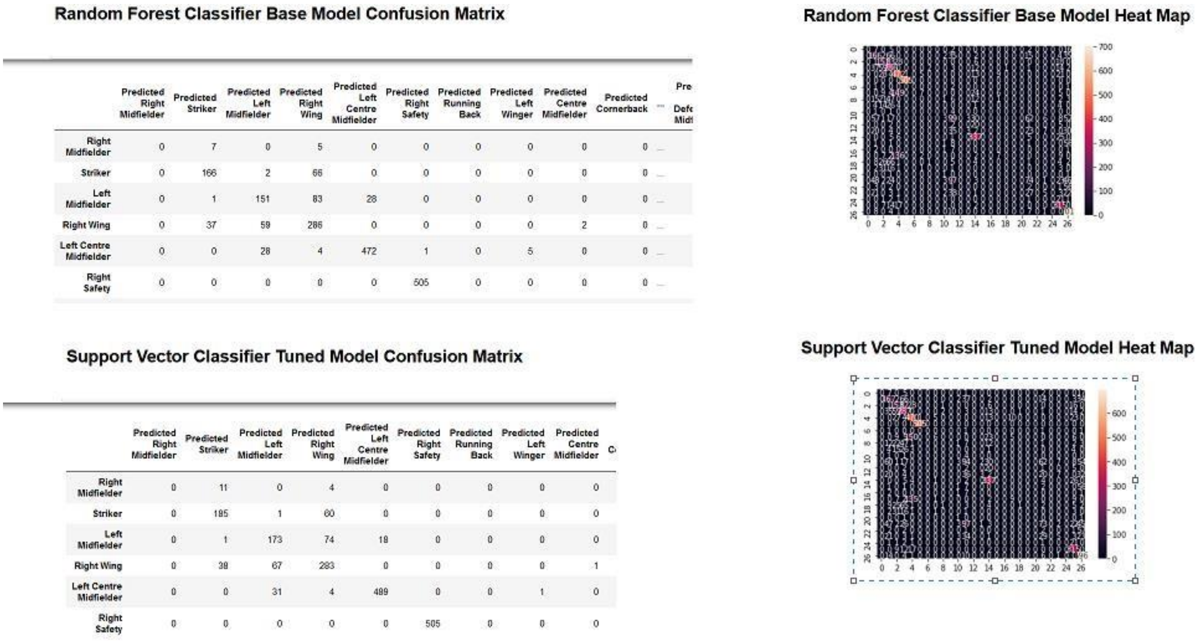


Evaluation Summary Graph for 27 different positions Classifier Tuned Model

Figure 8: Summary Table and Graphical Representation of all ensemble models predicting 27 different positions

Random Forest Base Classifier and Support Vector Tuned Classifier model perform better in evaluating with the evaluation metrics compared to all other 3 models in predicting the 27 different major and minor positions of the player with an accuracy of around 60% are shown inFigure 8 including both the evaluation summary table and graphical representations of the evaluation metrics. The performance of the model remains the same even after attempting to choose the best parameter values for the models using hyperparameter optimization techniques.

6.6 Comparison of the Best Random Forest Classifier Base and Support Vector Classifier Tuned Models in predicting the 27 Different Major and Minor Positions



Random Forest Classifier Base Model Confusion Matrix

	Predicted Right Midfielder	Predicted Striker	Predicted Left Midfielder	Predicted Right Wing	Predicted Left Centre Midfielder	Predicted Right Safety	Predicted Running Back	Predicted Left Winger	Predicted Centre Midfielder	Predicted Cornerback	...	Pre Defe Midf
Right Midfielder	0	7	0	5	0	0	0	0	0	0	...	
Striker	0	166	2	66	0	0	0	0	0	0	...	
Left Midfielder	0	1	151	83	28	0	0	0	0	0	...	
Right Wing	0	37	59	286	0	0	0	0	2	0	...	
Left Centre Midfielder	0	0	28	4	472	1	0	5	0	0	...	
Right Safety	0	0	0	0	0	505	0	0	0	0	...	

