Perspective

FOOTBALL'S BEST PLAYING ELEVEN

Khwaab Thareja (18BCE0930), Prathik Pugazhenthi (18BCE0940), Rahul Anand (18BCE0953)

[a] *Department of Computer Science Engineering, Vellore Institute Of Technology, Vellore*

A R T I C L E I N F O

Article history:
Review: 22-05-2021

Keywords:

Data-visualization
sports
football
prediction
statistics
analysis
fantasy football
machine learning
soccer
playing 11

A B S T R A C T

Team formation is regarded as a high-level tactic in any team sport. Football is a team game, but at the same time, we cannot ignore the performance of an individual player. It assigns players different responsibilities and indicates their active regions of play on the pitch, consequently enhancing the performance of the team greatly. Detailed analysis of team formations in football has become particularly indispensable for soccer fans and analysts. So, every player has a different responsibility, active position, etc. However, we aim to analyze each player based on their skill set, so we can compare the players on similarities and find out the best player suitable for that particular position. Our major aim is to figure out the best playing 11 with the help of data visualization.

1. Introduction

Team sports are described as groups that interact in a dynamic and interdependent manner. To Understand the individual role of each player in the process is highly relevant to uncover how any team operates. The collective behaviour within the team is frequently linked to performance outcomes in sports, the impact of individual players on the overall team performance requires further research. The basic aim of this project is to analyse all players of a particular category like Goalkeeping or Centre Forward based on different factors. The factors used are for e.g.: games played, age, proficiency, sweeping ability, etc. Based on the above values given in the dataset we will be visualizing the data and generating relevant graphs to find out the best player who can play for the team at a certain position. For doing the above-mentioned data visualization we will be using python and certain libraries mentioned ahead in the document. Football is the most watched and followed sport in the world it would be a great idea to actually try to predict the game results in advance! If in case the result is negative there is scope to analyse the team and go for some improvements in order to alter the result to the greatest extent possible. Also, we know that this is not possible to do in reality so we are implementing it on a football mobile game called Dream League Soccer.

Results in competitive gaming are based upon a lot of micro factors. Victory of players can be attributed largely to skill but also to careful selection of correct stats of objects and characters presented. In Dream League Soccer, players compete virtually by means of creating their ideal football teams which have players based upon real world statistics of football players around the world. This project tries to create a system to help classify and choose players by means of visual representation and help choose the best player for each role.

Till now, the individual roles of players in the game of soccer have only been studied at a match-level that considers the overall passing/other metric of a team i.e., only the overall team performance is considered. But so as to consider the structure of the game, a different approach is needed. This method helps in reflecting the actual interplay. There has to be a distinction between certain metric/characteristics, this is important to qualify for the different interaction phases. It is very difficult to calculate a well-known network metrics for the prediction, such as betweenness on play-level, so new adequate metrics are needed to predict accurately. The flow betweenness is introduced as a new indicator on play-level and is computed along with the flow centrality. It would be an added advantage to the overall performance of the team if we are able to predict which player is best in what position of the game. This will make sure that the team is not made on a hit and trial basis, this would focus on pure statistics.

Computer systems support several coaching activities, such as strategy development and performance analysis. It needs real-time interpretation of information. Only recently has better technology enabled systems to at the same time monitor both the position of players and the ball. Consequently, period of time game analysis systems should be ready to automatically acknowledge intentional activities in an exceedingly multiagent system. Our project acquires the player skills, and confirm strengths and weaknesses of team and the player. several approaches that analyse and predict results of international matches in soccer are related to statistical models incorporating many probably important covariates with regard to a national team's success, like the bookmakers' ratings or the Dream League football ranking. supported all matches from the four previous soccer games from 2000–2016, we compare the foremost common regression models that are related to the teams' covariate info with relation to their predictive performances.

A common machine learning tasks, that focuses on predicting a target variable in any previously unseen information, is classification. In general, this method is used to predict a target variable which is mostly the class variable of the dataset. This is done by building a classification model with the inputs of the train date, and then test that particular model to predict the value of the class category of the input/test information. this sort of information process is named supervised learning since the info process part is guided toward the category variable while building the model. we might first attempt to analyse individual players based on their sweeping, kicking and passing ability. we would attempt to understand the player's strong zone on the playing area using the idea of heatmap. The passing ability of players would be analysed by using line graph and displaying the successful and unsuccessful passes by blue and green line respectively

Later we would take a dataset consisting of players who play for a team and then try to visualize the entire team on the basis of their position like Goalkeeping, Forward, Midfielder and Defender, by creating variables that are derived from various attributes mentioned in the dataset. For example, new variables like Goal Stopping, Long Pass and Sweeping would be made using the dataset attributes and calculating their mean average and then using these variables the players would be sorted and then the top 5 for each position. The result would be in the form of visual plots like pie chart and bar graphs. As the end result, we would try to mention down all the players who should be included in the top 11.

Related Work

This section reviews works previously done in predicting a better side in the game.

[1] Uses player and team data collected from various web sources, and propose a professional program. It also included key analysis, non-parametric mathematical analysis, support vector support (SVM), and a computational learning algorithm to predict whether the team will win the game. Integration methods were improved in the SVM reference classification, and the accuracy of modelling estimates by integration exceeds 72%. Comparisons of multi-machine learning modes determine the chances of improving the accuracy of a game outcome prediction. The system evaluates model results, and addresses the challenges and concerns that exist in predicting a winning hockey game. **[2]** Demonstrates a standard prediction model for predicting English Premier League results. Using feature engineering and test data analysis, creating a factor in determining the most important factors for predicting the outcome of a football game, and as a result create a more accurate prediction system using machine learning. **[3]** Performs a detailed analysis of soccer matches and player performances and identifies the most important aspects of player performance that determine professional ratings and develop a series of sections with three different techniques. the most important performance metrics in the four gaming areas that measure the qualifications considered by an expert while giving

ratings is calculated and then find the most influential performance metrics for players to determine the outcome of a game and assess the extent to which the effect is reflected in the performance characteristics of the players and got around 74.12% prediction accuracy. [4] Explores the various methods of machine learning to predict the scores and results of football matches, using game events rather than the number of goals scored by each team and test various model design concepts and evaluate the performance of the models by looking. [5] Predicts the outcome of match given the home team and away team. Predictions are made based on a variety of important indicators that contain data from previous Premier League seasons. These important attributes may determine the outcome of a game. For predictive purposes, three different algorithms are used namely, Logistic Regression, XGBoost and Support Vector Machines and select the best of the three algorithms to predict that appropriate label. [6] Provides a critical measure of the use of machine learning in football related to offensive play.

Problem Statement

Predicting the best playing 11 in football and whether the home team would win or not in an effective way.

