Lambton College Mississauga
BDM 2053 - Big Data Algorithms and Statistic
Professor. Dhruwal Shah

Final Project Report

Henry Jones Inbaraj - C0863081
Meenakshi - C0864515
Javier A. Melo – C0871987
Amal Vandananikkal - C0863255
Kelvin Simon - C0866577

Table of Contents

1.Introduction:

The Big Data Analysis for Football Team Formation project aimed to use big data algorithms to analyse football player performance data and identify the best players for each position. The project team used this analysis to form the best possible team based on player statistics and performance metrics.

2.Problem statement:

The data is having the player's performance ratings in different parameters. With the given data, to assemble the best squad.

Plan the game strategy by predicting the players position based on their metrics and to rate the new players performance.

3.Risk Encountered:

- Data loss were prevented by backing up.
- Lack of client engagement was treated by having iterative approach. Every stage was shared with the stake holders regularly.
- To avoid wrong selection of Model , ML expert advice was considered and model evaluation were performed.

4.Challenges Encountered:

- Conflicts were managed properly and recorded as team minutes.
- Co-ordinating the team members were bit difficult but it was addressed seriously and were co-ordinated everyone for daily stand up.
- Domain Knowledge was shared to everyone in the team with the help of football expert.

5.Problem Approach:

Data is analysed and followed a standard approach like finding null values, description and dropping the unwanted columns.

Data is treated with the advice from the football expert. Data is transformed in to data frames and selected the features for each position.

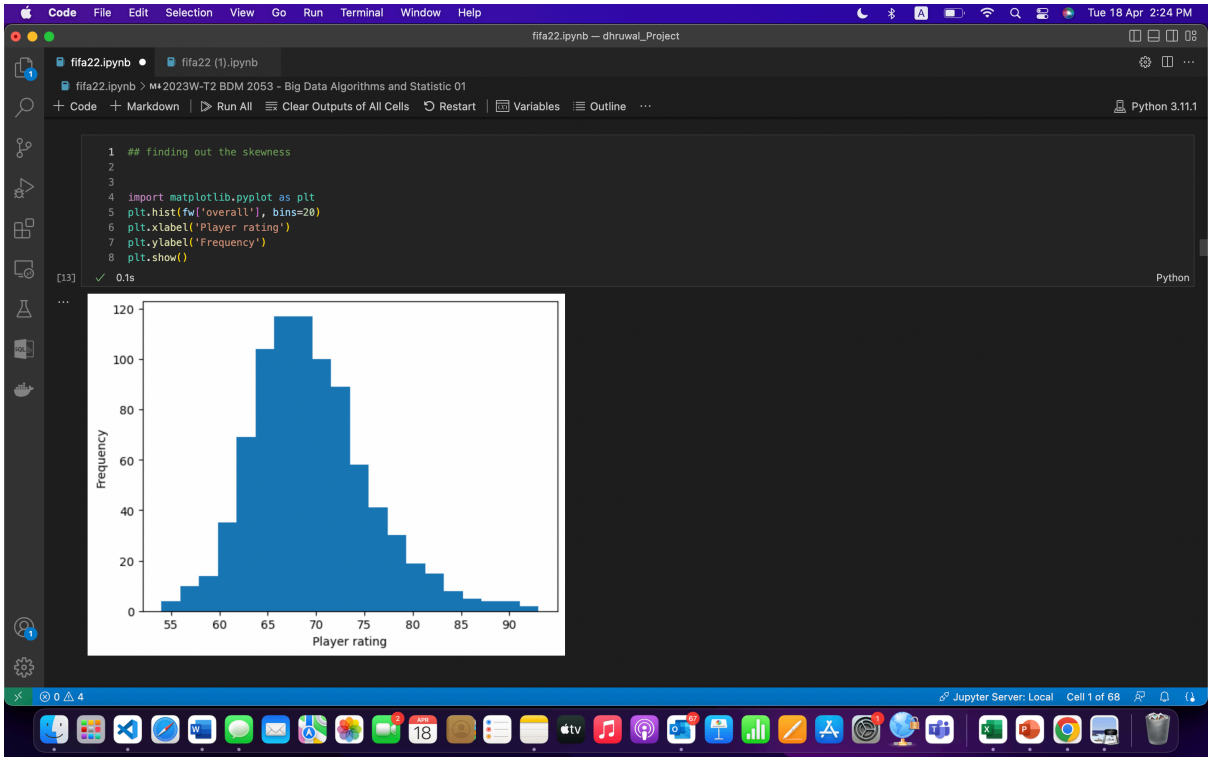
Best features for each position were figured out using correlation method.

Data frames were created for each position to find the best players to select in the team. It is based on the mean value of the all features of a player.

