DS 303 Project

Position Predictor in Football

Avish 180100023 | Rishika 190040090
Kaushikraj 20D100012 | Neeraj 20D070056

HONOR CODE

By the Honor Code, we pledge that this project is our own work. We have not copied from any sources available online.

Ideas are original, implementation is inspired from various libraries and articles.

Abstract

The Fédération Internationale de Football Association (FIFA) is a governing body of football (sometimes, especially in the USA, called soccer). FIFA is also a series of video games developed by EA Sports which faithfully reproduces the characteristics of real players. FIFA ratings of football players from the video game can be found at https://sofifa.com/ . Data from this website for 2019 were scrapped and made available at the Kaggle webpage FIFA 20 complete player dataset . We will use the data to build a predictive model for the evaluation of a player's position. Subsequently, we will use the model exploration and explanation methods to better understand the model's performance, as well as which variables and how to influence a player's value.

Problem Statement

The project objective is to provide statistical insights to football clubs worldwide. The project involves using EA Sports' FIFA video game data as a prototype for providing insight into real football. The actual data and statistics are expensive to obtain and often kept secret by the respective clubs' analytics teams.

The main focus of this project will be:

- Suggesting Playing positions based on player's statistics eg. age, strength, speed, agility, etc.
- Analyze how attributes like pace, shooting, dribbling, passing, etc. generally vary with different positions of players on the field.
- Analyzing how age affects the physical and skill-based statistics of players.

Technology landscape assessment

Most of the major football clubs have their statistics teams which are getting more and more important every season. Hiring a top statistics firm is increasingly becoming some of the club's top priority just like hiring a world-class manager.
Despite the acceptance statistics are getting in football, most of the known use of statistics seems to be rudimentary and not in-depth. passes completed, shots were taken, average traveled distance are some examples of popular statistics which influence player's stature and those are obviously justified but data can offer much more intuition than that. Young players' positions and market values are still determined by experienced scouts' intuitions and extensive trial and error methods.
There are a few firms that cater to the club's needs for in-depth data analysis. Most of the other data analysis firms focus on fantasy football and betting predictions rather than actual involvement in squad building and player development. There is a growing market for such service providers.

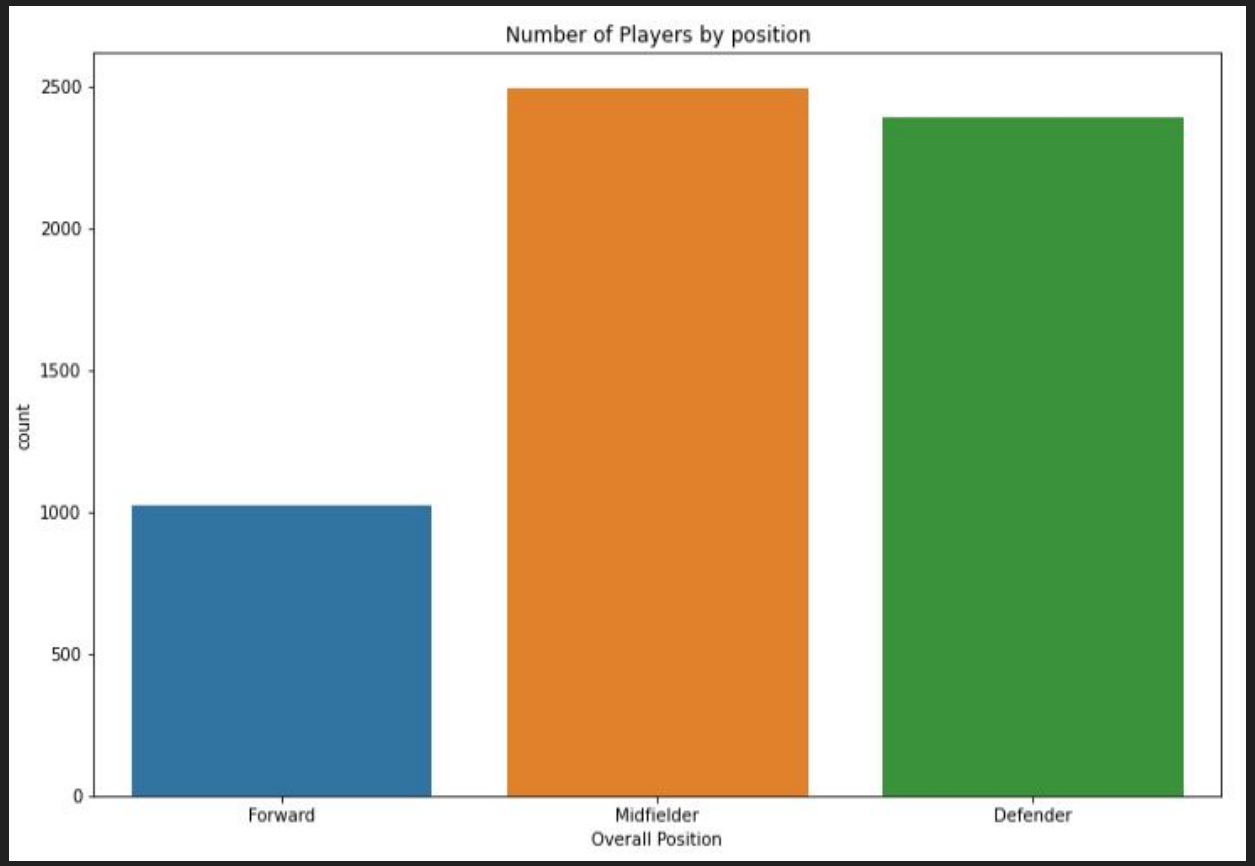
Data Sources and Details

- FIFA 19
- FIFA 20

Every player available in FIFA 19 and FIFA 20 with 100+ attributes, URL of the scraped player, player positions, with his role in the club and in the national team, Player attributes with statistics as Attacking, Skills, Defense, Mentality, GK Skills, etc. Player personal data like Nationality, Club, DateOfBirth, Wage, Salary, height, weight, age, etc.