Proposed Work

a) Player Analysis
We have taken a dataset that was basically a JSON file, which had information and data about a particular match. Here instead of having information in the form of matrix, which would cause a big chaos because at different timestamps we have different players at one particular position. So, we have stored data in the form of objects which basically represents collection of possessions that occurred in the match. This will help us analyze every single pass between the same team members and also between opponents. In order to analyze the passing ability of the player we draw a pitch using matplotlib. In this, we will try to analyze various visualizations. Later, a single player is extracted and is analyzed on a single attribute present in the dataset for example passing. By this we have an output that shows the players pass in 2 different colors based on whether the pass was successful or not. Also, in order to analyze which, place in the field is most visited by the player. Lastly, we create a function that takes the players name and the dataset as parameters and give an output of a field that visualizes both the passes and the heat-map.

b) Prediction

To predict whether a home team will win on its home ground or not, we have taken various datasets from years 2000-2016 and we have combined these datasets into a single dataset which will be used for prediction.
After Pre-processing the data, we have removed the data of the first 3 match weeks for each year because we will also consider the results of the last 3 matches that the home team plays to predict.

	FTR	HM1	HM2	HM3	AM1	AM2	AM3	DiffLP	FTHG	FTAG
0	NH	M	M	M	M	M	M	-7.0	0	0
1	NH	M	M	M	M	M	M	8.0	2	3
2	NH	M	M	M	M	M	M	6.0	2	2
3	H	M	M	M	M	M	M	-3.0	4	1
4	H	M	M	M	M	M	M	-13.0	3	0
...	...	...	...	...	...	...	...	...	...	...
5995	H	D	L	L	W	L	W	-4.0	3	2
5996	H	W	W	L	L	W	L	6.0	2	0
5997	NH	L	W	W	W	L	W	10.0	1	2
5998	H	W	W	W	W	D	D	2.0	3	1
5999	H	L	L	D	W	D	W	11.0	3	2

Fig 1 Dataset after Pre-processing

Then we converted the categorical attributes into binary values.

Feature values:

	HM1_D	HM1_L	HM1_M	HM1_W	HM2_D	HM2_L	HM2_M	HM2_W	HM3_D	HM3_L	...	AM2_L	AM2_M	AM2_W	AM3_D	AM3_L	AM3_M	AM3_W	DiffLP	FTHG	FTAG
0	0	0	1	0	0	0	1	0	0	0	...	0	1	0	0	0	1	0	-7.0	0	0
1	0	0	1	0	0	0	1	0	0	0	...	0	1	0	0	0	1	0	8.0	2	3
2	0	0	1	0	0	0	1	0	0	0	...	0	1	0	0	0	1	0	6.0	2	2
3	0	0	1	0	0	0	1	0	0	0	...	0	1	0	0	0	1	0	-3.0	4	1
4	0	0	1	0	0	0	1	0	0	0	...	0	1	0	0	0	1	0	-13.0	3	0

Fig 2 Dataset with binary values

```
Total number of matches: 6000
Number of features: 9
Number of matches won by home team: 2787
Win rate of home team: 46.45%
```

Fig 3 Home team win/lose prediction dataset

c) Technique

we have used three different classifiers- Logistic Regression, Random Forest and XGBClassifier and compared their accuracy. Random Forest Classifier tops the list with an accuracy of 98.30% and we got the F1 score of the model as 0.9833.

d)Method

We analyzed which country has the highest number of players that are into soccer. Using the counterplot function, we plot a pie chart and figure out the result. Here we find out that England is most active in the game of football. In order to achieve the aim of finding out the best playing 11, we try to get a dataset that has every player information and consist of 26 attributes that are essential for the comparison of players for example passing, etc. With these attributes we try to generate new variables depending on the position for which we are trying to find the best player for. For example to find the best goalkeeper, we create 2 variables namely shot_stopper and sweeping which are formed by taking the weighted average of all the attributes and giving more weightage to GK_reflexes and GK_kicking. After calculating the 2 variables we try to sort it in the descending order and then predict the top 5 by specifying the type of player required. The graphs are plotted in the form of bar graphs and pie charts. By using different such variables for forward, midfielder and defenders we try to analyse what skills are required for different positions like Centre forward, Left Cornerback, Left Winger etc. We plot different graphs for these variables and find out the best players for every position.

Methodology Implementation

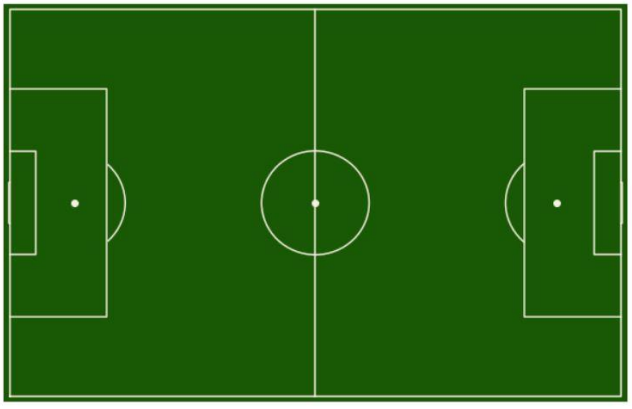


Fig 4 football pitch with dimensions 120m X 80m using Matplotlib.

We have used three classifiers -
I. Logistic Regression

II. XGB Classifier

III. Random Forest Classifier.

The highest accuracy score is **98.30%** on **Random Forest Classifier**.

The highest F1 score was attained by **Random Forest Classifier** which was **0.9833**.

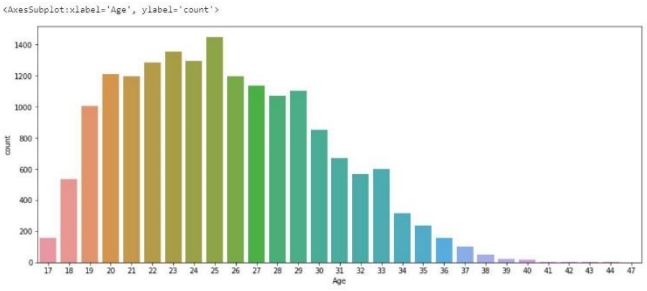


Fig 5 Plot representing number of players playing at a particular age.

From the above bar graph, we can see that the graph follows the maxwell curve and the peak is at 25 years i.e., maximum players are of the age 25 years.

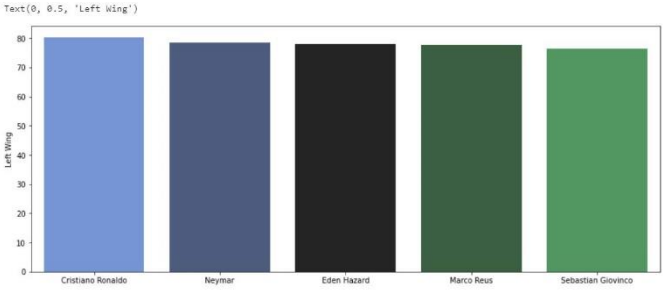


Fig 6 Plot representing the top 5 players for left wing.