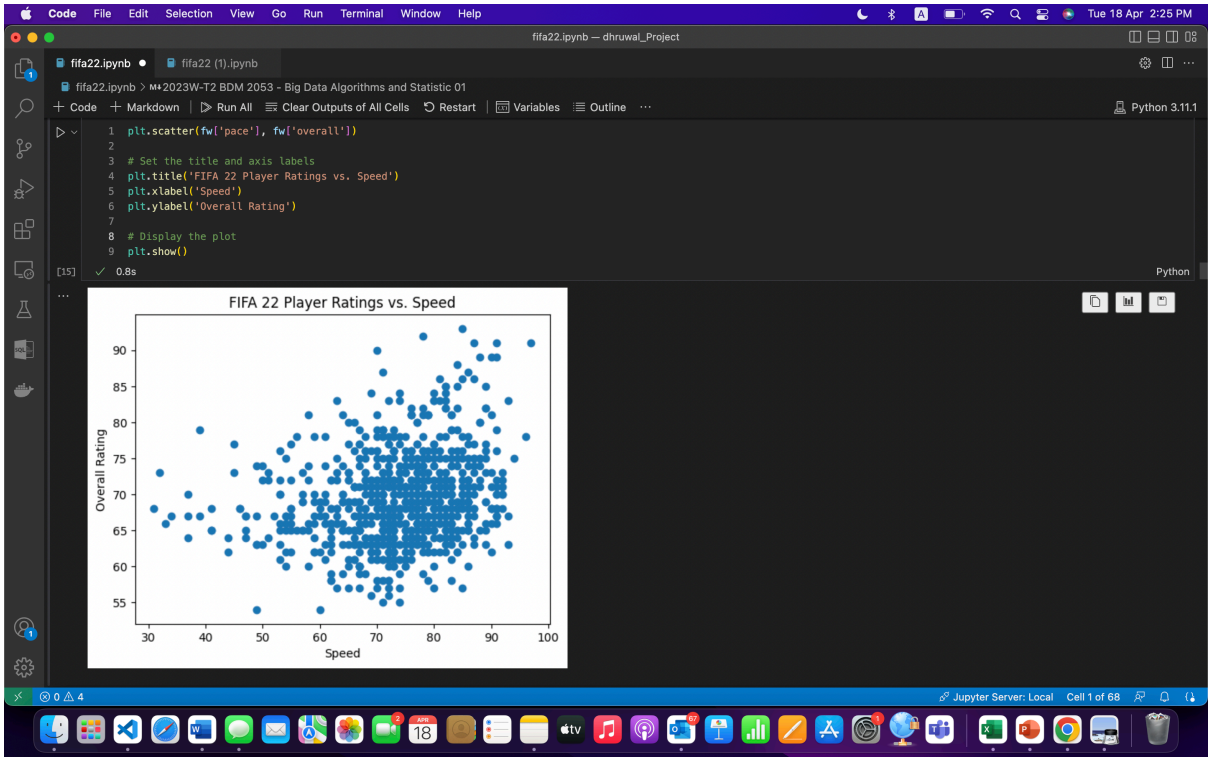
Knn model is used to predict the players position to plan the game strategy

5.1 Exploring the data:

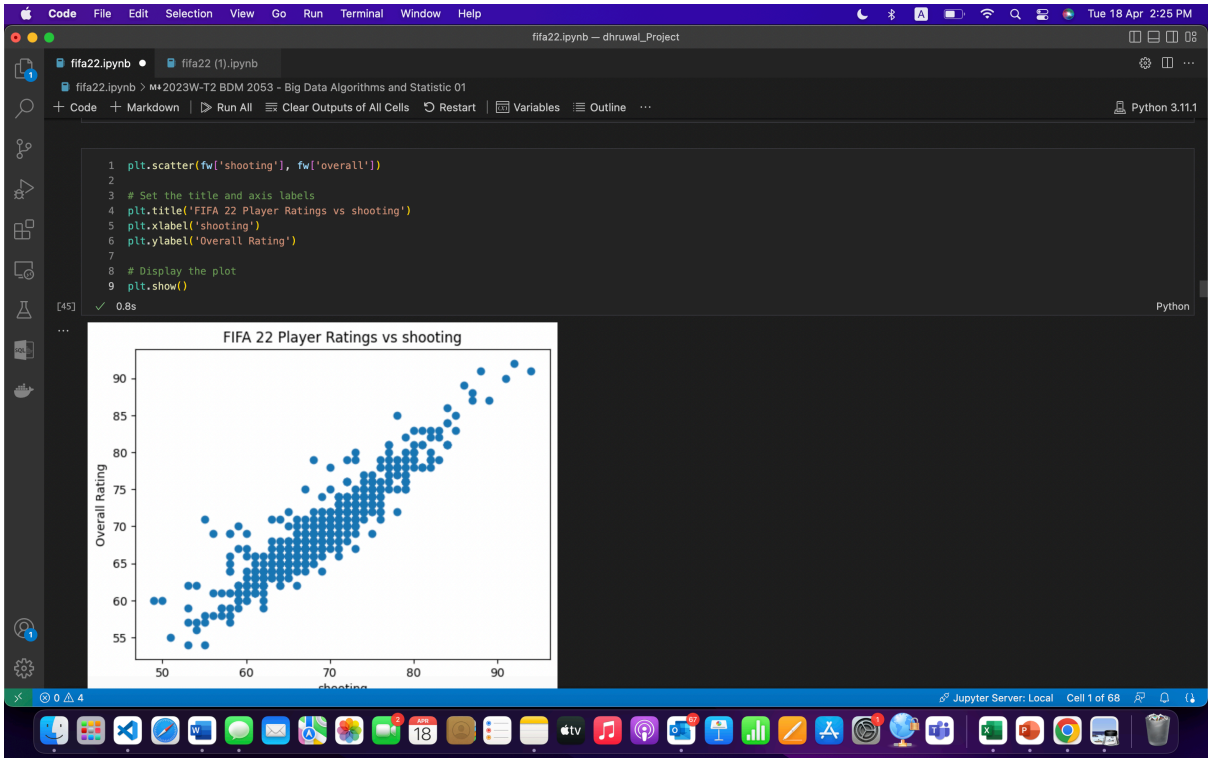
Importing the necessary libraries like pandas, matplotlib. pyplot, sklearn. neighbors, klearn.model_selection, numpy and sklearn. preprocessing. Finding the null values in the data to treat it. Dropping unwanted columns like value_eur, wage_eur. Filling the null values with zeros. Creating the dataframe for Forward positions with the 'club_position column' to find the best features affecting the 'overall_rating'. Finding the statics like mean, mode, max, min and standard deviation for the overall rating. Finding the skewness to know the data distribution.



Finding the correlation and plotting using scatter plots to give a visual treat. Correlation between ‘speed’ and ‘overall_rating’ is not good.



