Major Project Report

On

Qatar 2022 FIFA World Cup Predictor

Submitted by:
Vikram Badhan (vb2174)
Deepti Goutham (dg3781)
Akshat Sharma (as14680)

Table of Contents

1.INTRODUCTION 1

1.1 Problem statement 2

2. High-Level Design Diagram 2

3. Strategy and Goal Approach 4

3.1 Schema details, Data types, and table description 5

3.1.1 Team Schema 5

3.1.2 Player Schema 6

4. Hypothesis and Report 7

4.1 Player Analysis 7

4.2 Time Series Analysis For Players 9

4.3 FIFA - Team Analysis 8

4.4 ML Algorithms - Players 13

4.4.1 Random Forest 13

4.4.2 Gradient- Boosted Tree Classifier regression 13

4.5 ML Algorithms - Teams 15

4.5.1 Random Forest 15

4.5.2 Naïve Bayes 15

5 Conclusion 17

6 Future Work 18

1.Introduction

The FIFA World Cup is a quadrennial international soccer tournament contested by the senior men's national teams of the members of Fédération Internationale de Football Association (FIFA), the sport's global governing body. The tournament has been held since 1930, except for 1942 and 1946 due to World War II.

The problem facing the organizers of the FIFA World Cup is selecting the teams that will participate in the tournament and to schedule and staging the event in a way that maximizes the enjoyment of the players and fans and minimizes any negative impact on the host country. This includes selecting a host nation, building, or renovating stadiums and other infrastructure, and coordinating the various logistical and operational aspects of the tournament. It also involves promoting the event and generating revenue through sponsorships, ticket sales, and media rights.

1.1 Problem Statement

We are trying to predict the logistical impact of the FIFA World Cup and the change in trends across the dynamics of the teams and players involved in this prestigious tournament, creating a Machine Learning model which predicts the ***best players of the tournament*** and predicts ***which teams will win in the different stages of the tournament***.

2. High-Level Design Diagram

Apache Spark is an open-source distributed computing system that is designed for large-scale data processing and analytics. It is built on top of the Hadoop distributed file system and is designed to be faster and more flexible than Hadoop MapReduce, the original framework for distributed data processing on Hadoop. Spark works by dividing a large dataset into smaller chunks called "partitions" and distributing them across a cluster of

computers for parallel processing. Each partition is processed by a separate "executor" on a worker node in the cluster. The executors are responsible for executing the tasks assigned to them by the "driver" program, which is the main program that controls the Spark application. Spark has a number of features that make it well-suited for large-scale data processing and analytics, including in-memory processing, support for a wide range of data sources and formats, and a rich set of libraries and APIs for various data processing and machine learning tasks.

Google Colab is particularly useful for data science and machine learning tasks, as it comes pre-installed with a number of useful libraries and tools, such as NumPy, pandas, scikit-learn, and TensorFlow. It also provides access to GPUs and TPUs (Tensor Processing Units), which can be used to accelerate the training of machine learning models.

Spark ML is widely used in a variety of applications, including recommendation systems, fraud detection, and predictive maintenance. It is an important tool for data scientists and developers working with large datasets and looking to build efficient machine-learning models.

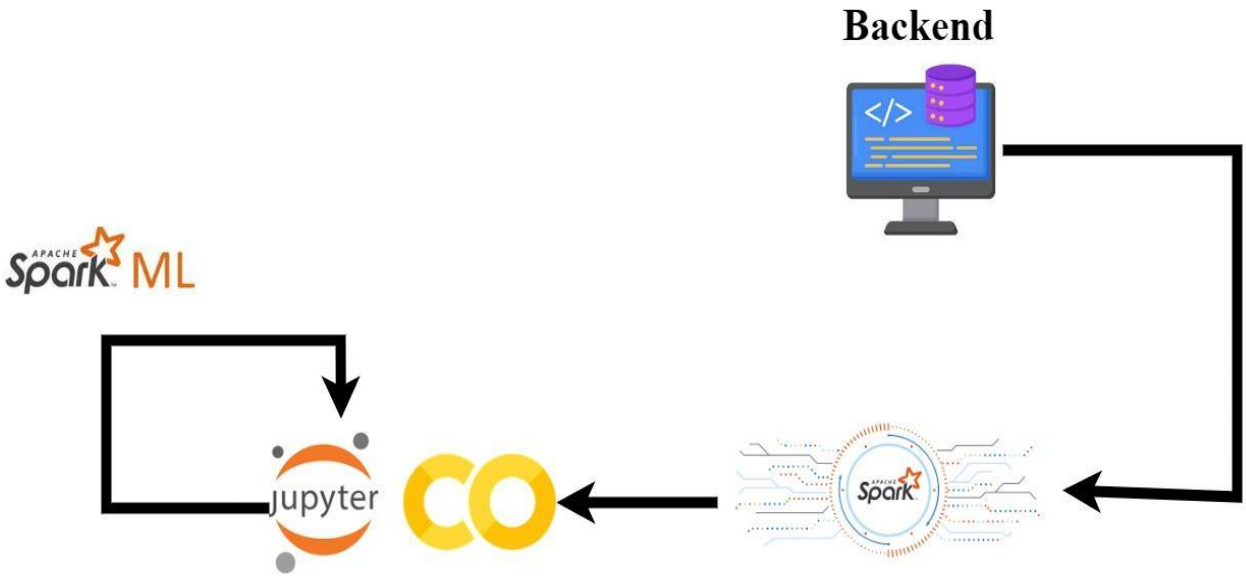


Fig 2.1 : High Level Diagram

3.Strategy and Goal Approach

1) Gather Data**:**

The first step would be to gather as much data as possible about past World Cup matches, as well as about the teams that will be participating in the Qatar World Cup. This data could include information about the teams' strengths and weaknesses, their past performance in World Cup matches, and other relevant factors.