Data Set Details and Preprocessing

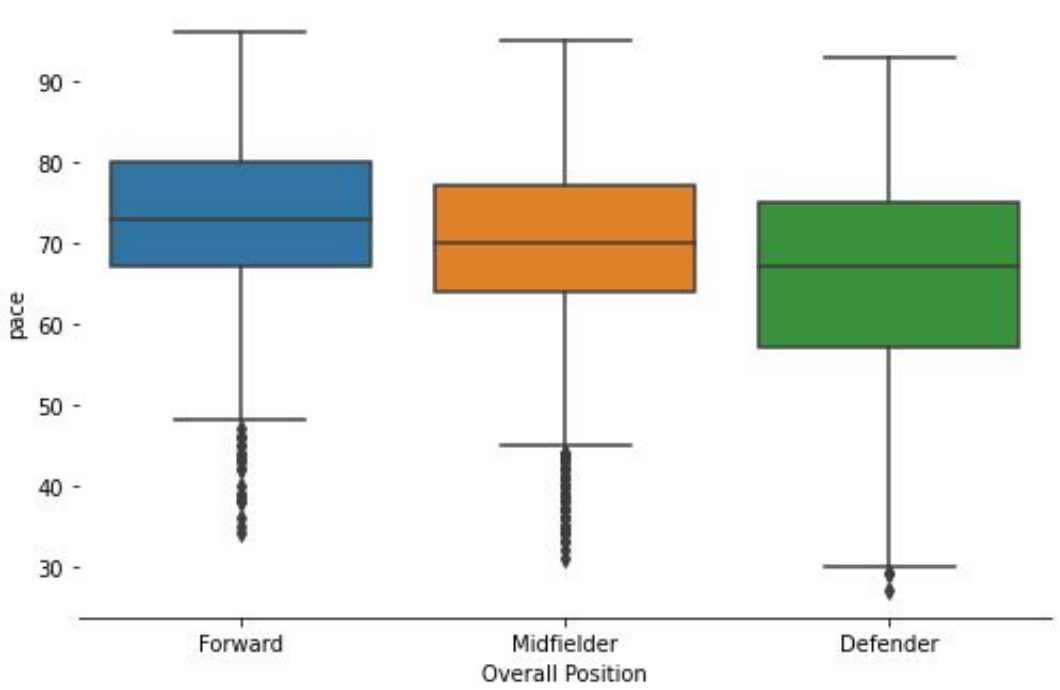
To build our model we had to remove the object type data like short name, ID, player URL, Last name, Player traits (playmaker, long-distance shooter, diver, leader, etc.), etc. The numerical data remained is thus used for the analysis. For calculating the position we used the 'team_position' attribute. All kinds of midfielders ie Attacking, right, left, defensive, etc. are given the position 'Midfielder' and similar for forwards and defenders.