Random Forest Classifier Base Model Heat Map

Support Vector Classifier Tuned Model Confusion Matrix

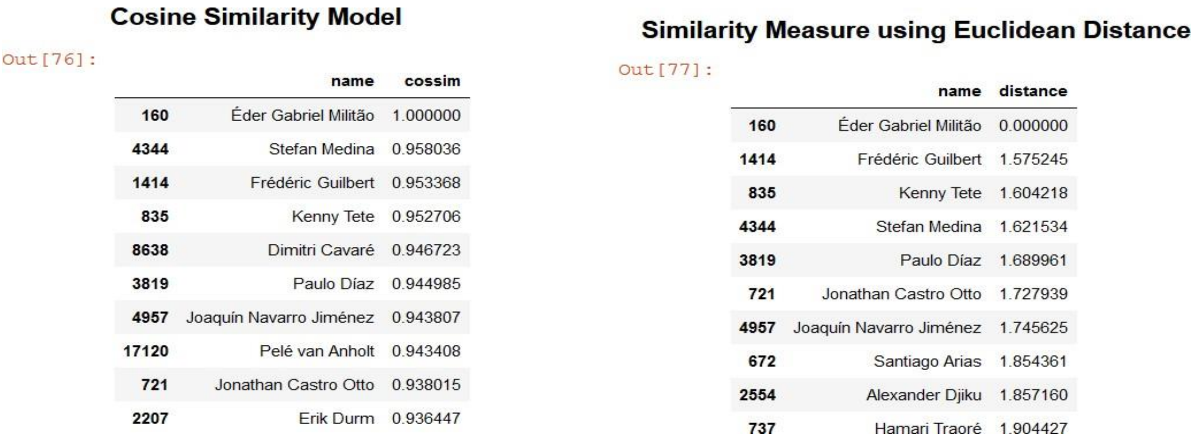
	Predicted Right Midfielder	Predicted Striker	Predicted Left Midfielder	Predicted Right Wing	Predicted Left Centre Midfielder	Predicted Right Safety	Predicted Running Back	Predicted Left Winger	Predicted Centre Midfielder	C
Right Midfielder	0	11	0	4	0	0	0	0	0	
Striker	0	185	1	60	0	0	0	0	0	
Left Midfielder	0	1	173	74	18	0	0	0	0	
Right Wing	0	38	67	283	0	0	0	0	1	
Left Centre Midfielder	0	0	31	4	489	0	0	1	0	
Right Safety	0	0	0	0	0	505	0	0	0	

Support Vector Classifier Tuned Model Heat Map

Figure 9: Confusion matrix and Heat Map for Best Random Forest Classifier Base Model and Support Vector Classifier Tuned Model predicting 27 different positions

Figure 9 shows the comparison of the random forest base classifier and the Support Vector tuned classifier in predicting the 27 different major and minor positions by discussing the confusion matrix and the heat map. Here, there is not much difference between the base model and the tuned model after selecting the best parameters using hyperparameter techniques. The accuracy remains the same at around 60% for both the base and the tuned model. The confusion matrix in Figure 9 helps to understand how many player positions are truly classified and how many are misclassified and can able to track the misclassification exactly in particular positions.

6.7 Comparision of the Cosine Similarity and Euclidean Distance approach in finding the similar players



Cosine Similarity Model

Out[76]:

	name	cossim
160	Éder Gabriel Militão	1.000000
4344	Stefan Medina	0.958036
1414	Frédéric Guilbert	0.953368
835	Kenny Tete	0.952706
8638	Dimitri Cavaré	0.946723
3819	Paulo Díaz	0.944985
4957	Joaquín Navarro Jiménez	0.943807
17120	Pelé van Anholt	0.943408
721	Jonathan Castro Otto	0.938015
2207	Erik Durm	0.936447