Correlation between ‘shooting’ and ‘overall_rating’ is good (positive correlation)



Factors affecting for each position based on the correlation value with the 'overall rating':

Forward Position	skill_ball_control
	movement_reactions
	mentality_positioning
	shooting
	dribbling
	mentality_composure
	attacking_short_passing
	skill_dribbling
	attacking_finishing
	potential
	power_long_shots
	passing
	power_shot_power
	attacking_volleys
	mentality_vision

Defence Position	Movement_reactions
	Defending
	Defending_standing_tackle
	Mentality_interceptions
	Defending_sliding_tackle
	Defending_marking_awareness
	Mentality_composure
	Attacking_short_passing
	Potential
	Skill_ball_control
	Skill_long_passing
	Passing
	Mentality_aggression
	Attacking_heading_accuracy

Midfielder Position	skill_ball_control
	movement_reactions
	attacking_short_passing
	passing
	dribbling
	mentality_composure
	potential
	skill_dribbling
	mentality_vision
	skill_long_passing
	shooting
	power_shot_power
	mentality_positioning
	power_long_shots
	attacking_crossing
	skill_curve
	attacking_finishing

Goal keeper	goalkeeping_reflexes
	Goalkeeping_diving
	Goalkeeping_positioning
	Goalkeeping_handling
	Movement_reactions
	Potential
	Goalkeeping_kicking
	Power_shot_power

Best player for the forward position were identified by creating a column called ‘rating’ where it is measure of mean of the affecting features of the player.
Best forward position Players

Cristiano Ronaldo
R. Lewandowski
K. Mbappé,
H. Kane
K. Benzema
S. Agüero
W. Ben Yedder
Roberto Firmino
K. Volland
R. Lukaku

Best defence players:

Sergio Ramos
Marquinhos
V. van Dijk
M. Hummels
Rúben Dias,
A. Laporte
D. Alaba
Piqué
L. Bonucci
H. Maguire

Best midfielder players:

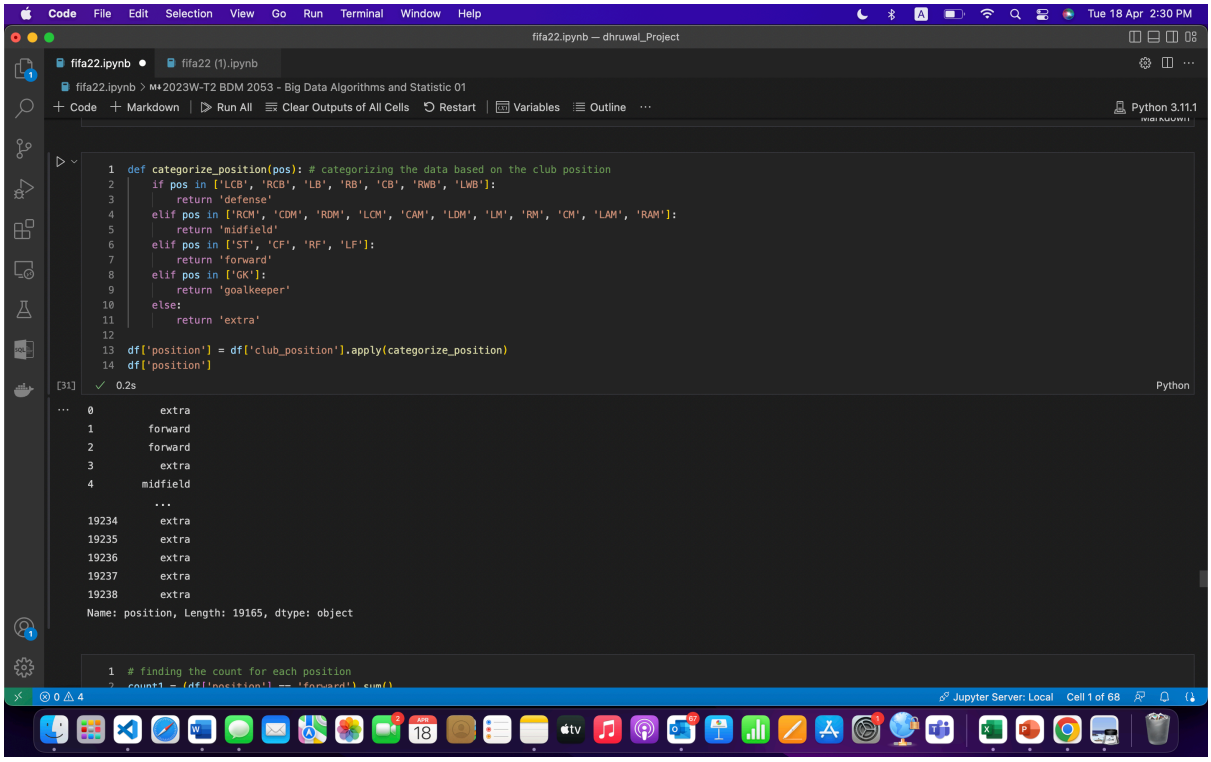
K. De Bruyne
Bruno Fernandes
T. Kroos
P. Dybala
Parejo
P. Pogba
L. Modrić
M. Reus
T. Müller
İ. Gündoğan

Best Goal Keepers:

M. Neuer,
Ederson
M. ter Stegen
J. Oblak
Alisson
G. Donnarumma
M. Maignan
T. Courtois
K. Casteels
Y. Sommer

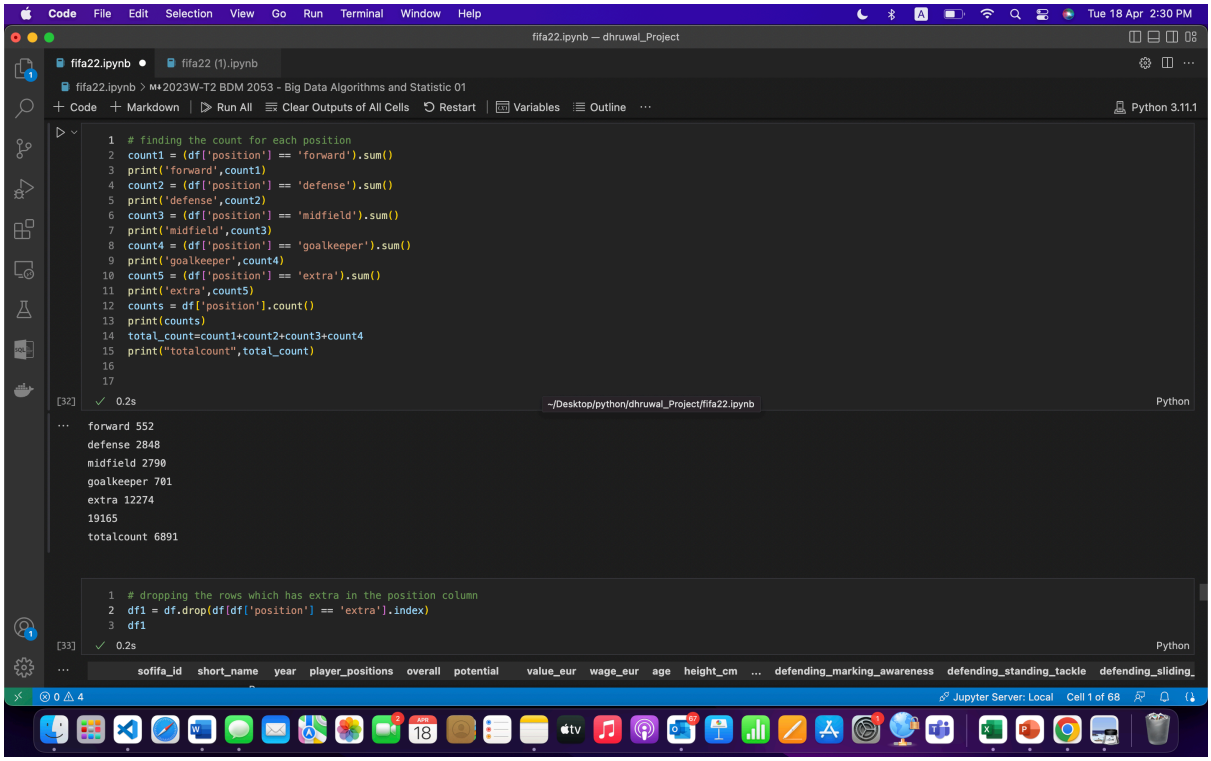
Pre-processing for data Modelling:

Categorizing the data by labelling based on their positions



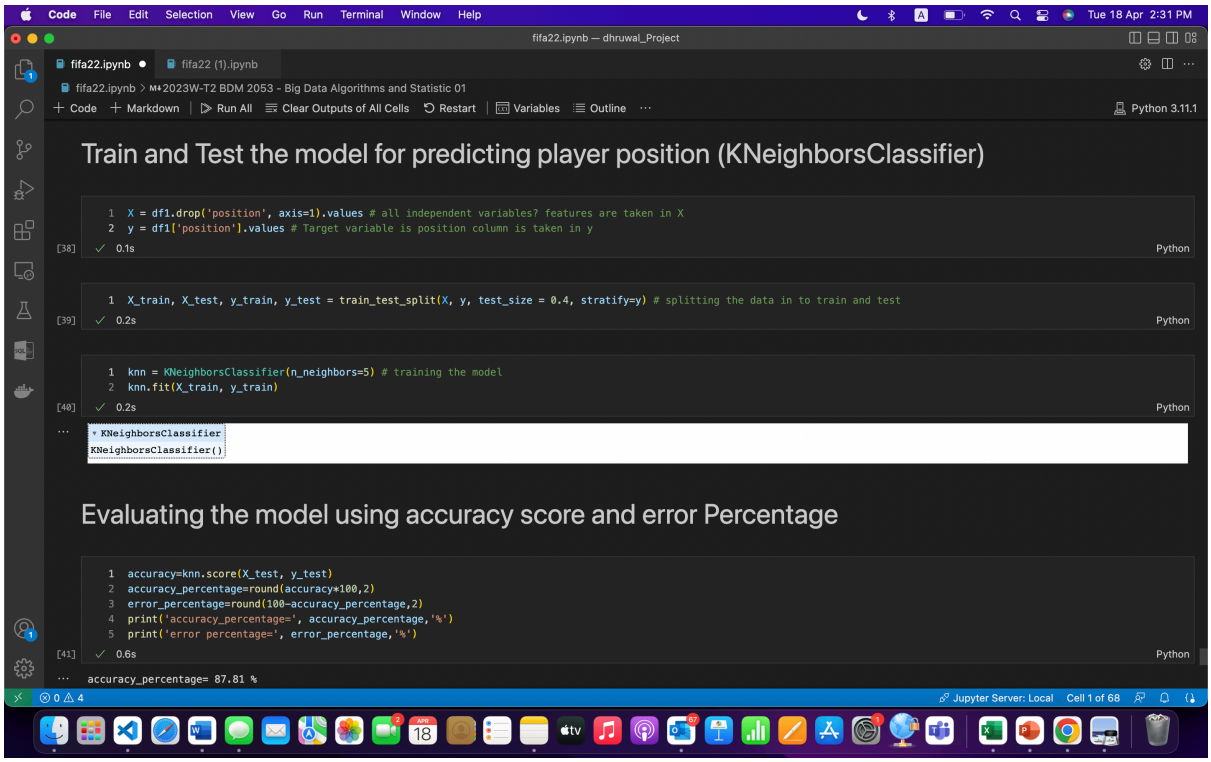
Counting the number of players for each positions:

Positions(Labelling)	**Count(No. of players)**
Forward players	552
Defence players	2848
midfielders	2790
goalkeepers	701
extra	12274



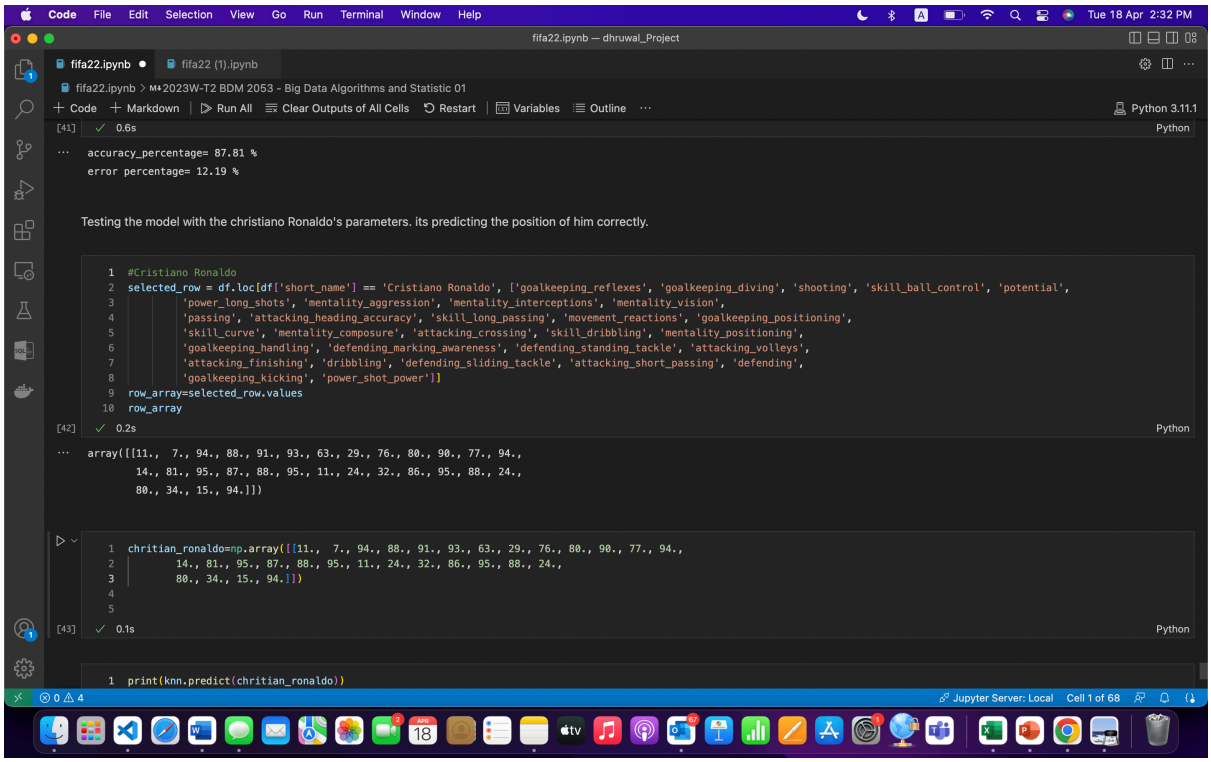
5.3 New Data Frame for Predicting Model:

- Model will predict the players position based on their parameters / features.
- Knn Model is a classification model which is best model to use it categorical value. Data was split into train and test.