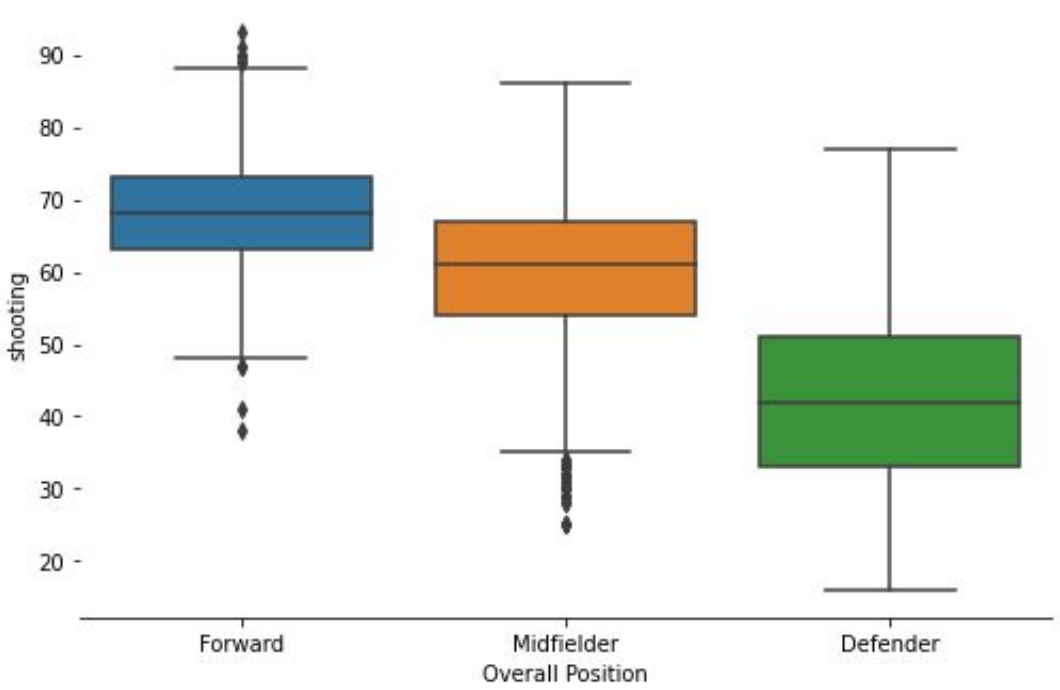


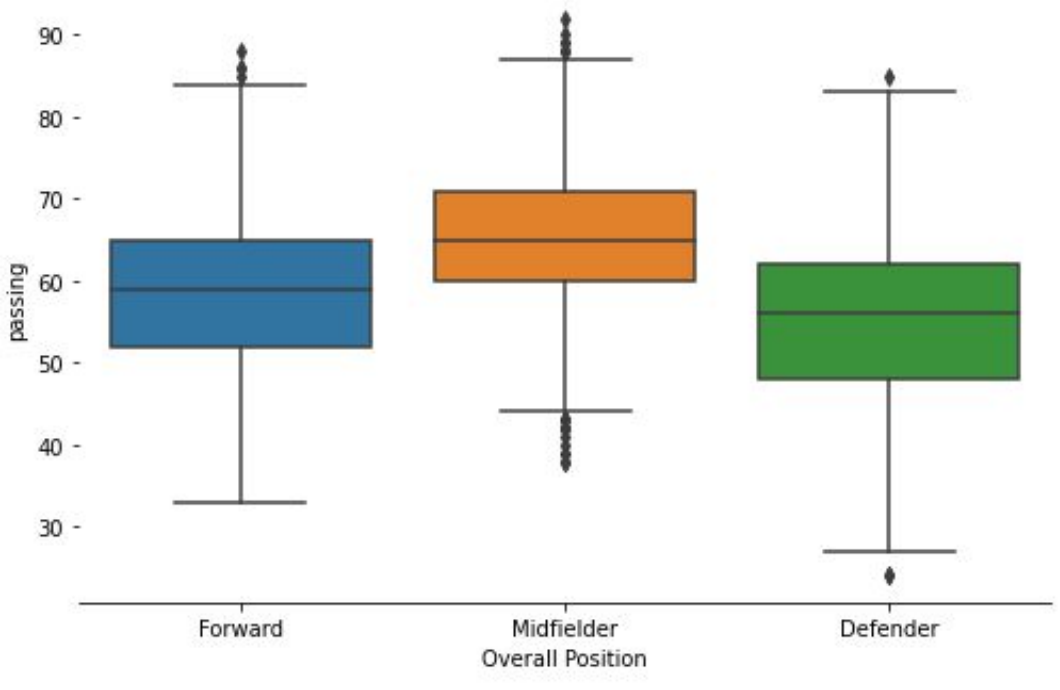
Data Set Details and Preprocessing

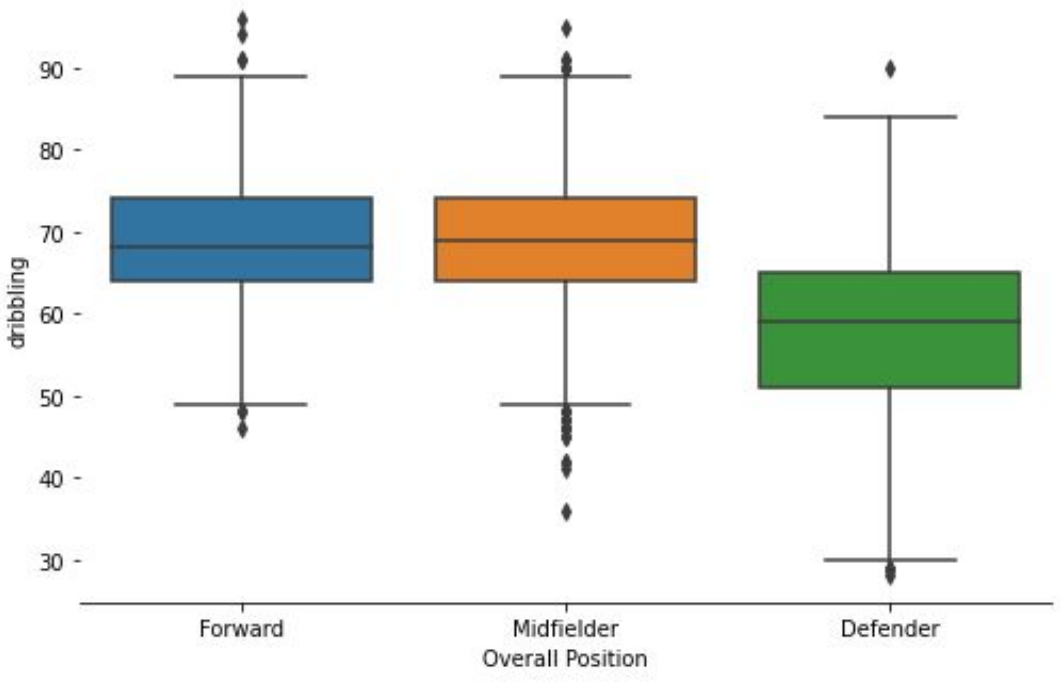
The data set also had ‘SUB’ and ‘RES’ (substitute and reserves) as ‘team_position’, due to which accuracy of any model will be very less because differentiating a midfielder substitute and a playing 11 midfielder has other aspects such as overall team statistics, the kind of players it contains, etc. We have not considered such attributes in our analysis and hence we have removed the substitutes and midfielders from our analysis which reduces the data set to nearly 5000-6000 players from 13000 players.