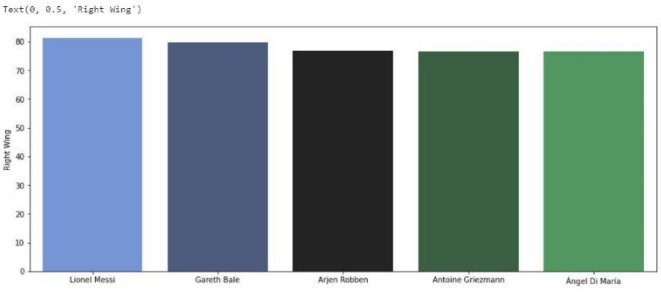


Fig 7 Plot representing the top 5 players for right wing.

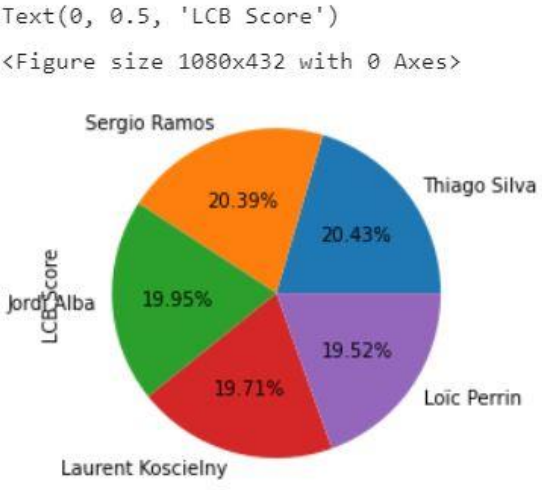


Fig 8 Plot representing the top 5 players for left center back.

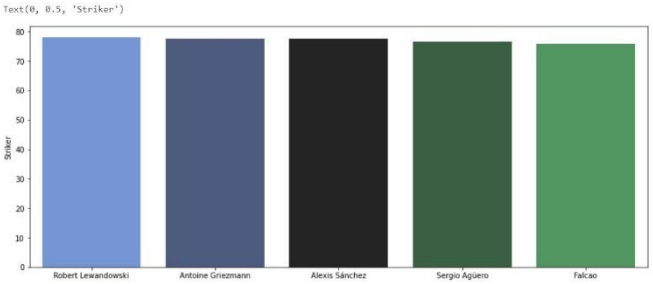


Fig 9 Plot representing the top 5 players for Striker.

Comparison with existing methods

We are using Machine Learning algorithms to predict the best playing 11. We are using data objects instead of data cells that is in general used by the previous methods. Data objects hold information wrt time stamps which helps in visualizing better. These data objects are used to analyze the ball possession. We have used Random forest classifiers which gives a better result compared to the previously used methods like XGBClassifier, SVM, Logistic Regression. We have tried to improve the prediction accuracy to above 90%. The boost in the prediction accuracy is mainly due to the change in the format of data input and the model.

```
Training a RandomForestClassifier using a training set size of 5940. . .
Trained model in 0.4177 seconds
Made predictions in 0.0126 seconds.
F1 score and accuracy score for test set: 0.9818 , 0.9833.
```

Fig 12 Prediction accuracy and F1 score of the three models.

Additionally, we try to create dependent variables from already present independentattributes for getting better insights of the data. For e.g., to predict the best goalkeeper we create attributes like shot_stopper and sweeping which are formed by taking the weighted average of all the attributes and giving more weightage to GK_reflexes and GK_kicking.

Result

We analyze player attributes for different positions like striker, defender, etc. And finally get a list of players for our playing 11 along with substitutes. This is visualized with the name of the player in that particular position.

Fig 10 Final plot representing the best playing 11 wrt their positions.

Conclusion and Future work

The focus of this study lies in an analysis at play-level. this implies that interaction in every ball possession is examined rather than evaluating an aggregative passing matrix at match-level. The field positions are evaluated to facilitate an analysis of the individual contribution of players in our study. Multiple players can also be allotted to the same position. Average metric values are reported to gauge the performance of the playing positions in this case. the ultimate classification is in line with the previous studies that specialize in players in soccer. we found the most effective team attainable for every position. There was an in-depth visualization of all players using different parameters like ages, heights, weights, agility. The visualization enabled us to show and compare athletes based on completely different parameters dynamically. For future work we've a far larger dataset than we've and better machine learning algorithms to induce far more correct, precise and quick results

References

[P1] gu, w., foster, k., shang, j., & wei, l. a game-predicting expert system using big data and machine learning

[P2] baboota rahul, kaur harleen. predictive analysis and modelling football results using machine learning approach for english premier league

[P3] gunjan kumar. machine learning for soccer analytics corentin herbinet. predicting football results using machine learning techniques

[P4] sushant ,deepanshu rana . premier league match result prediction using machine learning

[P5] mat harold, floris goes, stephan nopps. machine learning in men's professional football: current applications and future directions for improving attacking play

[P6] sayed muhammad yonus saiedy,muhammad aslam hemmat qachmas,amanullah faqiri predicting epl football matches results using machine learning algorithms