Similarity Measure using Euclidean Distance

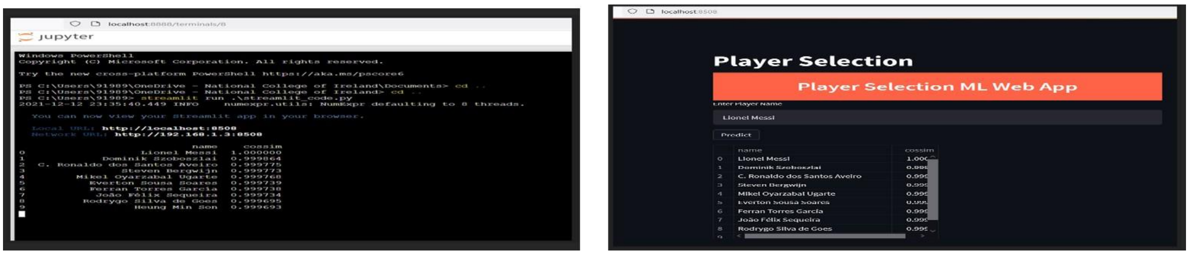
Out[77]:

	name	distance
160	Éder Gabriel Militão	0.000000
1414	Frédéric Guilbert	1.575245
835	Kenny Tete	1.604218
4344	Stefan Medina	1.621534
3819	Paulo Díaz	1.689961
721	Jonathan Castro Otto	1.727939
4957	Joaquín Navarro Jiménez	1.745625
672	Santiago Arias	1.854361
2554	Alexander Djiku	1.857160
737	Hamari Traoré	1.904427

Figure 10: Evaluation Matrix for Cosine Similarity and Euclidean Distance in finding the similar players

Figure 10, Shows the comparison of cosine similarity and Euclidean distance approach in finding similar players. The Euclidean distance approach finds similar players based on a more number of attributes. But, the idea is to get a similar player not based on more number of attributes. But, to find a similar player even if the particular player is low in ratings. Cosine similarity performs better when compared to the euclidean distance approach with the above aspect and helps to find the more similar player and in replacing the particular player.

6.8 Web App Developed Using Stream lit Python Library



```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

PS C:\Users\91989\OneDrive - National College of Ireland\Documents> cd ..
PS C:\Users\91989\OneDrive - National College of Ireland> cd ..
PS C:\Users\91989> streamlit run .\streamlit_code.py
2021-12-12 23:35:40.449 INFO    numexpr.utils: NumExpr defaulting to 8 threads.

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8508
  Network URL: http://192.168.1.3:8508

                         name    cossim
0                Lionel Messi  1.000000
1          Dominik Szoboszlai  0.999864
2  C. Ronaldo dos Santos Aveiro  0.999775
3             Steven Bergwijn  0.999773
4       Mikel Oyarzabal Ugarte  0.999768
5         Everton Sousa Soares  0.999739
6         Ferran Torres García  0.999738
7          João Félix Sequeira  0.999734
8        Rodrygo Silva de Goes  0.999695
9               Heung Min Son  0.999693
```

Player Selection

Player Selection ML Web App

Enter Player Name

Lionel Messi

Predict

	name	cossim
0	Lionel Messi	1.000
1	Dominik Szoboszlai	0.999
2	C. Ronaldo dos Santos Aveiro	0.999
3	Steven Bergwijn	0.999
4	Mikel Oyarzabal Ugarte	0.999
5	Everton Sousa Soares	0.999
6	Ferran Torres García	0.999
7	João Félix Sequeira	0.999
8	Rodrygo Silva de Goes	0.999

Figure 11: Web App Developed Using Stream lit and running the web app in the jupyter notebook terminal

Finally, a web app is created for the cosine similarity approach in finding similar players for replacement based on skills and other attributes. This web app shown in Figure 11 is developed using a stream lit python inbuild library and created a custom function to pass the required attributes from the similarity data frame. Once the user enters the particular player name and by that time custom function developed uses the required attributes from the similarity.csv file and displays the top 10 similar players for the value passed. The Similarity.csv file is just in which the data is transferred after pre-processing to the CSV file from the main similarity measure code. Figure 11 shows how to start up the web app on the local server by running the particular streamlit_code.py file. While running, make sure that all the files like fifa21.csv, similarity.csv, and streamlit_code.py files are in the same current working directory along with all the three .ipynb files.