Attributes Variation with Position

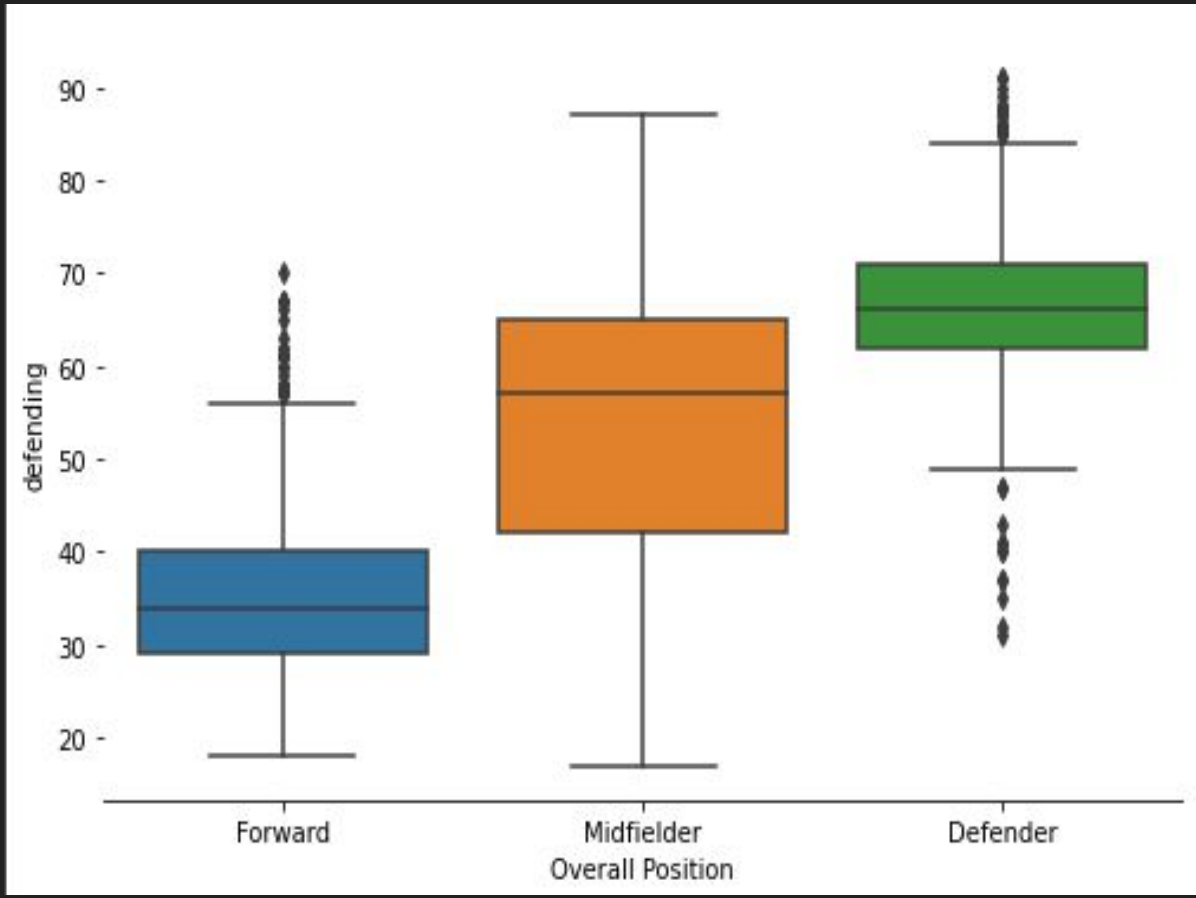


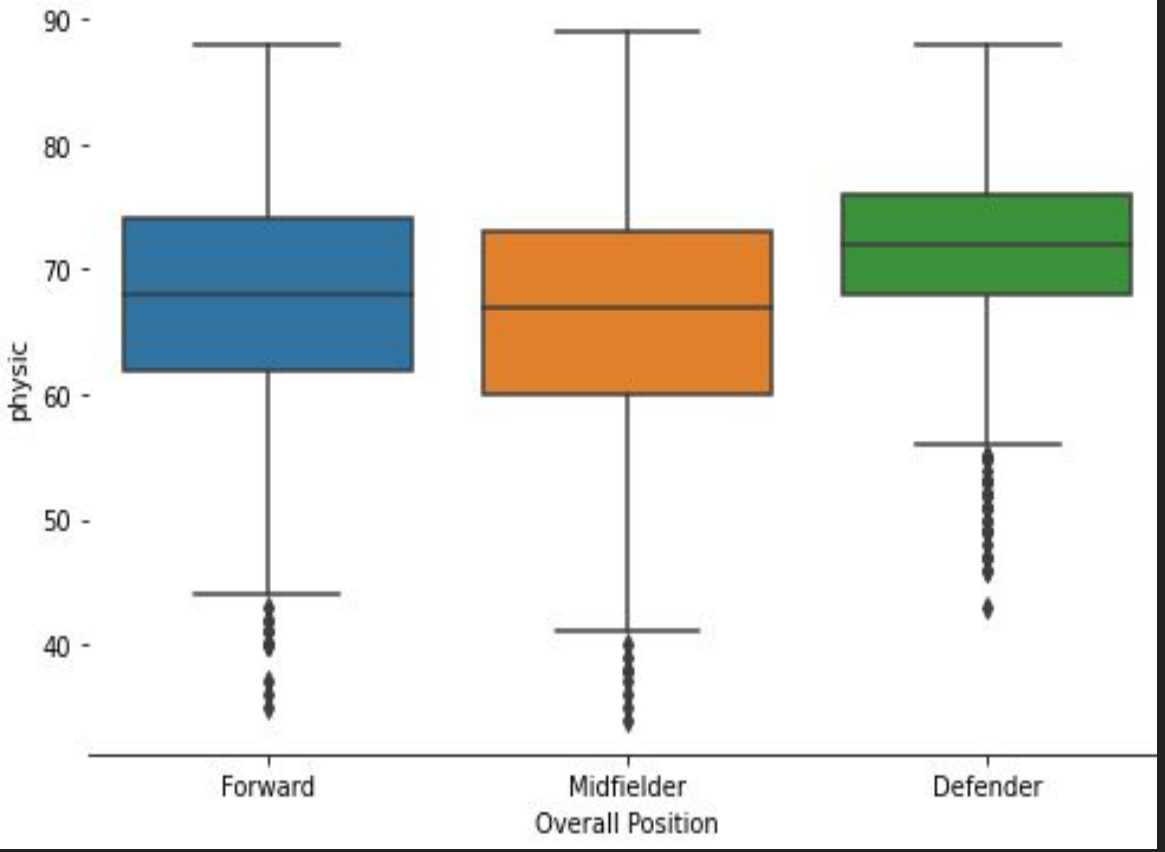




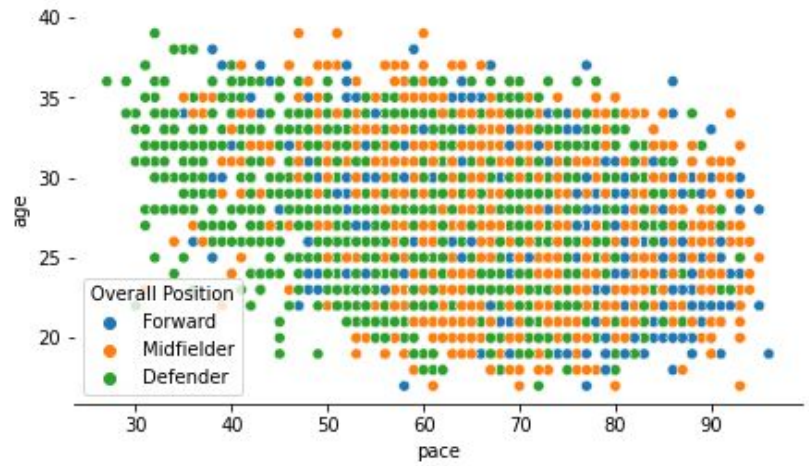