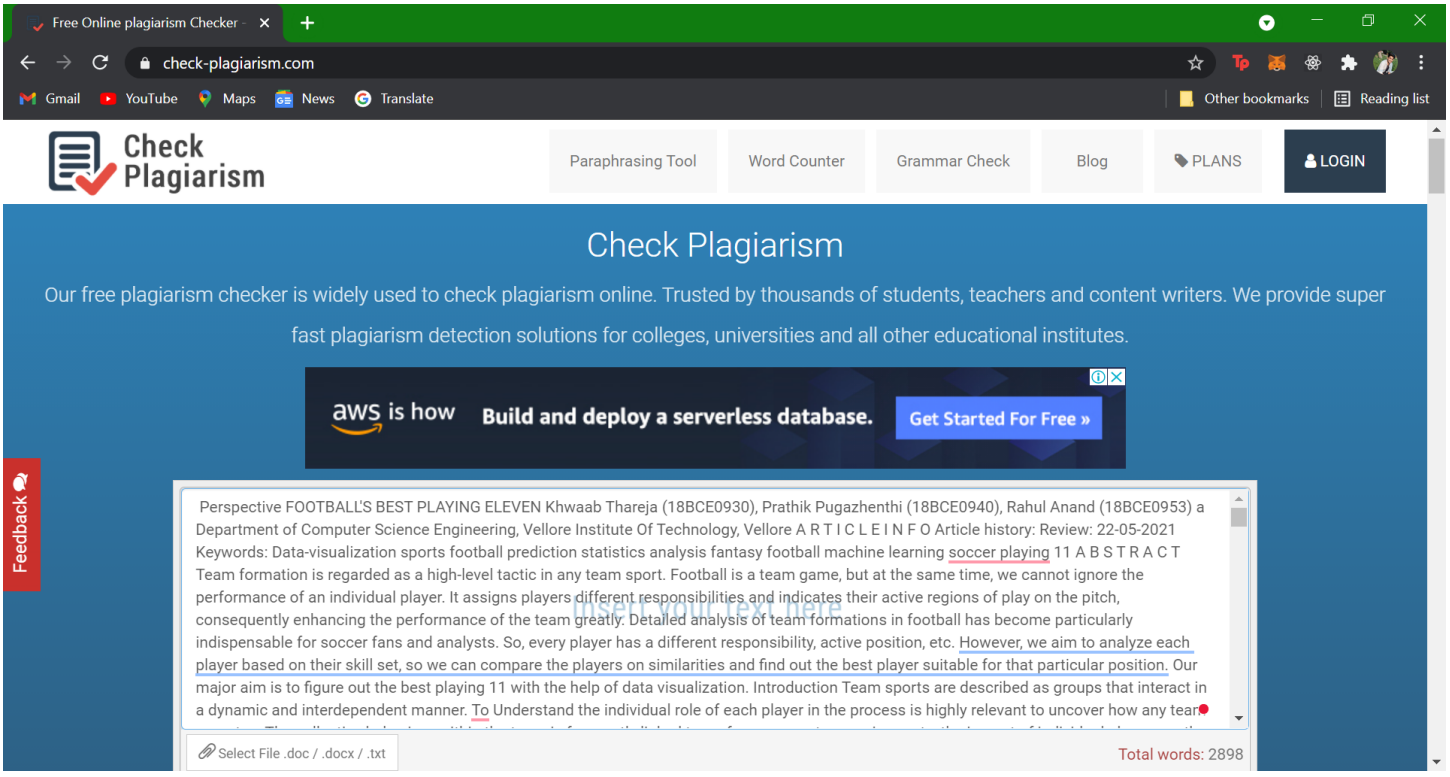
[P7] ayush tripathi[1*], rashidul islam[2], vatsal khandor[3], vijayabharathi murugan[4] prediction of ipl matches using machine learning while tackling ambiguity in results

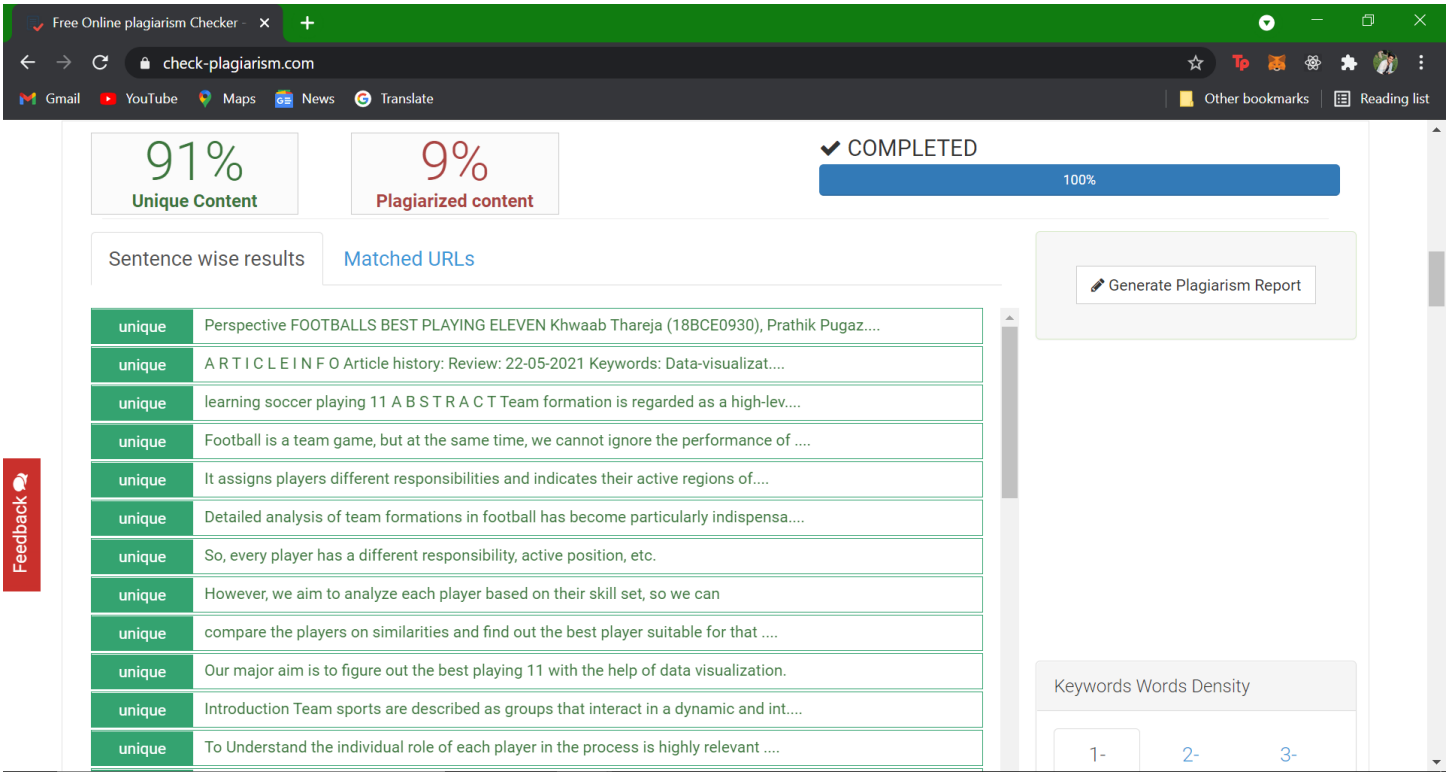
[P8] tarak kharrat, javier lópez peña, i.g. mchale.plus-minus player ratings for soccer