Once after running the streamlit_code.py file to start up the Player Selection ML web app, the web app appears as shown in Figure 11. The user can enter the player name as required and click predict and it will generate the top 10 similar players based on the skills and other attributes of that player. So, those similar players can be replaced in the team in place of that particular player.

6.9 Discussion

This research aimed to predict the position of the players with the help of machine learning techniques and predict the similar players based on particular skills from the machinelearning models. I have achieved different results and seen that Random Forest gave the best accuracy compared to all other machine learning algorithms in predicting the 4 major positions and the tuned support vector classifier model performs better in predicting the 27 different positions. so throughout the investigation, I have found out that I can predict similar players based on the skills and features had by another player with the help of machine learning techniques. Also, I have found out that the cosine similarity measure is better compared to Euclidean distance as the features related to the players do not lie in the same direction which is why the cosine similarity measure is a better technique compared to the euclidean distance measure. Similar players can be predicted with the help of cosine similarity measures in real-life applications and also the position of the players can be predicted with the Random Forest technique.

I have also performed model tuning in all the algorithms and have seen that there is no huge improvement compared to the base model where I have taken default parameters so I can share that after the application of the best parameters the accuracy of the model is slightlyor not improved in some algorithms.

In this research, deep learning algorithms are not performed to predict the position of the players, and this part is left to future work where I can experiment with deep learning algorithms to see if there is any improvement in accuracy compared to the machine learning models.

7 Conclusion and Future Work

7.1 Conclusion

Football player prediction can be helpful to sports management where the player can beeasily replaced based on the skills and it can be beneficial as algorithmic prediction can be farbetter accuracy compared to the human prediction. This is why machine learning and other approaches had been significantly studied and discussed in the sports domain nowadays to increase the efficiency of hiring different team members based on their skills and other features. In our research, I have performed different machine learning algorithms to predict the position of the players and also perform similarity measures to find similar players based on their skills and other features. I have found out that the random forest model performs better in predicting the 4 major positions and the tuned support vector classifier model performs better in predicting the 27 different positions compared to four other algorithms and cosine the similarity measure is found to be a better technique to find out similar players based on skills.

I have also explored the data and found that the size of the data is quite large which is why I have performed a feature selection technique to select the top 30 features to predict the position of the players. Since the data is imbalanced all the classes do not contain the same number of attributes which can be problematic as imbalanced data do not give good accuracy while training. This is why I have performed the ADASYN algorithm to create synthetic points to balance the data in all the classes so that an equal number of attributes can be trained in the model to give the desired output.

Our research topic was to find out if machine learning techniques are efficient enough to predict the position of the players based on skills and other features and I have found out that machine learning techniques are beneficial in predicting the position of the Players as they are nearly 90% accurate in predicting the position of the players and can save a lot of time compared to manual techniques. This technique can also be beneficial as they can process a large number of data related to different players and can give prediction results in no time. This is why machine learning and approaches are used by many domain experts as they save a lot of time and give accurate results which are not possible to be processed by manual approaches.

The model tuning did not give better accuracy compared to the basic model with default parameters so I can say that model tuning is not working in machine learning algorithms but can be helpful in other future work such as deep learning algorithms which can be studied with hyperparameter tuning.