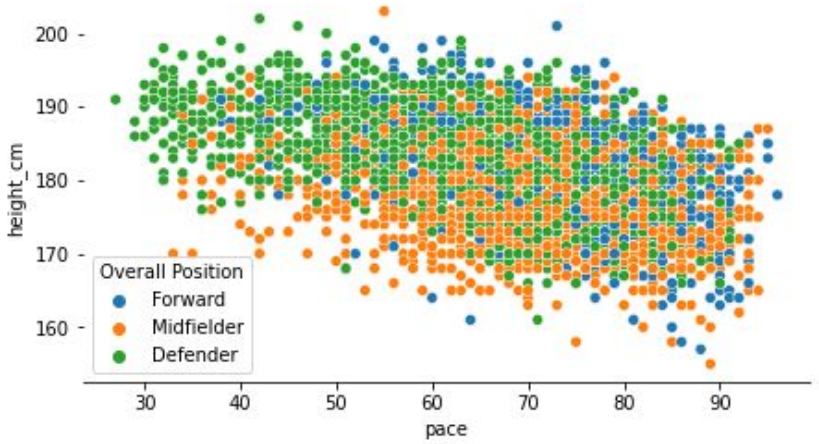
Attributes Variation with Position

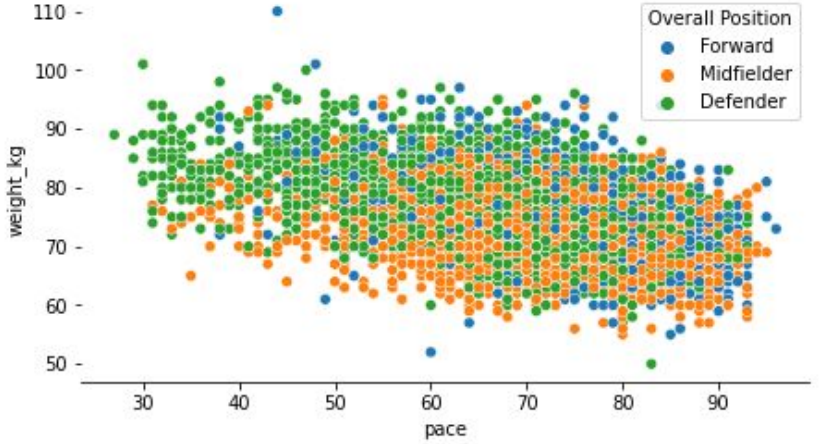


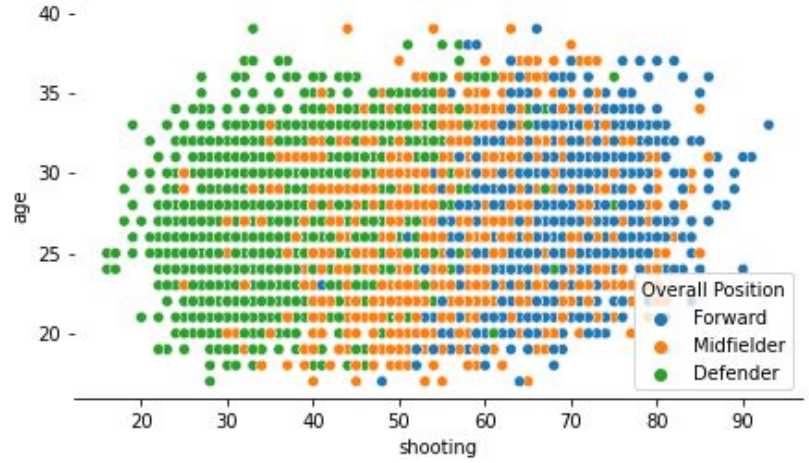


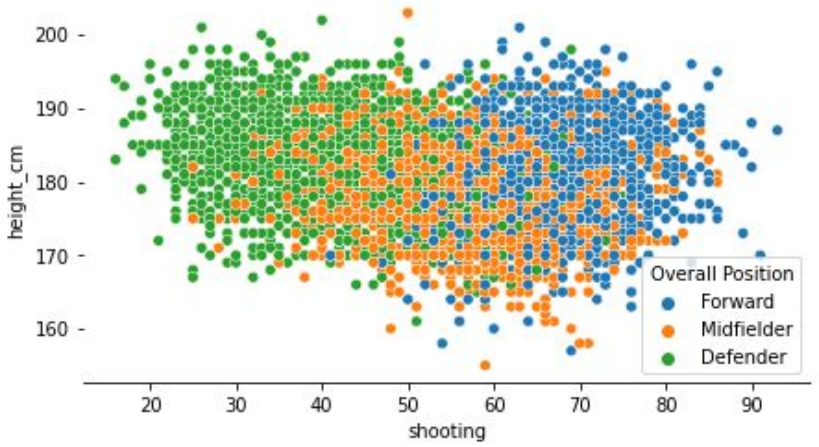