2) Pre-Process the Data**:**

Once you have collected the data, you will need to pre-process it to make it suitable for analysis. This may involve cleaning the data, filling in missing values, and transforming the data into a format that is easier to work with.

3) Explore the Data**:**

Next, you will need to explore the data to understand patterns and trends that could be useful for predicting the winner of the Qatar World Cup. You could use visualizations and statistical analysis techniques to identify important features or variables that could be used to build a predictive model.

4) Build a Predictive Model**:**

Based on the insights you have gained from exploring the data, you can then build a predictive model that uses machine learning algorithms to forecast the outcome of World Cup matches. There are many different algorithms you could use for this task, such as decision trees, random forests, or neural networks.

5) Make Predictions**:**
Once you have a validated model, you can use it to make predictions about the winner of the Qatar World Cup. You may want to update the model as new data becomes available, or as the tournament progresses, to improve its accuracy.

3.1 Schema details, Data types, and table description

3.1.1.Team Schema-

```
root
 |-- date: timestamp (nullable = true)
 |-- home_team: string (nullable = true)
 |-- away_team: string (nullable = true)
 |-- home_team_continent: string (nullable = true)
 |-- away_team_continent: string (nullable = true)
 |-- home_team_fifa_rank: integer (nullable = true)
 |-- away_team_fifa_rank: integer (nullable = true)
 |-- home_team_total_fifa_points: integer (nullable = true)
 |-- away_team_total_fifa_points: integer (nullable = true)
 |-- home_team_score: integer (nullable = true)
 |-- away_team_score: integer (nullable = true)
 |-- tournament: string (nullable = true)
 |-- city: string (nullable = true)
 |-- country: string (nullable = true)
 |-- neutral_location: boolean (nullable = true)
 |-- shoot_out: string (nullable = true)
 |-- home_team_result: string (nullable = true)
 |-- home_team_goalkeeper_score: double (nullable = true)
 |-- away_team_goalkeeper_score: double (nullable = true)
 |-- home_team_mean_defense_score: double (nullable = true)
 |-- home_team_mean_offense_score: double (nullable = true)
 |-- home_team_mean_midfield_score: double (nullable = true)
 |-- away_team_mean_defense_score: double (nullable = true)
 |-- away_team_mean_offense_score: double (nullable = true)
 |-- away_team_mean_midfield_score: double (nullable = true)
```

We use this data to predict the top 10 teams based on the overall performance of each team There are over 25 columns in our data we use all of them as key features in our analysis. We compare two teams to classify which team wins, losses, or ends in a draw.

Our end result is stored in “Home Team Result”, Where the outcome is classified in “0,1,2” as “loss, tie, win”.

3.1.2.Player Schema-

```
root
 |-- label: integer (nullable = true)
 |-- value_eur: double (nullable = true)
 |-- age: integer (nullable = true)
 |-- wage_eur: double (nullable = true)
 |-- league_level: double (nullable = true)
 |-- weak_foot: integer (nullable = true)
 |-- skill_moves: integer (nullable = true)
 |-- international_reputation: integer (nullable = true)
 |-- pace: double (nullable = true)
 |-- shooting: double (nullable = true)
 |-- passing: double (nullable = true)
 |-- dribbling: double (nullable = true)
 |-- defending: double (nullable = true)
 |-- physic: double (nullable = true)
 |-- attacking_crossing: integer (nullable = true)
 |-- attacking_finishing: integer (nullable = true)
 |-- attacking_heading_accuracy: integer (nullable = true)
 |-- attacking_short_passing: integer (nullable = true)
 |-- attacking_volleys: integer (nullable = true)
 |-- skill_dribbling: integer (nullable = true)
 |-- skill_curve: integer (nullable = true)
 |-- skill_fk_accuracy: integer (nullable = true)
 |-- skill_long_passing: integer (nullable = true)
 |-- skill_ball_control: integer (nullable = true)
 |-- movement_acceleration: integer (nullable = true)
 |-- movement_sprint_speed: integer (nullable = true)
 |-- movement_agility: integer (nullable = true)
 |-- movement_reactions: integer (nullable = true)
 |-- movement_balance: integer (nullable = true)
 |-- power_shot_power: integer (nullable = true)
 |-- power_jumping: integer (nullable = true)
 |-- power_stamina: integer (nullable = true)
 |-- power_strength: integer (nullable = true)
 |-- power_long_shots: integer (nullable = true)
 |-- mentality_aggression: integer (nullable = true)
 |-- mentality_interceptions: integer (nullable = true)
 |-- mentality_positioning: integer (nullable = true)
 |-- mentality_vision: integer (nullable = true)
 |-- mentality_penalties: integer (nullable = true)
 |-- mentality_composure: double (nullable = true)
 |-- defending_marking_awareness: integer (nullable = true)
 |-- defending_standing_tackle: integer (nullable = true)
 |-- defending_sliding_tackle: integer (nullable = true)
```

Based on various features like shooting, passing, attacking skills, and the player's physic an overall score out of 100 is calculated. This is done by taking the average of these skills divided by the number of skills. The result is stored in "overall" which we had calculated during data preparation.

This is later predicted by our two machine learning models which are explained in this report. Our aim here is two come up with a list of the top 10 players based on their overall score over the years 2017 – 2022.