In the literature review section, I have some researchers using other machine learning approaches and got different accuracies from which I can say that previous studies are successful enough in predicting the position of the players based on skills and other features are also different types of data differs in predictions as they contain different features that may or may not be important enough to predict suitable player based on skills and features.
The distance similarity matrix is studied to find the difference between cosine similarity measure and euclidean measure and I have found that cosine similarity measure is beneficial compared to euclidean distance as they can easily find the similar points that are not collinear. The realistic data do not contain such features which are collinear and are very rare which is why I can conclude in our research that the cosine similarity matrix is the best metric to find similar players based on skills and features.

7.2 Future Work

In this research, only machine learning algorithms are studied and various other deep learning approaches tend to give better results in this type of data which is left unstudied in this research. So the future work includes the implementation of different deep learning algorithms to predict the position of the players concerning 27 different positions and also balance the data by adding more number data if possible and adding other features that might be beneficial to predict the position of the players in a much better way.

Also, the model can be deployed in real-life applications with the help of various software and tools so that domain experts and sports analysts can easily predict the position of the player by using the interface that can predict similar players based on the skills in a much easier way.

This research was beneficial in predicting the position of the player with the help of machine learning techniques. What I did not achieve is a better prediction above 90% and that can be achieved with the help of various deep learning algorithms such as Artificial Neural Network, Convolution Neural Networks, and other techniques. These techniques are proven beneficial in predicting data that contain both categorical and numerical attributes which is why this is a must future work that can be decided by different domain experts so that the prediction power of a player increases with the help of such techniques.

Acknowledgment

I would first want to thank my thesis Supervisor Mr. Aaloka Anant and the NCI School of the computing department for their constant support throughout this research project during this covid tough phase of life. I extend my sincere thanks to all professors in NCI for their knowledge sharing all time and finally want to thank my family members, and friends who are always encouraging me in this research project.

References

Abreu, D. and Sannikov, Y. (2014). An algorithm for two-player repeated games with perfect monitoring. *Theoretical Economics*, 9(2), pp.313–338.

Baerg, A. (2016). Big Data, Sport, and the Digital Divide. *Journal of Sport and Social Issues*, 41(1), pp.3–20.

Berrar, D., Lopes, P. and Dubitzky, W. (2018). Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine Learning*, 108(1), pp.97–126.

Betasolo, M.L. (2016). Culture – Educational Paradigm Shift Learning Methodologies Derived from Axiomatic Design Principle. *Procedia CIRP*, 53, pp.179–186.

Bolturk, E. and Kahraman, C. (2018). A novel interval-valued neutrosophic AHP with cosine similarity measure. *Soft Computing*, 22(15), pp.4941–4958.

Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, 81(3-4), pp.429–450.

Chen, H. (2019). Neural Network Algorithm in Predicting Football Match Outcome Based on Player Ability Index. *Advances in Physical Education*, 09(04), pp.215–222.

Cho, Y., Yoon, J. and Lee, S. (2018). Using social network analysis and gradient boosting to develop a soccer win–lose prediction model. *Engineering Applications of Artificial Intelligence*, 72, pp.228–240.

F.I., A. and J.C, O. (2015). English Premier League (EPL) Soccer Matches Prediction using An Adaptive Neuro-Fuzzy Inference System (ANFIS). *Transactions on Machine Learning and Artificial Intelligence*.

Falch, T.L. and Elster, A.C. (2016). Machine learning-based auto-tuning for enhanced performance portability of OpenCL applications. *Concurrency and Computation: Practice and Experience*, 29(8), p.e4029.

Fan, D., Kim, H., Kim, J., Liu, Y. and Huang, Q. (2019). Multi-Task Learning Using Task Dependencies for Face Attributes Prediction. *Applied Sciences*, 9(12), p.2535.

Guan, S. and Wang, X. (2021). Optimization analysis of football match prediction model based on neural network. *Neural Computing and Applications*.

Hari Chandana, Ch. and Bala Krishna, G. (2021). Breast cancer detection using random forest classifier. *Materials Today: Proceedings*.

Haugen, K.K. (2017). Equilibrium team selection in football under win or profit maximisation. *Mathematics for Application*, 6(2), pp.161–170.