More Insight into Data set

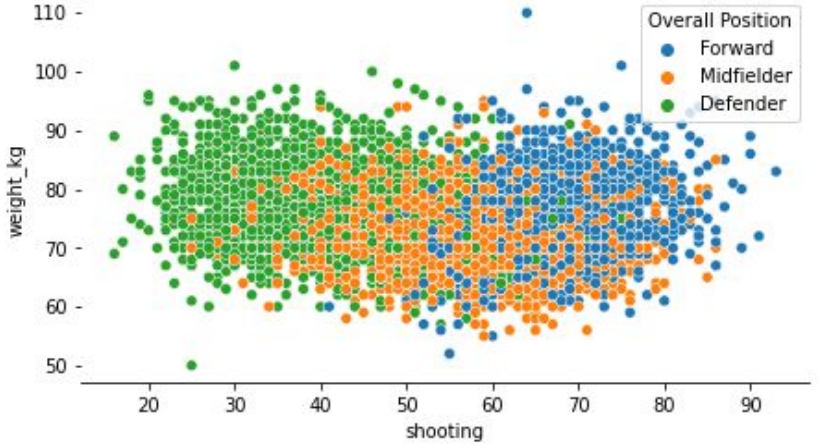




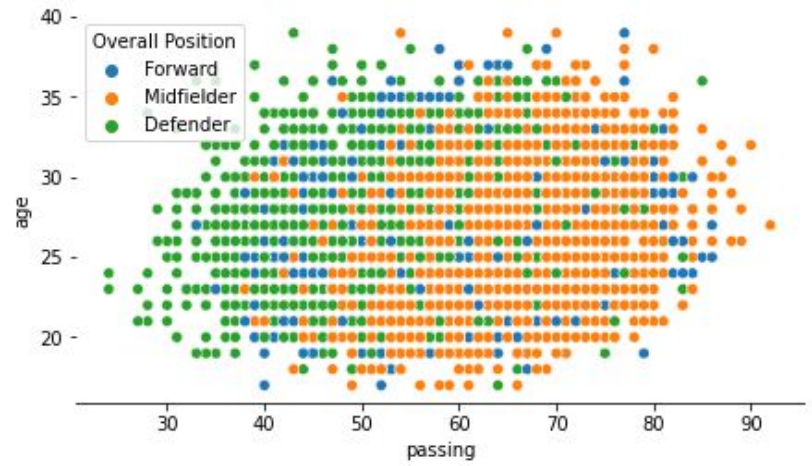


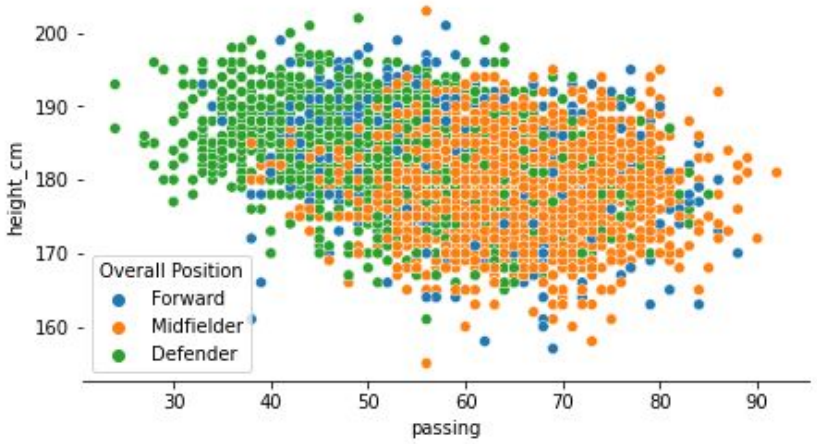