4.Hypothesis & Report

4.1. Player Analysis:

In order to do the player analysis, we utilized a player dataset for 2022 that we prepared. It consists of 19500 players in total.

1) Top 10 Players of 2022 :

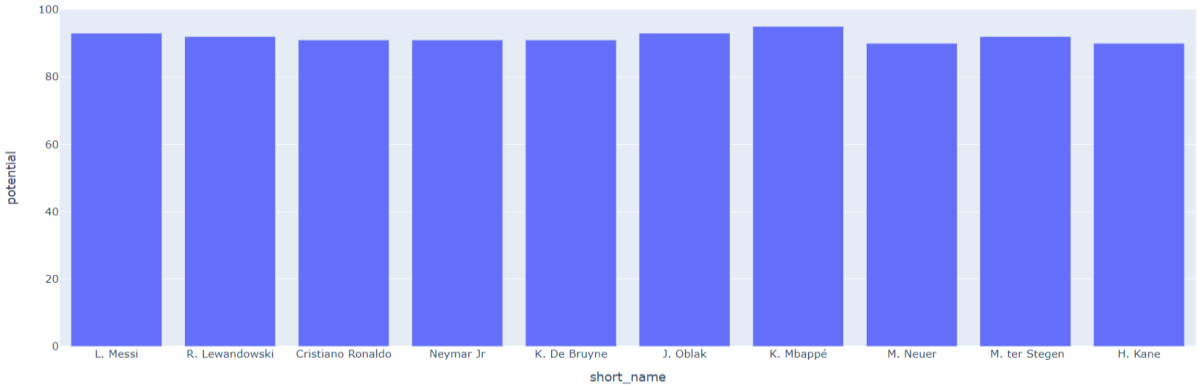


Fig 4.1 : Top 10 players of 2022

Here, we have filtered out the top 10 players of the world in the year 2022 based on the features called “potential” and “overall”.

2) Top 10 Goalkeepers and Strikers of 2022 :

short_name	nationality_name
J. Oblak	Slovenia
M. Neuer	Germany
M. ter Stegen	Germany
T. Courtois	Belgium
Ederson	Brazil
Alisson	Brazil
G. Donnarumma	Italy
K. Navas	Costa Rica
H. Lloris	France
W. Szczęsny	Poland

Fig 4.2 : Top 10 Goalkeepers

short_name	nationality_name
R. Lewandowski	Poland
H. Kane	England
L. Suárez	Uruguay
R. Lukaku	Belgium
E. Haaland	Norway
S. Agüero	Argentina
C. Immobile	Italy
J. Vardy	England
E. Cavani	Uruguay
L. Martínez	Argentina

Fig 4.3 : Top 10 Strikers

Similarly, here we have a dataframe which consists of the top 10 goalkeepers and strikers of 2022. We filtered it out using the features called “player_position” and “overall”.

3) Top 10 Midfielders and Defenders of 2022 :

short_name	nationality_name
T. Kroos	Germany
L. Modrić	Croatia
Parejo	Spain
N. Barella	Italy
Merino	Spain
Arthur	Brazil
F. Valverde	Uruguay
I. Rakitić	Croatia
M. Pjanić	Bosnia and Herzeg...
Rúben Neves	Portugal

Fig 4.4 : Top 10 Midfielders

short_name	nationality_name
V. van Dijk	Netherlands
Sergio Ramos	Spain
Rúben Dias	Portugal
G. Chiellini	Italy
M. Hummels	Germany
K. Koulibaly	Senegal
R. Varane	France
A. Laporte	Spain
M. Škriniar	Slovakia
Thiago Silva	Brazil

Fig 4.5 : Top 10 Defenders

The data frames above show the top 10 midfielders and defenders in the year 2022. We filtered it out using the features called “player_position” and “overall”.

4) Top 10 Valued and paid players of 2022 :

short_name	nationality_name
K. Mbappé	France
E. Haaland	Norway
H. Kane	England
Neymar Jr	Brazil
K. De Bruyne	Belgium
R. Lewandowski	Poland
F. de Jong	Netherlands
G. Donnarumma	Italy
J. Sancho	England
T. Alexander-Arnold	England

Fig 4.7 : Top 10 Value players

short_name	nationality_name
K. Benzema	France
K. De Bruyne	Belgium
L. Messi	Argentina
Casemiro	Brazil
T. Kroos	Germany
R. Sterling	England
Neymar Jr	Brazil
R. Lewandowski	Poland
Cristiano Ronaldo	Portugal
S. Mané	Senegal

Fig 4.8 : Top 10 Paid Players

Here, we see the top 10 value and paid players of the year 2022 based on the feature called “Value” and “wage” respectively.

4.2 Time Series Analysis of Players

In this Project, we were able to do a Time series Analysis on all the players. For this purpose, we used a dataset of players from 2015 to 2022 and performed a time series analysis to check the trends.

1) Time Series analysis for finding the change in the value of Players across the years

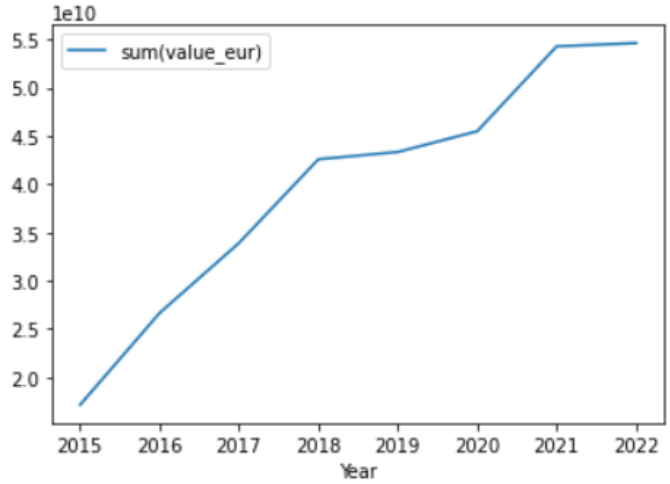


Fig 4.9 : change in value of players

This shows the worth of the sport over the span of 2015 to 2022. And, it is gradually on the rise over the years.