Iso, S., Ishitsuka, K., Onishi, K. and Matsuoka, T. (2019). GPR data interpretation by the deep learning with coloring data. *BUTSURI-TANSA(Geophysical Exploration)*, 72(0), pp.68–77.

Jayalath, K.P. (2018). A machine learning approach to analyze ODI cricket predictors. *Journal of Sports Analytics*, 4(1), pp.73–84.

Kerns, L.X. (2017). Simultaneous confidence bands for log-logistic regression with applications in risk assessment. *Biometrical Journal*, 59(3), pp.420–429.

Kumar, P., Kumar Sehgal, V. and Singh Chauhan, D. (2011). Knowledge Discovery in Databases KDD with Images: A Novel Approach toward Image Mining and Processing. *International Journal of Computer Applications*, 27(6), pp.10–13.

Lee, H., Jeong, Y. and Kim, S. (2016). Induced Rule-Based Fuzzy Inference System from Support Vector Machine Classifier for Anomalous Propagation Echo Detection. *International Journal of Machine Learning and Computing*, 6(2), pp.92–96.

Ma, L. and Ladisch, M. (2019). Evaluation complacency or evaluation inertia? A study of evaluative metrics and research practices in Irish universities. *Research Evaluation*.

Ozceylan, E.O. (2016). A MATHEMATICAL MODEL USING AHP PRIORITIES FOR SOCCER PLAYER SELECTION: A CASE STUDY. *South African Journal of Industrial Engineering*, 27(2).

PANAIT, C. and BUCINSCHI, V. (2018). MYERS-BRIGGS TYPE INDICATOR INFLUENCE IN TEAM BUILDINGS. *Review of the Air Force Academy*, 16(1), pp.89–94.

Qingwen, T. (2020). Football Player Performance Prediction Based on Combined Kernel Function Correlation Vector Machine. *Dynamic Systems and Applications*, 29(5).

Rahayu, S., Bharata Adji, T. and Akhmad Setiawan, N. (2017). Penghitungan k-NN pada Adaptive Synthetic-Nominal (ADASYN-N) dan Adaptive Synthetic-kNN (ADASYN-kNN) untuk Data Nominal-Multi Kategori. *Jurnal Otomasi Kontrol dan Instrumentasi*, 9(2), p.119.

S., S. and D., D.S. (2018). Rule Generation for Gallbladder Cancer Prediction Using Decision Tree Classifier. *Bonfring International Journal of Data Mining*, 8(1), pp.01–03.

Shajila, R. and Vimala, S. (2017). Graceful Labelling for Complete Bipartite Fuzzy Graphs. *British Journal of Mathematics & Computer Science*, 22(2), pp.1–9.

Sharma, M., Mittal, P., Garg, N. and Jain, P. (2019). Data Analysis FIFA World Cup Data Set. *Indian Journal of Science and Technology*, 12(39), pp.1–4.

Sheikh, Y. (2017). Effective Feature Selection for Feature Possessing Group Structure. *International Journal Of Engineering And Computer Science*.

Shim, I., Oh, T.-H. and Kweon, I. (2018). High-Fidelity Depth Upsampling Using the Self-Learning Framework. *Sensors*, 19(1), p.81.

Strnad, D., Nerat, A. and Kohek, Š. (2015). Neural network models for group behavior prediction: a case of soccer match attendance. *Neural Computing and Applications*, 28(2), pp.287–300.

Vasiliu, D., Dey, T. and Dryden, I.L. (2018). Penalized Euclidean distance regression. *Stat*, 7(1), p.e175.

Zaini, M.Z. and Salimin, N. (2020). Coaching Games for Understanding (CGfU) Module for Invasion Games in Football. *International Journal of Academic Research in Business and Social Sciences*, 10(4).

Zhang, Z. (2017). Text Feature Selection based on Feature Dispersion Degree and Feature Concentration Degree. *International Journal of Performability Engineering*.