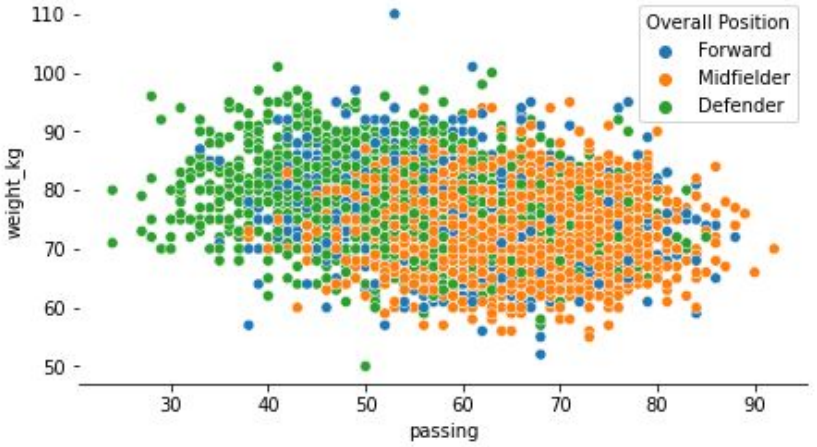


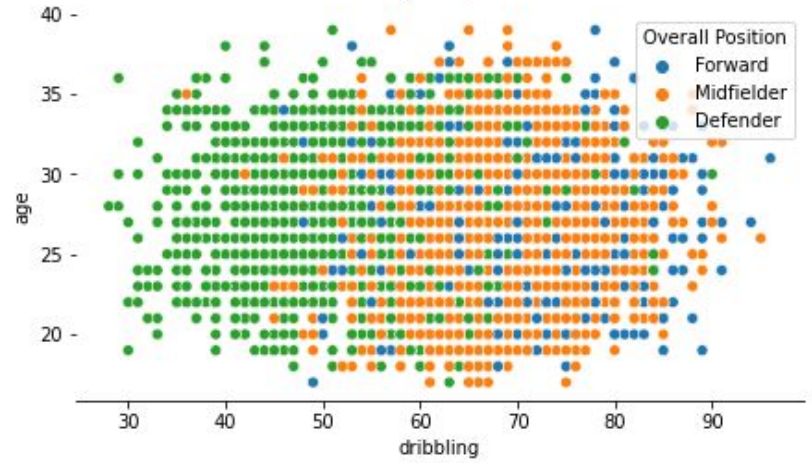


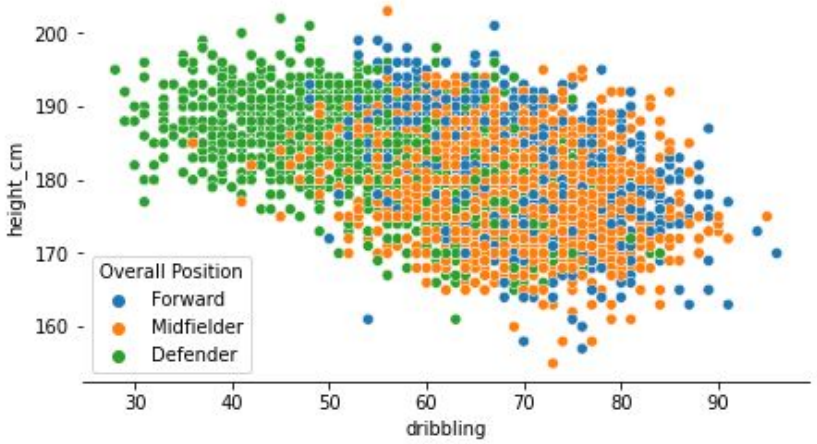
More Insight into Data set

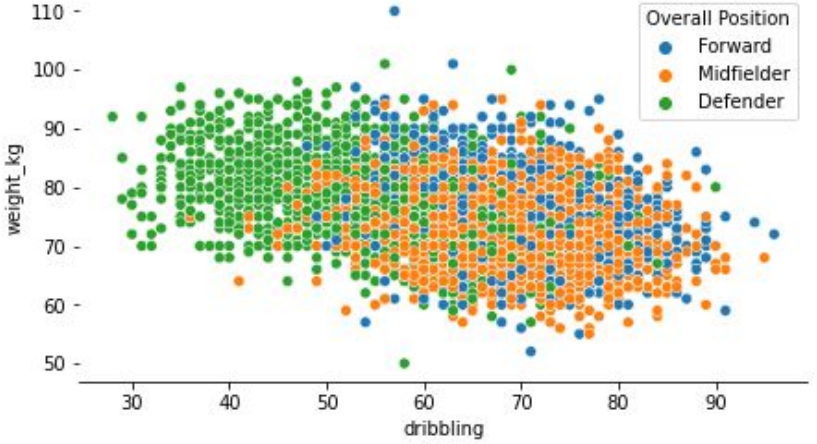