2) Time series analysis for finding changes in player wages across the years:

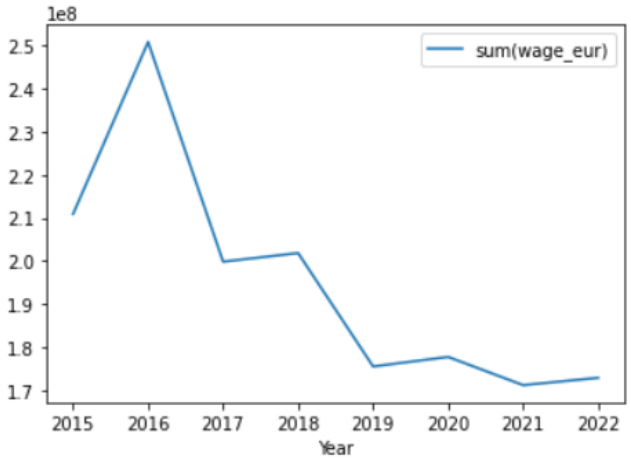


Fig 4.10 : change in player wage

In the chart above, we see a sharp increase in the player wages between 2015 to 2017, however, the wages decrease between the years 2017 to 2022.

3) Time Series analysis for finding the change in the average value of the players :

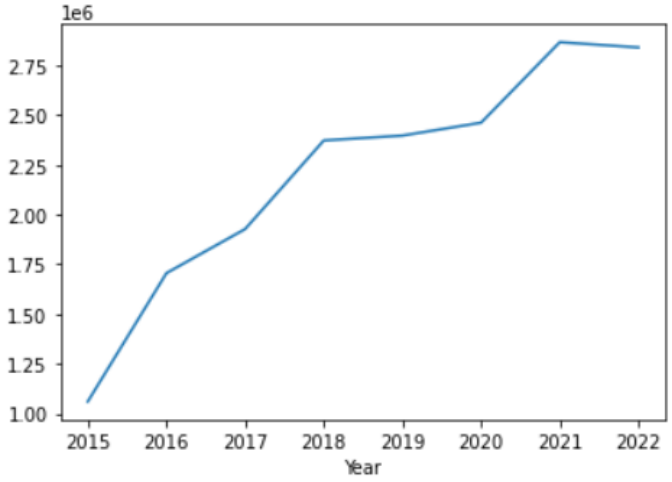


Fig 4.11: change in the average value of players

This graph is a sanity check for the average value of players across the years. It is the same as the value graph that we saw above. It basically shows that the average value of players is uniform around the curve.

4) Time Series Analysis for finding the change in the average wage of players :

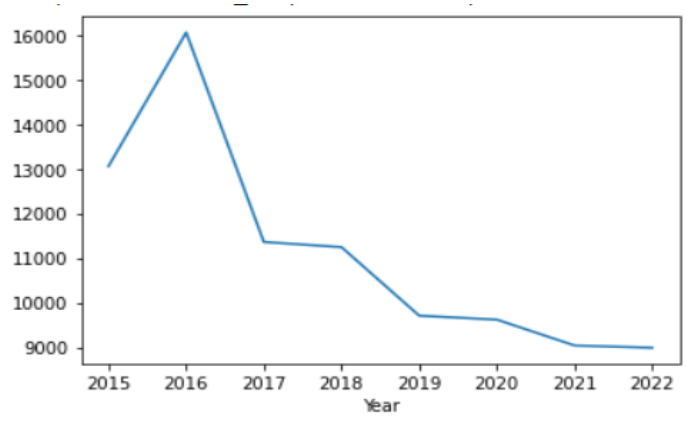


Fig 4.12: change in the average wage of players

This graph is a sanity check for average wages of the players over the years. We see a sharp increase in the player wages between 2015 to 2017, however, the wages decrease between the years 2017 to 2022.

5) Time Series Analysis for finding the change in average overall stats of a player :

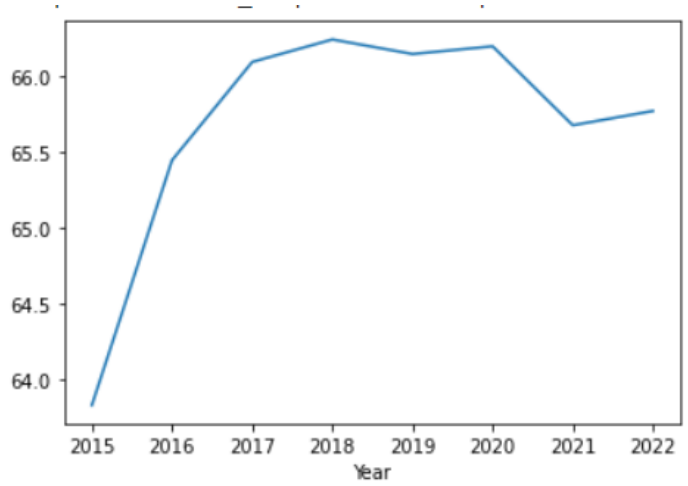


Fig 4.13 : change in overall stats of a player

In this chart, we see a steady increase in the average overall of the players which means that players are pushing towards a better game and they are getting better every year.