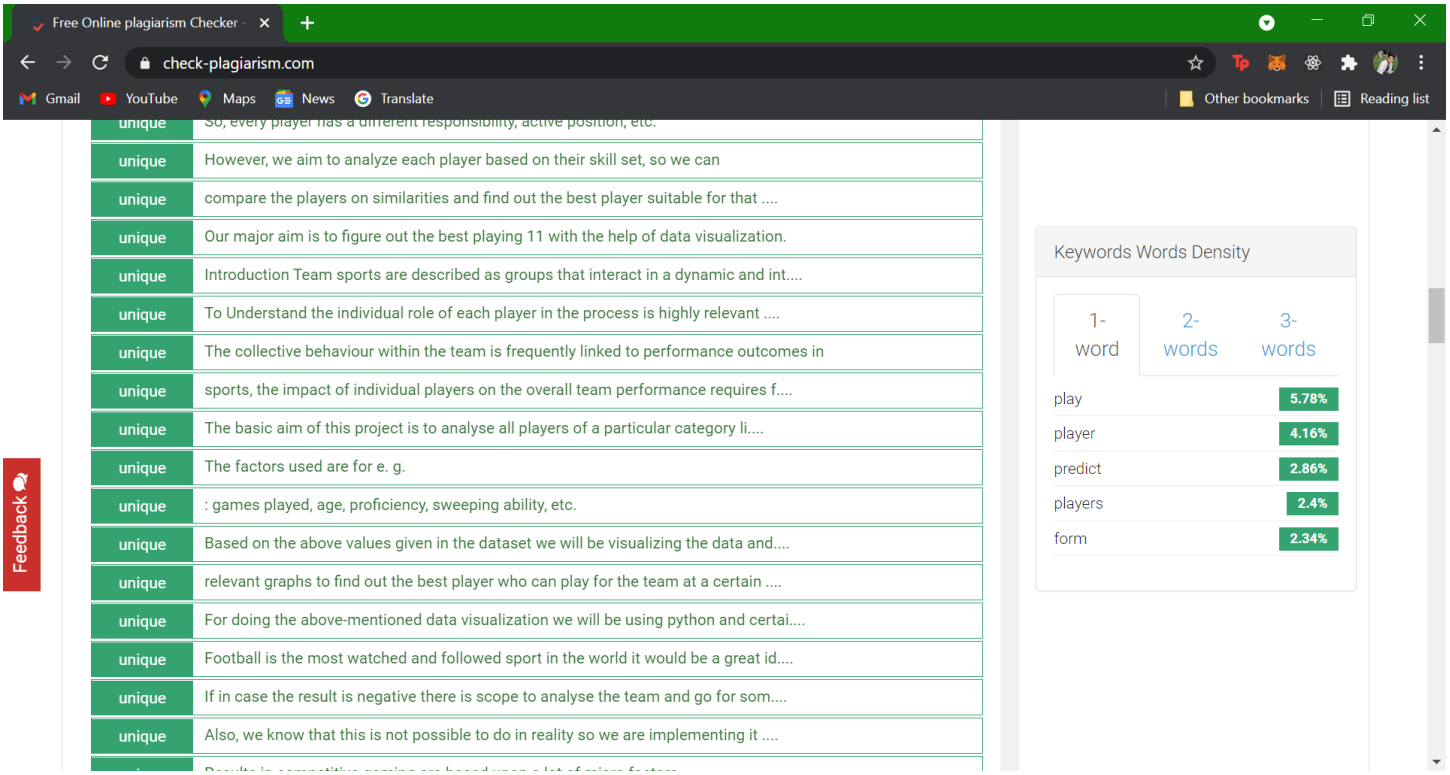
[P9] i.g. mchale, phil scarf, david e. folker
on the development of a soccer player performance rating system for the english premier league

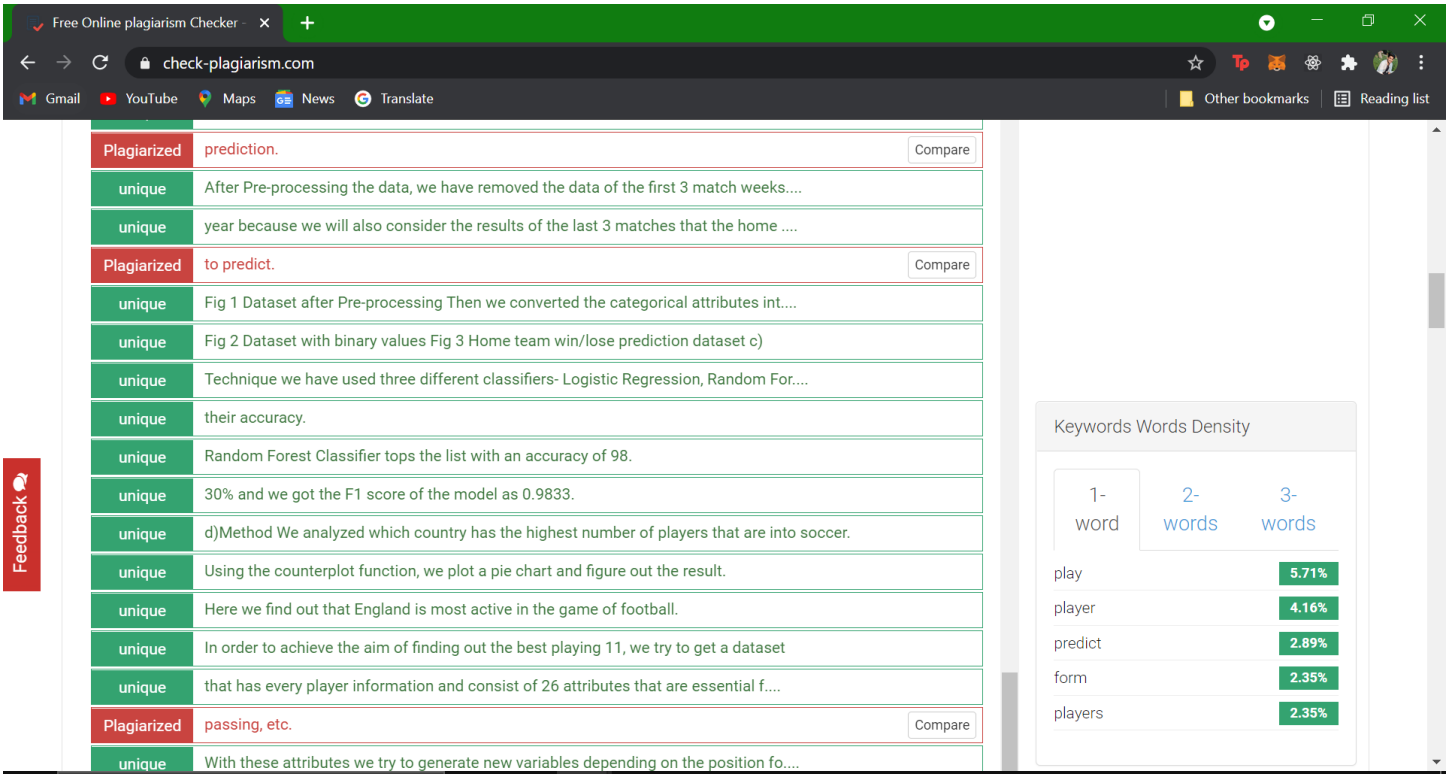
[P10] halvard arntzen, lars magnus hvattum predicting match outcomes in association football using team ratings and player ratings