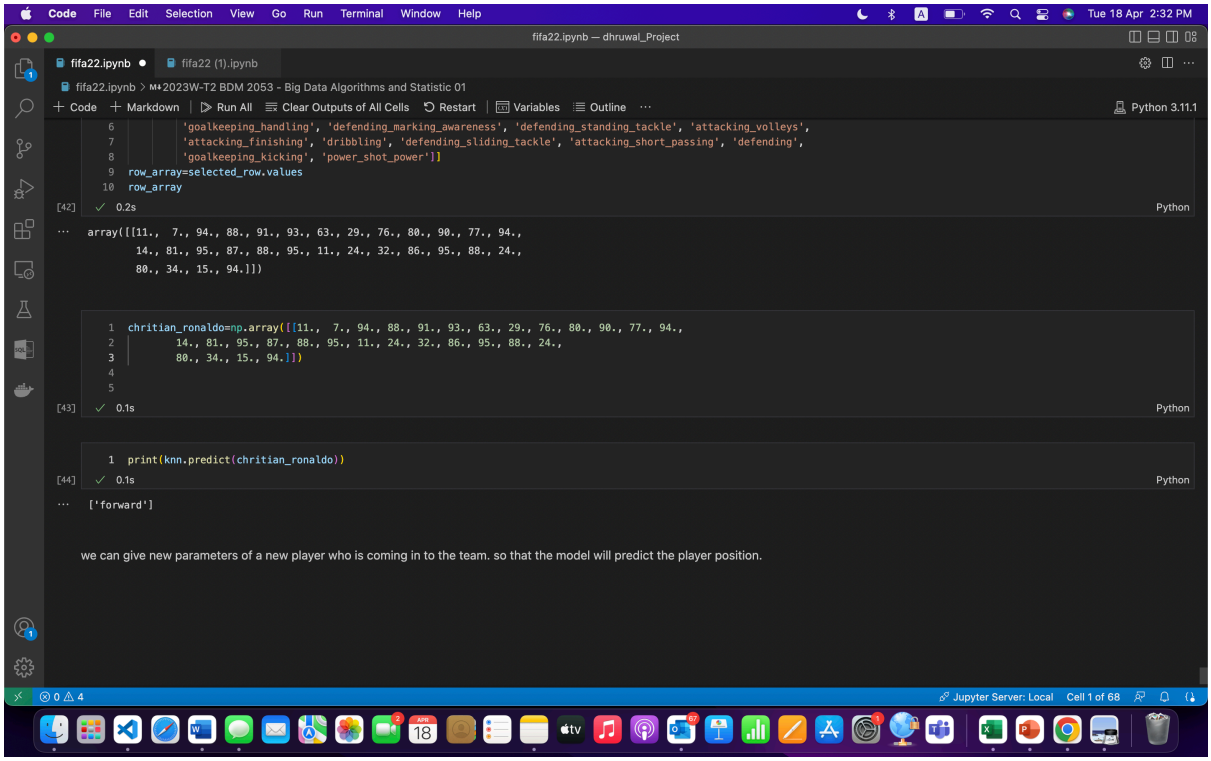


5.4 Model Evaluation:

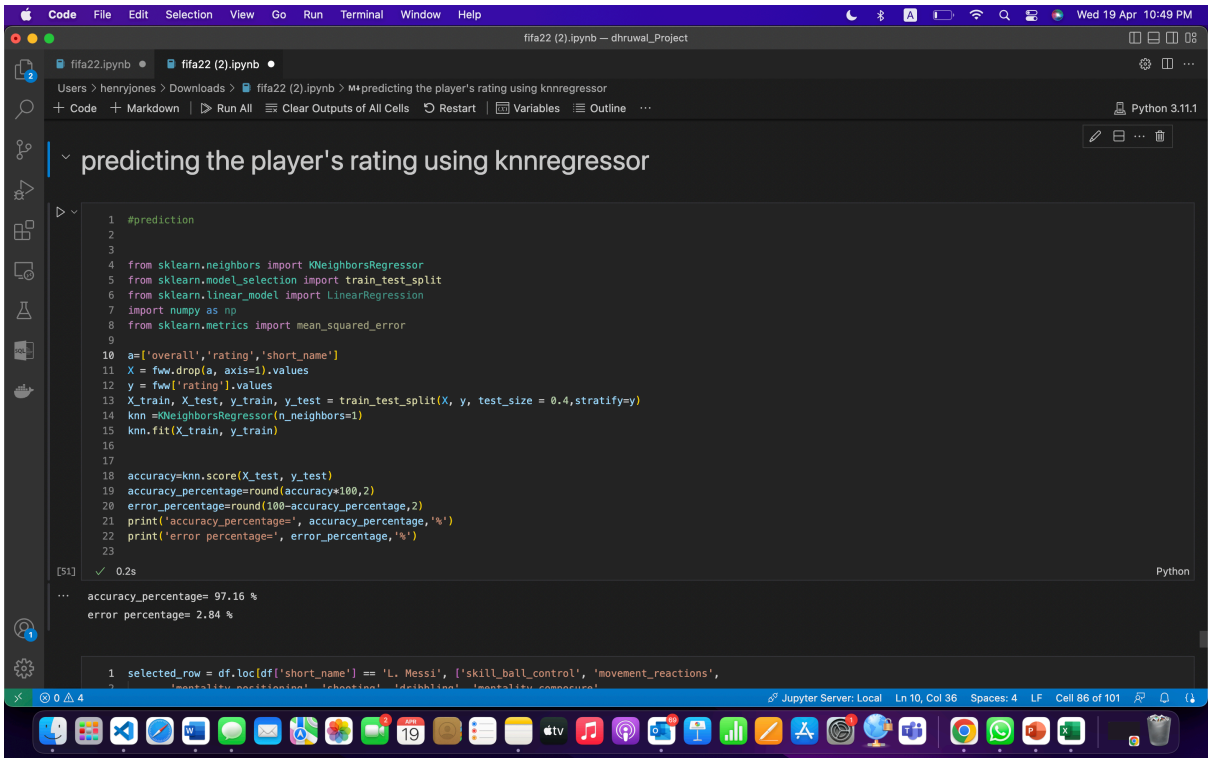
Model is performing well and it was evaluated with accuracy of 87.81% and error percentage of 12.19%



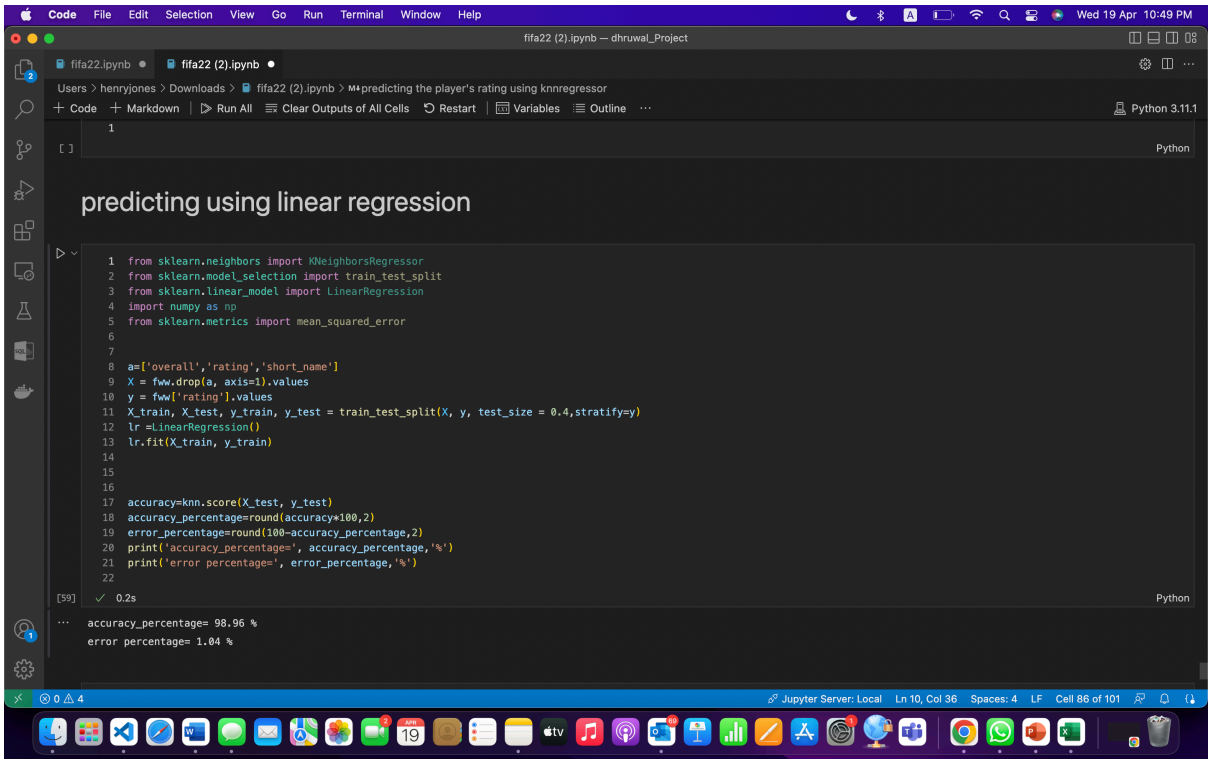
Tested the model with the parameters of the player it predicts the player position accurately. Model correctly predicts the Cristiano Ronaldo's position.