6)Time Series Analysis for finding the change in the Average potential of players :

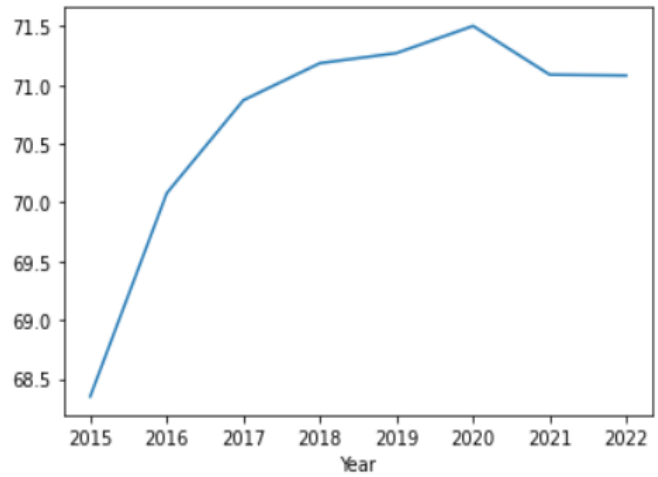


Fig 4.14 : change in average potential of players

In this chart, we see a steady increase in the average potential of the players which means that players are pushing towards a better game and they are getting fitter and faster.

7) Time Series Analysis for finding the change in average age of the players :

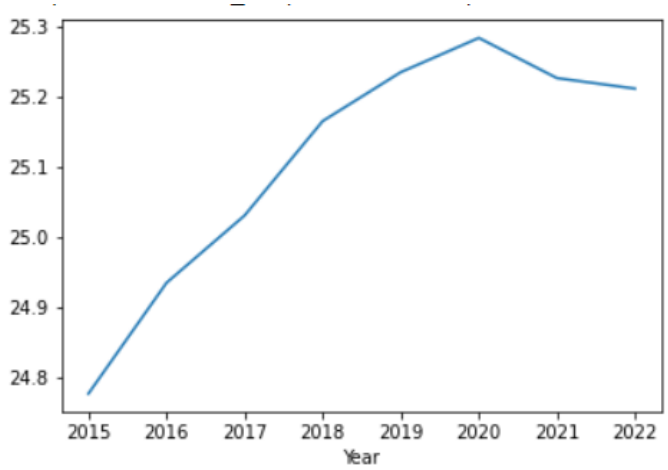


Fig 4.15 : change in average age of player

In this chart, we see that the average age is increasing until 2020. Senior players during covid in the year 2020-2021 were taking more money, and so younger players who were cheaper to purchase and maintain were brought into the game.

8) Time Series Analysis for finding the change in average height of the players :

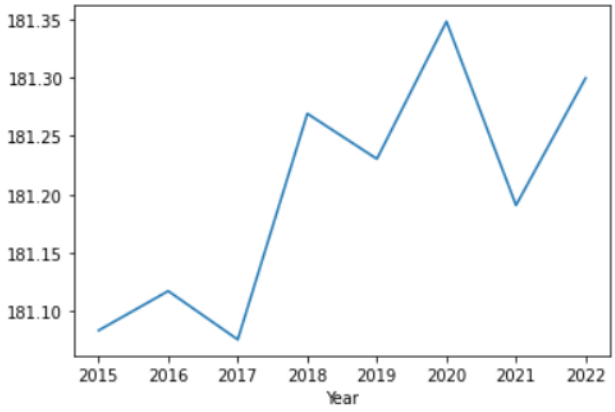


Fig 4.16 : change in average height of players

In this chart, there is no correlation between an increase or decrease in the average height of the players as it is varying across the entire plot. Therefore, there is no trend of increasing or decreasing height in football.

9) Time Series Analysis for finding the change in Average weight of players :