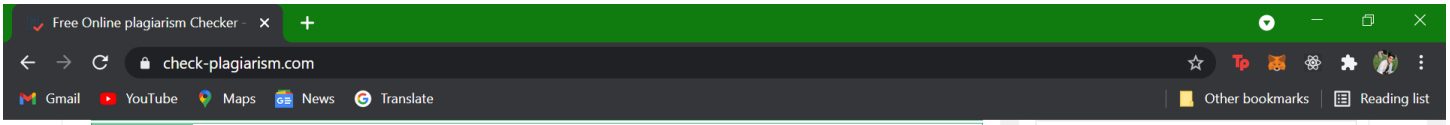
PLAGIARISM REPORT






Free Online plagiarism Checker
check-plagiarism.com
Gmail
YouTube
Maps
News
Translate
Other bookmarks
Reading list
unique
So, every player has a different responsibility, active position, etc.
unique
However, we aim to analyze each player based on their skill set, so we can
unique
compare the players on similarities and find out the best player suitable for that ....
unique
Our major aim is to figure out the best playing 11 with the help of data visualization.
unique
Introduction Team sports are described as groups that interact in a dynamic and int....
unique
To Understand the individual role of each player in the process is highly relevant ....
unique
The collective behaviour within the team is frequently linked to performance outcomes in
unique
sports, the impact of individual players on the overall team performance requires f....
unique
The basic aim of this project is to analyse all players of a particular category li....
unique
The factors used are for e. g.
unique
: games played, age, proficiency, sweeping ability, etc.
unique
Based on the above values given in the dataset we will be visualizing the data and....
unique
relevant graphs to find out the best player who can play for the team at a certain ....
unique
For doing the above-mentioned data visualization we will be using python and certai....
unique
Football is the most watched and followed sport in the world it would be a great id....
unique
If in case the result is negative there is scope to analyse the team and go for som....
unique
Also, we know that this is not possible to do in reality so we are implementing it ....
Feedback
Keywords Words Density
1-word
2-words
3-words
play 5.78%
player 4.16%
predict 2.86%
players 2.4%
form 2.34%


Free Online plagiarism Checker
check-plagiarism.com
Gmail
YouTube
Maps
News
Translate
Other bookmarks
Reading list
Plagiarized
prediction.
Compare
unique
After Pre-processing the data, we have removed the data of the first 3 match weeks....
unique
year because we will also consider the results of the last 3 matches that the home ....
Plagiarized
to predict.
Compare
unique
Fig 1 Dataset after Pre-processing Then we converted the categorical attributes int....
unique
Fig 2 Dataset with binary values Fig 3 Home team win/lose prediction dataset c)
unique
Technique we have used three different classifiers- Logistic Regression, Random For....
unique
their accuracy.
unique
Random Forest Classifier tops the list with an accuracy of 98.
unique
30% and we got the F1 score of the model as 0.9833.
unique
d)Method We analyzed which country has the highest number of players that are into soccer.
unique
Using the counterplot function, we plot a pie chart and figure out the result.
unique
Here we find out that England is most active in the game of football.
unique
In order to achieve the aim of finding out the best playing 11, we try to get a dataset
unique
that has every player information and consist of 26 attributes that are essential f....
Plagiarized
passing, etc.
Compare
unique
With these attributes we try to generate new variables depending on the position fo....
Feedback
Keywords Words Density
1-word
2-words
3-words
play 5.71%
player 4.16%
predict 2.89%
form 2.35%
players 2.35%



unique	In Dream League Soccer, players compete virtually by means of creating their ideal....
unique	teams which have players based upon real world statistics of football players aroun....
unique	This project tries to create a system to help classify and choose players by means ....
unique	Till now, the individual roles of players in the game of soccer have only been stud....
unique	, only the overall team performance is considered.
unique	But so as to consider the structure of the game, a different approach is needed i. e.
unique	, a play-by-play network analysis.
unique	This method helps in reflecting the actual interplay.
unique	There has to be a distinction between certain metric/characteristics, this is impor....
unique	It is very difficult to calculate a well-known network metrics for the prediction,
unique	such as betweenness on play-level, so new adequate metrics are needed to predict ac....
unique	The flow betweenness is introduced as a new indicator on play-level and is computed....
unique	It would be an added advantage to the overall performance of the team if we are abl....
unique	This will make sure that the team is not made on a hit and trial basis, this would ....
unique	Computer systems support several coaching activities, such as strategy development ....
unique	It needs real-time interpretation of information.
unique	Only recently has better technology enabled systems to at the same time monitor bot....