Predicting the Players Rating using Knn Regressor



Predicting the players rating using Linear regression



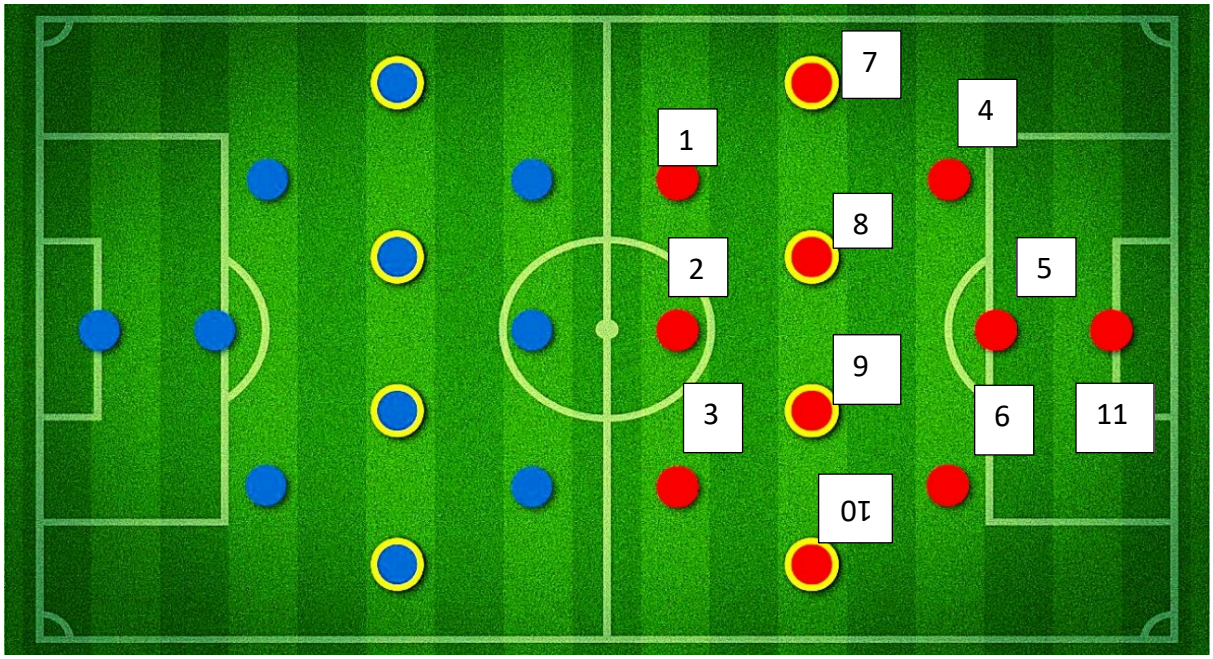
Two algorithm were used to predict the players rating. Linear regression model performed better with the accuracy of 98.96% and the error percentage is 1.04 % whereas KNN regressor model accuracy is 97.16% and error percentage is 2.84%. Linear regression model performed better.

6. Outcome:

6.1 Best team list

Top players for each positions are taken to form the team.

S.No	Players	Positions
1.	Cristiano Ronaldo	Forward position
2.	R. Lewandowski	Forward position
3.	K. Mbappé	Forward position
4.	Sergio Ramos	Defence Position
5.	Marquinhos	Defence Position
6.	V. van Dijk	Defence Position
7.	P. Dybala	Mid Fielder
8.	T. Kroos	Mid Fielder
9.	K. De Bruyne	Mid Fielder
10.	Bruno Fernandes	Mid Fielder
11.	M. Neuer	Goal Keeper



6.2 Predicting Player's Position:

A machine learning approach used for classification and regression analysis is called K-Nearest Neighbours (KNN). In KNN, a new data point's class is determined by the training data's k-nearest neighbours' dominant class. The k-nearest neighbours are chosen, and the class label is then assigned based on the majority of their labels using the algorithm, which first calculates the distances between the new data point and every other point in the training set. The classifier's accuracy may be impacted by the choice of k value, so making that decision is crucial. KNN is a flexible method that is both straightforward and efficient.

Knn Model is trained to predict the player position. This will help the franchise or team to put the player in correct position if the player joins the team newly. Players position is one of the key factor to affect the game result. Hence it will be very much helpful to increase the standard.