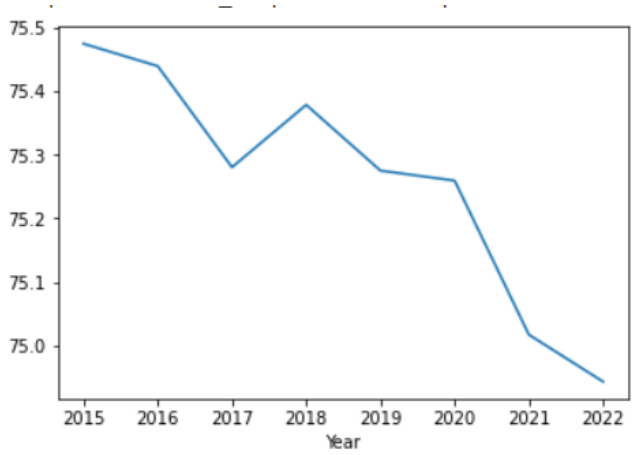
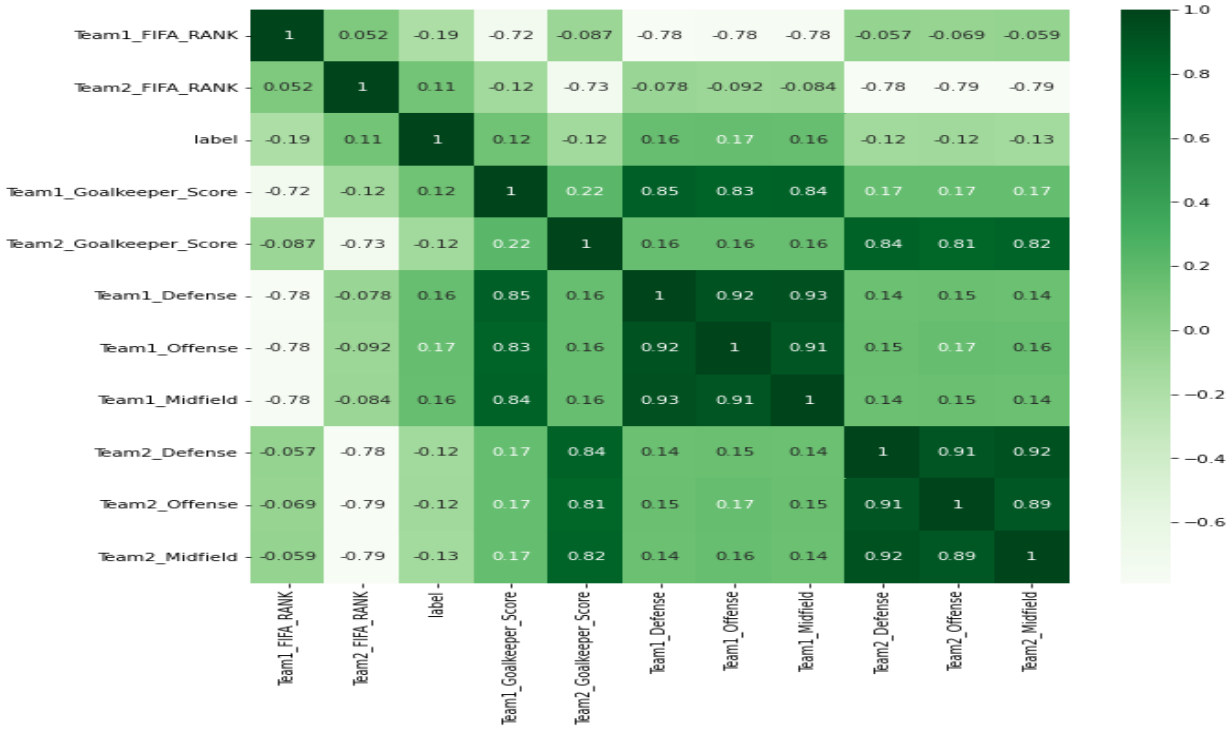


Fig 4.17 : change in average weight of players

In this graph, we see that the average weight of players is on a constant decline which means that the sport is turning towards faster players.

4.3 FIFA - Team Analysis



Above is the correlation matrix for our data set between two teams “Team 1” and “Team 2”. All the features are closely correlated. We use this to list the top 10 teams that have strong attacking skills and teams that are good at defending.

1) Top 10 Attacking teams :