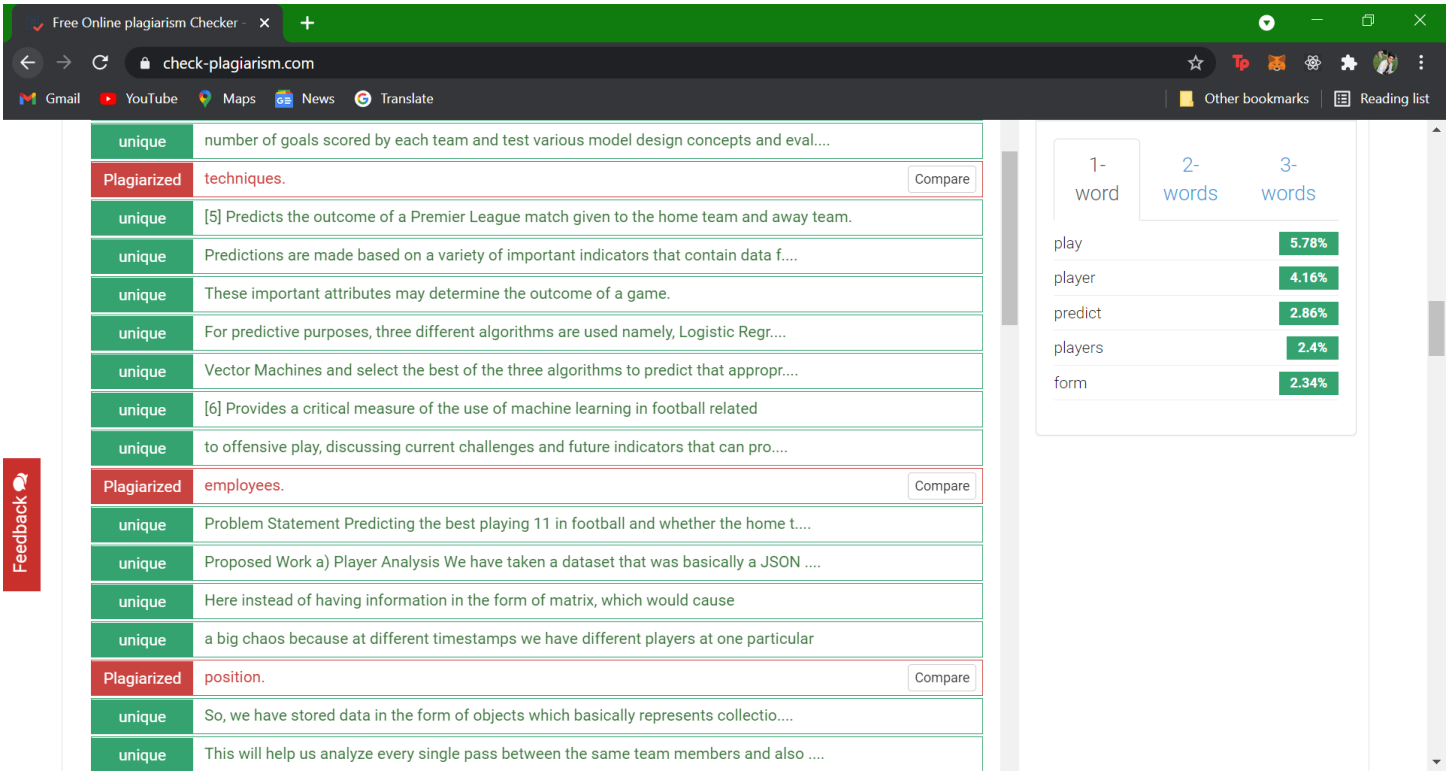
Feedback

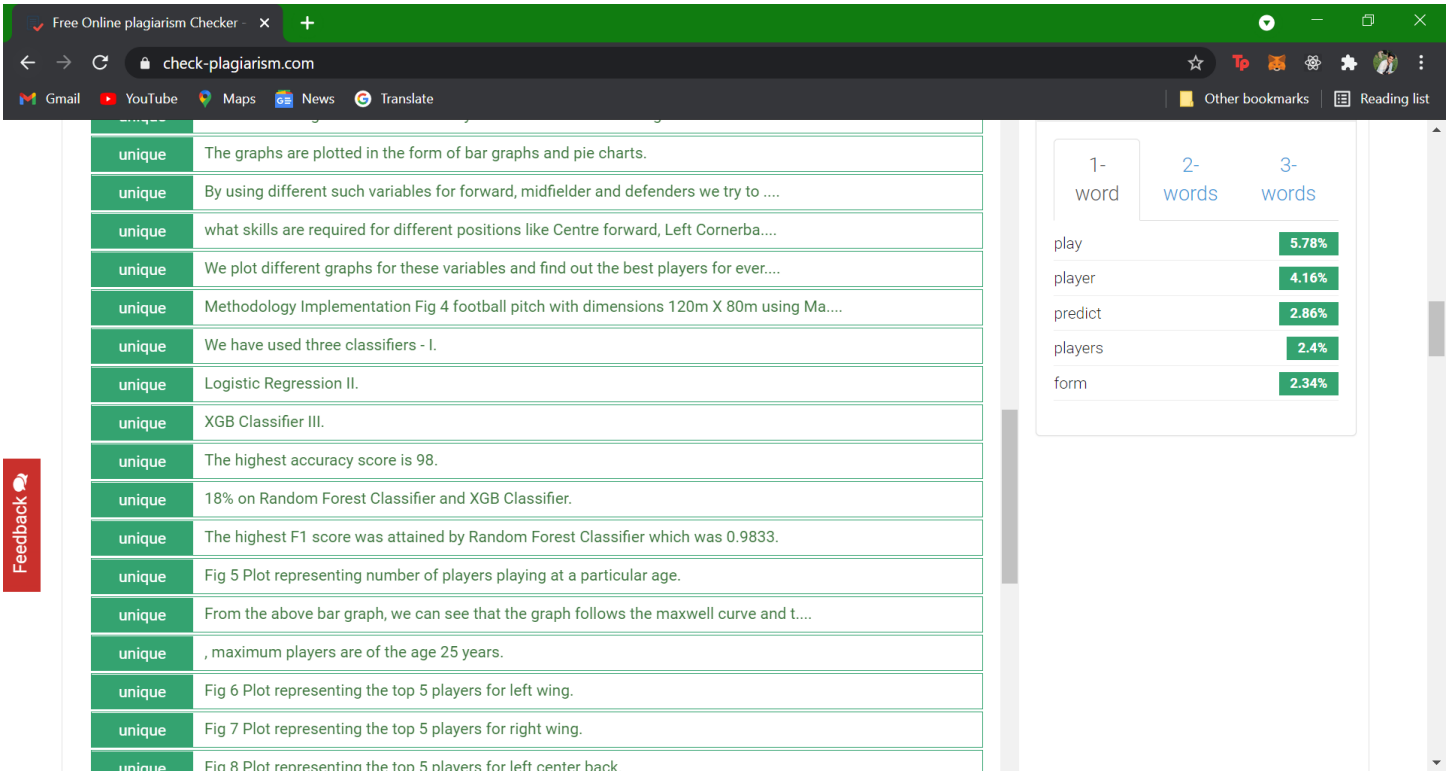
1-word	2-words	3-words
play	5.78%	
player	4.16%	
predict	2.86%	
players	2.4%	
form	2.34%	