6.3 Predicting Player's rating:

KNN regressor:

Machine learning algorithms for regression tasks include the K-Nearest Neighbours (KNN) regressor. By averaging the output values of a new data point's k-nearest neighbours in the training data, the algorithm may predict the output variable of that new data point. In order to accomplish this, the algorithm computes the distances between the new data point and each of the other points in the training set, chooses the k-nearest neighbours, and predicts the output value using the average of those neighbours' output values. The accuracy of the regressor can be impacted by the k value selection, which is crucial. KNN regressor is a straightforward yet powerful method that may be applied to many regression issues.

Linear Regression:

A machine learning approach called linear regression is used to forecast a continuous output variable based on one or more input variables. Finding a linear link between the input and output variables is how it operates. By reducing the sum of the squared errors between the expected and actual values of the output variable, the algorithm calculates the linear equation's parameters. The model can be used to forecast the output variable for fresh input values once the parameters have been estimated. It is possible to estimate future sales and predict property prices using the straightforward yet effective linear regression algorithm.

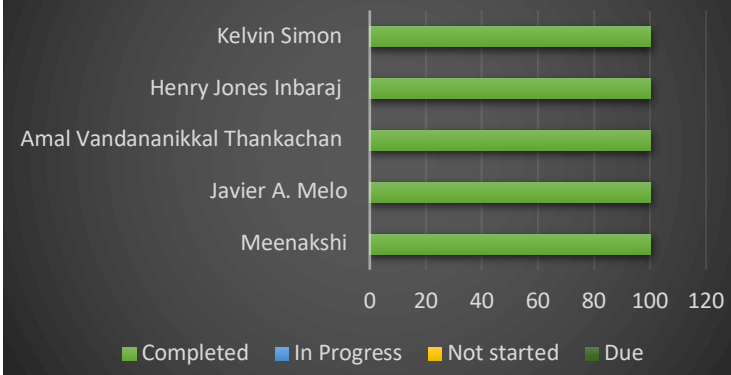
Two algorithm were used to predict the players rating but linear regression model performed better with the accuracy of 98.96% and the error percentage is 1.04 % whereas KNN regressor model accuracy is 97.16% and error percentage is 2.84%.

7. Project Status Final Report

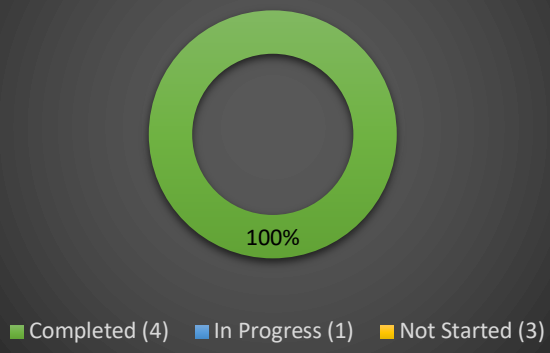
Big Data Analysis for Football Team Formation

Project Status: Completed. **Date:** 13th April

Workload Status

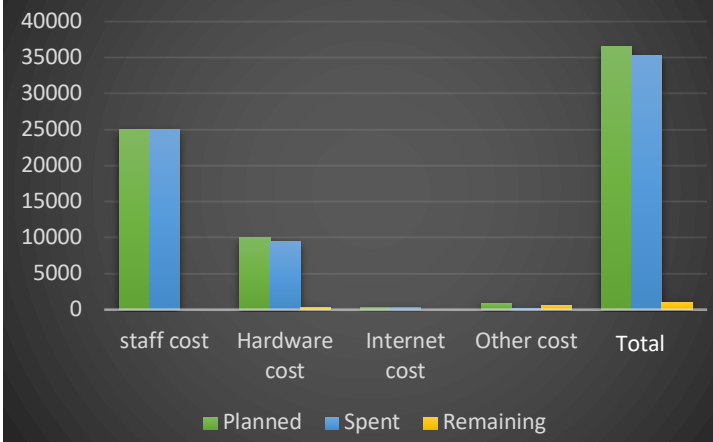


Milestone Achieved



Project Title (03/16/2023)	Finding the data (03/23/2023)	Project objective and planning (03/30/2023)	Explore data (04/04/2023)	Model Selection (04/10/2023)	Train and Test the Data (04/11/2023)	Model Evaluation (04/12/2023)	Final Report (04/13/2023)

Cost



Overall Health:

Time	Completed within deadline
Tasks	All task completed
Workload	0 tasks overdue
Progress	100% complete
Cost	4% under Budget

Summary and Key Highlights:

- Top Players were identified by finding the statistics.
- Best team was formed.
- Game plan based on predicting player ‘s position
- Model predicting the player’s rating to elevate the game standard.
- Project was successfully completed within the deadline

8. Conclusion:

We have achieved all our milestone, mitigated all risks, bounced from the challenges, cost efficiently budget was under control and successfully completed the project. Our Project found the best players and recommended the best team out of it. We have also two models to predict the player's position to plan the game strategy and to find the player's rating to elevate the game standard.

This will definitely will bring a humongous change in the industry to make the game interesting which will increase the game popularity, views and the sponsorship.

We have planned to implement this project to other games as well to benefit the industry more.

Reference:

Data set link: https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset?select=players_22.csv

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

https://moodle.queenscollege.ca/moodle/pluginfile.php/1625917/mod_resource/content/7/Week5_Modified.pdf

https://moodle.queenscollege.ca/moodle/pluginfile.php/1625908/mod_resource/content/5/Week2_Modified.pdf

https://moodle.queenscollege.ca/moodle/pluginfile.php/1625905/mod_resource/content/1/Week1%20notes%20.pdf

https://moodle.queenscollege.ca/moodle/pluginfile.php/1625926/mod_resource/content/9/Week7_Modified.pdf