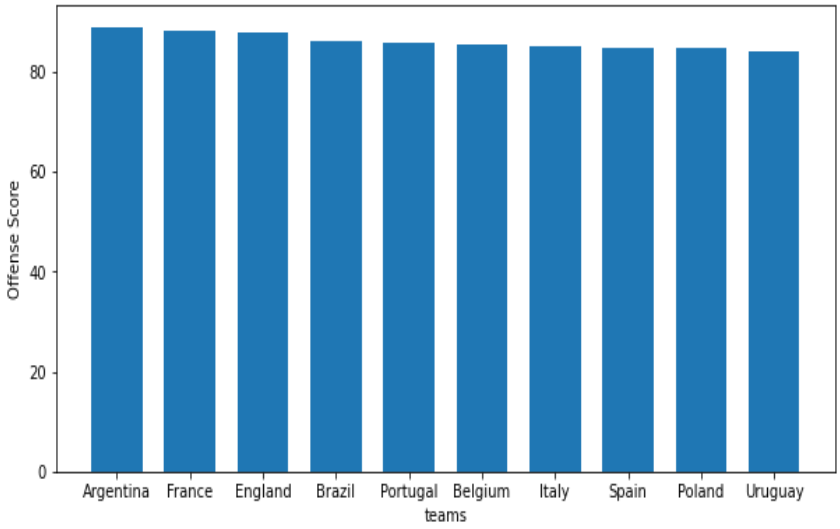


Fig 4.19 : Top 10 attacking teams

2) Top 10 defending teams

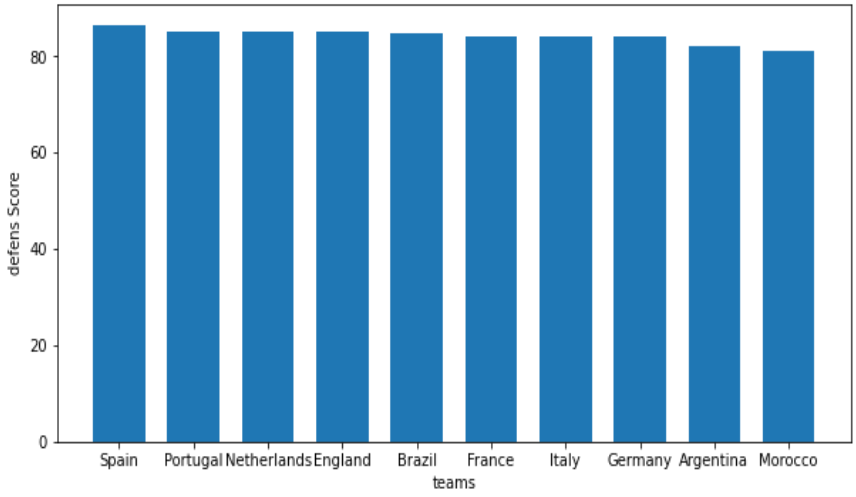


Fig 4.20 : Top 10 defending Teams

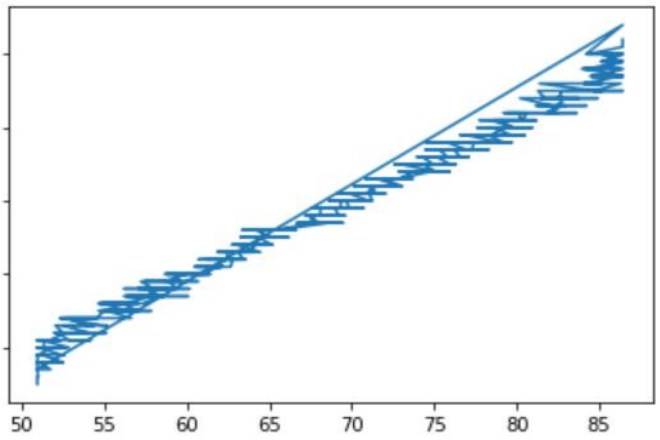
4.4 ML Algorithms – Players

4.4.1 Random forest

Mean absolute error = 0.50, r2 score = 0.99

The straight line in between represents the original value. The slight variations towards the edges represent the predicted value. Here we performed a regression random forest algorithm. The accuracy achieved is 99%.

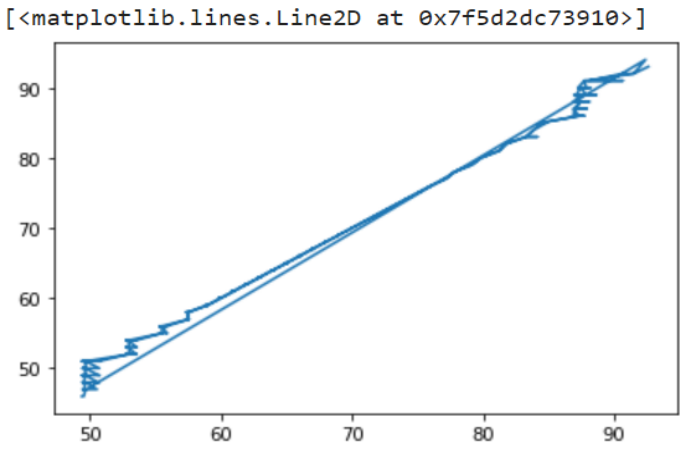
features	label	prediction
[45.0,70000.0,18....	45	50.94443694241347
[46.0,40000.0,18....	46	50.94443694241347
[46.0,50000.0,18....	46	50.94443694241347
[46.0,60000.0,17....	46	50.94443694241347
[46.0,60000.0,18....	46	50.94443694241347
[46.0,60000.0,18....	46	51.003781049129024
[47.0,60000.0,17....	47	50.94443694241347
[47.0,60000.0,17....	47	50.94443694241347
[47.0,60000.0,18....	47	50.94443694241347
[47.0,60000.0,18....	47	50.94443694241347
[47.0,60000.0,19....	47	50.94443694241347
[47.0,70000.0,18....	47	50.94443694241347
[48.0,35000.0,18....	48	50.94443694241347
[48.0,45000.0,17....	48	50.94443694241347
[48.0,45000.0,18....	48	50.94443694241347
[48.0,45000.0,20....	48	51.03863917045974
[48.0,50000.0,17....	48	50.94443694241347
[48.0,60000.0,17....	48	50.94443694241347
[48.0,60000.0,17....	48	50.94443694241347
[48.0,60000.0,18....	48	50.94443694241347



4.4.2 Gradient-boosted tree classifier regression

Mean absolute error = 0.96, Root Mean Squared Error(RMSE) = 0.2**.** The straight line in between represents the original value. The slight variations towards the edges represent the predicted value. Here we performed a regression gradient-boosted algorithm. Both models performed equally well.

prediction	label	features
49.411996935609324	46	[46.0,60000.0,17....
49.411996935609324	46	[46.0,60000.0,17....
49.411996935609324	46	[46.0,60000.0,17....
49.547134365493626	46	[46.0,70000.0,17....
49.681938117704526	47	[47.0,40000.0,19....
49.64589503412328	47	[47.0,50000.0,18....
49.63195458685835	47	[47.0,50000.0,19....
49.697026199686015	47	[47.0,50000.0,22....
49.47706854843699	47	[47.0,60000.0,17....
49.547134365493626	47	[47.0,60000.0,17....



4.5 ML Algorithms - Teams

4.5.1 Random Forest

```
+--------------+-----+--------------------+
|predictedLabel|label|            features|
+--------------+-----+--------------------+
|             0|    0|[130.0,58.0,0.0,5...|
|             0|    0|[75.0,16.0,0.0,72...|
|             0|    0|[65.0,2.0,0.0,82....|
|             0|    0|[66.0,7.0,0.0,52....|
|             0|    0|[92.0,11.0,0.0,51...|
|             2|    2|[95.0,55.0,2.0,72...|
|             2|    2|[89.0,59.0,2.0,72...|
|             1|    1|[47.0,43.0,1.0,70...|
|             2|    2|[75.0,58.0,2.0,70...|
|             1|    1|[69.0,80.0,1.0,70...|
|             2|    2|[75.0,79.0,2.0,70...|
|             0|    0|[205.0,3.0,0.0,50...|
|             0|    0|[156.0,3.0,0.0,50...|
|             0|    0|[186.0,15.0,0.0,5...|
|             0|    0|[165.0,2.0,0.0,50...|
|             1|    1|[9.0,61.0,1.0,81....|
|             1|    1|[8.0,36.0,1.0,81....|
|             1|    1|[10.0,62.0,1.0,81...|
|             1|    1|[3.0,16.0,1.0,79....|
|             1|    1|[5.0,21.0,1.0,79....|
+--------------+-----+--------------------+
only showing top 20 rows
```

```
Accuracy = 0.9536494405966969
Test Error = 0.0463506
```

we used PySpark ml libraries to implement our machine-learning algorithms. Here our aim was to train the Machine learning model that will help predict the top Teams in the World Cup matches.