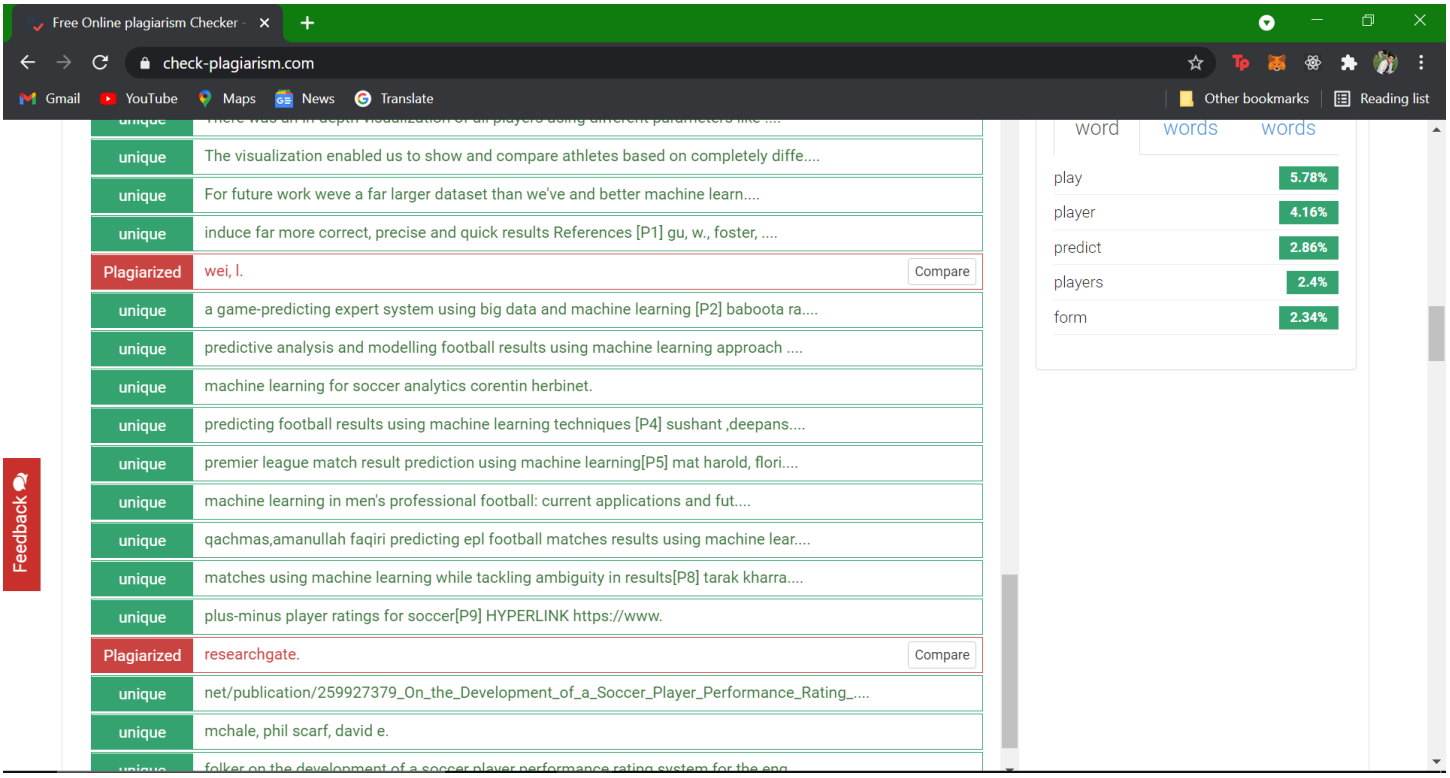
unique	we might first attempt to analyse individual players based on their sweeping, kicki....
unique	we would attempt to understand the player's strong zone on the playing area us....
unique	The passing ability of players would be analysed by using line graph and displaying....
unique	respectively Later we would take a dataset consisting of players who play for a tea....
unique	the basis of their position like Goalkeeping, Forward, Midfielder and Defender, by ....
unique	For example, new variables like Goal Stopping, Long Pass and Sweeping would be mad....
unique	calculating their mean average and then using these variables the players would be ....
unique	The result would be in the form of visual plots like pie chart and bar graphs.
unique	As the end result, we would try to mention down all the players who should be included
unique	in the top 11 and ultimately, we would be showing all the players placed on the foo....
unique	Related Work This section reviews works previously done in predicting a better side....
unique	[1] Uses player and team data collected from various web sources, and propose
unique	a professional program to better predict NHL match results and improve employment a....
unique	It also included key analysis, non-parametric mathematical analysis, support vecto....
unique	and a computational learning algorithm to predict whether the team will win the
unique	Integration methods were improved in the SVM reference classification, and the accu....
unique	Comparisons of multi-machine learning modes determine the chances of improving the

Feedback

1-word	2-words	3-words
play	5.78%	
player	4.16%	
predict	2.86%	
players	2.4%	
form	2.34%	


Free Online plagiarism Checker
check-plagiarism.com
Gmail
YouTube
Maps
News
Translate
Other bookmarks
Reading list
unique number of goals scored by each team and test various model design concepts and eval....
Plagiarized techniques. Compare
unique [5] Predicts the outcome of a Premier League match given to the home team and away team.
unique Predictions are made based on a variety of important indicators that contain data f....
unique These important attributes may determine the outcome of a game.
unique For predictive purposes, three different algorithms are used namely, Logistic Regr....
unique Vector Machines and select the best of the three algorithms to predict that appropr....
unique [6] Provides a critical measure of the use of machine learning in football related
unique to offensive play, discussing current challenges and future indicators that can pro....
Plagiarized employees. Compare
unique Problem Statement Predicting the best playing 11 in football and whether the home t....
unique Proposed Work a) Player Analysis We have taken a dataset that was basically a JSON ....
unique Here instead of having information in the form of matrix, which would cause
unique a big chaos because at different timestamps we have different players at one particular
Plagiarized position. Compare
unique So, we have stored data in the form of objects which basically represents collectio....
unique This will help us analyze every single pass between the same team members and also ....
Feedback
1-word
2-words
3-words
play 5.78%
player 4.16%
predict 2.86%
players 2.4%
form 2.34%


Free Online plagiarism Checker
check-plagiarism.com
Gmail
YouTube
Maps
News
Translate
Other bookmarks
Reading list
unique The graphs are plotted in the form of bar graphs and pie charts.
unique By using different such variables for forward, midfielder and defenders we try to ....
unique what skills are required for different positions like Centre forward, Left Cornerba....
unique We plot different graphs for these variables and find out the best players for ever....
unique Methodology Implementation Fig 4 football pitch with dimensions 120m X 80m using Ma....
unique We have used three classifiers - I.
unique Logistic Regression II.
unique XGB Classifier III.
unique The highest accuracy score is 98.
unique 18% on Random Forest Classifier and XGB Classifier.
unique The highest F1 score was attained by Random Forest Classifier which was 0.9833.
unique Fig 5 Plot representing number of players playing at a particular age.
unique From the above bar graph, we can see that the graph follows the maxwell curve and t....
unique , maximum players are of the age 25 years.
unique Fig 6 Plot representing the top 5 players for left wing.
unique Fig 7 Plot representing the top 5 players for right wing.
unique Fig 8 Plot representing the top 5 players for left center back
Feedback
1-word
2-words
3-words
play 5.78%
player 4.16%
predict 2.86%
players 2.4%
form 2.34%


Free Online plagiarism Checker
check-plagiarism.com
Gmail
YouTube
Maps
News
Translate
Other bookmarks
Reading list
unique
The visualization enabled us to show and compare athletes based on completely diffe....
unique
For future work weve a far larger dataset than we've and better machine learn....
unique
induce far more correct, precise and quick results References [P1] gu, w., foster, ...
Plagiarized
wei, l.
Compare
unique
a game-predicting expert system using big data and machine learning [P2] baboota ra....
unique
predictive analysis and modelling football results using machine learning approach ....
unique
machine learning for soccer analytics corentin herbinet.
unique
predicting football results using machine learning techniques [P4] sushant ,deepans....
unique
premier league match result prediction using machine learning[P5] mat harold, flori....
unique
machine learning in men's professional football: current applications and fut....
unique
qachmas,amanullah faqiri predicting epl football matches results using machine lear....
unique
matches using machine learning while tackling ambiguity in results[P8] tarak kharra....
unique
plus-minus player ratings for soccer[P9] HYPERLINK https://www.
Plagiarized
researchgate.
Compare
unique
net/publication/259927379_On_the_Development_of_a_Soccer_Player_Performance_Rating_....
unique
mchale, phil scarf, david e.
word
words
words
play
5.78%
player
4.16%
predict
2.86%
players
2.4%
form
2.34%
Feedback