4.5.2 Naive Bayes

```
+----------+-----+--------------------+
|prediction|label|            features|
+----------+-----+--------------------+
|       2.0|    0|[1.0,2.0,0.0,90.0...|
|       2.0|    0|[1.0,2.0,0.0,92.0...|
|       2.0|    0|[1.0,4.0,0.0,89.0...|
|       2.0|    1|[1.0,5.0,1.0,79.0...|
|       2.0|    1|[1.0,5.0,1.0,86.0...|
|       2.0|    2|[1.0,5.0,2.0,89.0...|
|       2.0|    0|[1.0,7.0,0.0,89.0...|
|       2.0|    0|[1.0,8.0,0.0,91.0...|
|       2.0|    1|[1.0,8.0,1.0,86.0...|
|       2.0|    1|[1.0,8.0,1.0,86.0...|
|       2.0|    1|[1.0,8.0,1.0,88.0...|
|       2.0|    1|[1.0,8.0,1.0,91.0...|
|       2.0|    1|[1.0,9.0,1.0,86.0...|
|       2.0|    1|[1.0,9.0,1.0,86.0...|
|       2.0|    2|[1.0,9.0,2.0,86.0...|
|       2.0|    1|[1.0,10.0,1.0,86....|
|       2.0|    1|[1.0,11.0,1.0,86....|
|       2.0|    1|[1.0,12.0,1.0,82....|
|       2.0|    1|[1.0,12.0,1.0,86....|
|       2.0|    1|[1.0,13.0,1.0,86....|
+----------+-----+--------------------+
only showing top 20 rows
```

```
print("Accuracy: " + str(evaluator.evaluate(predictionAndLabels)))
```

```
Accuracy: 0.5967828418230563
```

we are predicting outcomes between two teams, team 1 and team 2 the team result has 3 classifications implying a draw, win, or a loss. In terms of accuracy, Random Forest is identified as the best model with 95% accuracy.

5.Conclusions:

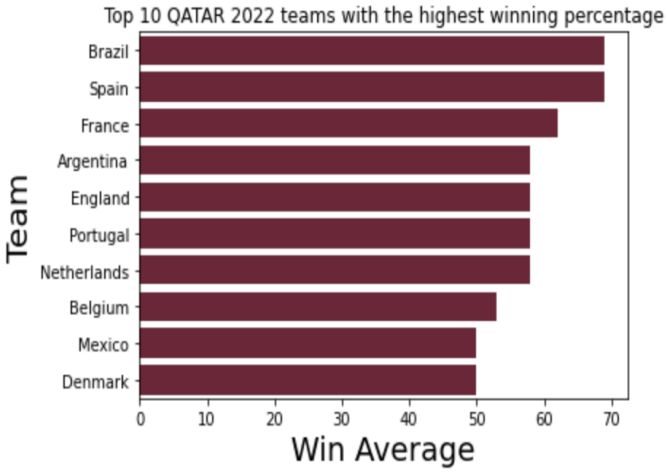
Based on various features like shooting, passing, and attacking skills an overall score out of 100 is calculated. Which is stored in the "Overall" column based on which the players are ranked.

Top 10 Players based on overall performance as per our analysis

short_name	nationality_name	age	year	potential
L. Messi	Argentina	34	2022	93
R. Lewandowski	Poland	32	2022	92
Cristiano Ronaldo	Portugal	36	2022	91
Neymar Jr	Brazil	29	2022	91
K. De Bruyne	Belgium	30	2022	91
J. Oblak	Slovenia	28	2022	93
K. Mbappé	France	22	2022	95
M. Neuer	Germany	35	2022	90
M. ter Stegen	Germany	29	2022	92
H. Kane	England	27	2022	90

only showing top 10 rows

Top 10 Teams based on overall performance as per our analysis



6. Future work

1) Incorporate Real-time Data:

One potential direction for future work could be to incorporate real-time data into the predictive model. This could include data about the current form and performance of the teams, as well as information about injuries or other factors that could impact the outcome of the matches.

2) Expand the Dataset :

Another option could be to expand the dataset to include more data from past World Cup matches, as well as data from other football tournaments and leagues. This could help improve the accuracy of the model by providing a larger and more diverse sample of data to work with.

3) Explore new Algorithms :

Another potential direction for future work could be to explore new machine-learning algorithms that may be more effective at predicting the winner of the Qatar World Cup. There are many different algorithms available, and some may be better suited to this task than others.

4) Enhance the visualization of the result :

Finally, you could consider enhancing the visualization of the results to make it easier for users to understand and interpret the predictions. This could involve creating more interactive and intuitive visualizations, such as graphs, charts, or maps.

5) New Technologies :

Technologies like Tableau, Dask, and Pyspark can be used to further explore the scope of the project.