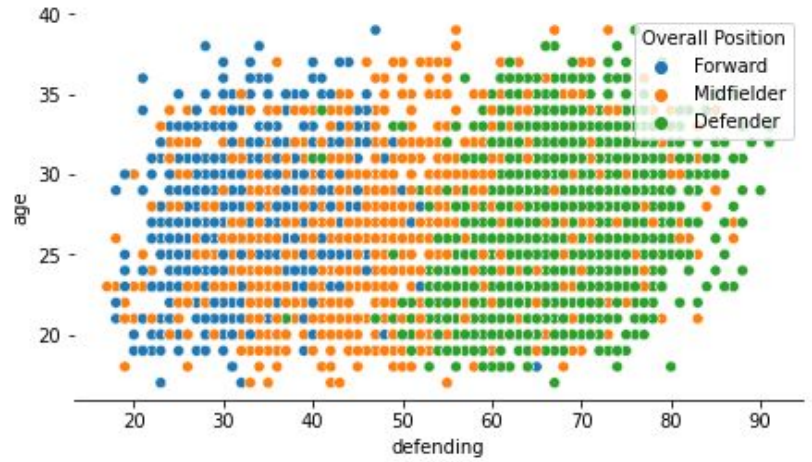


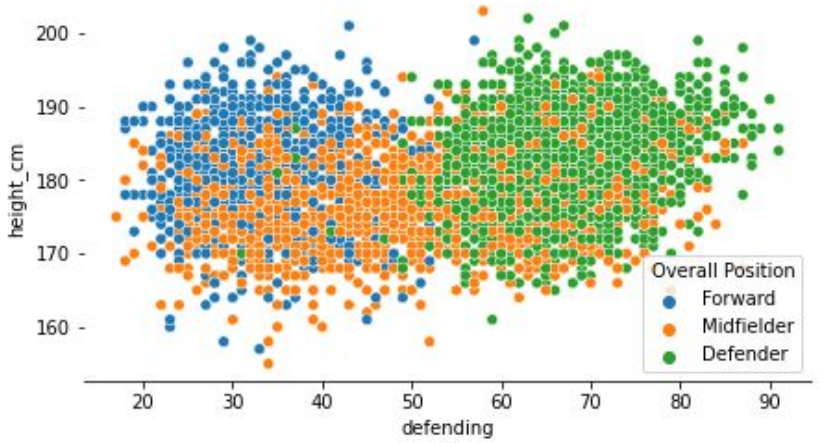


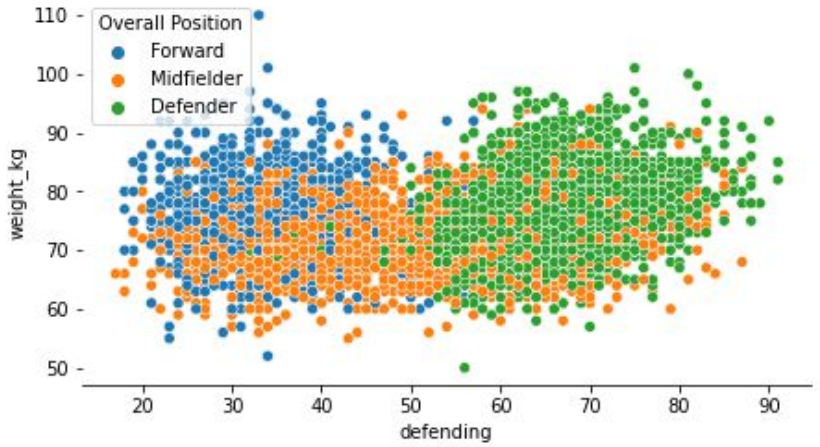


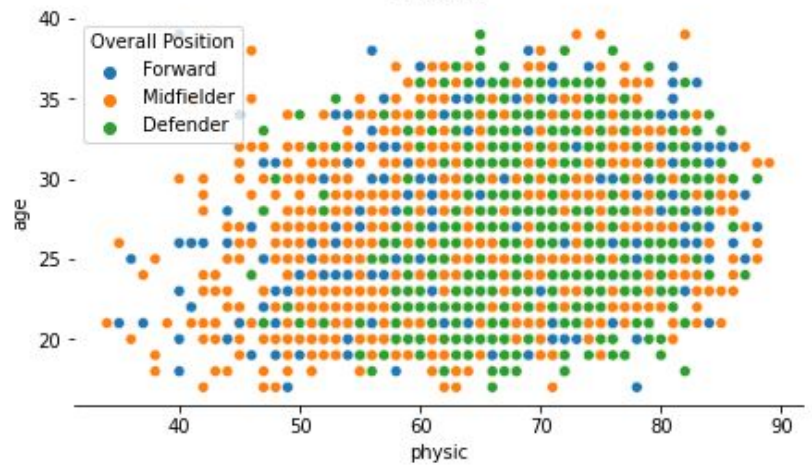


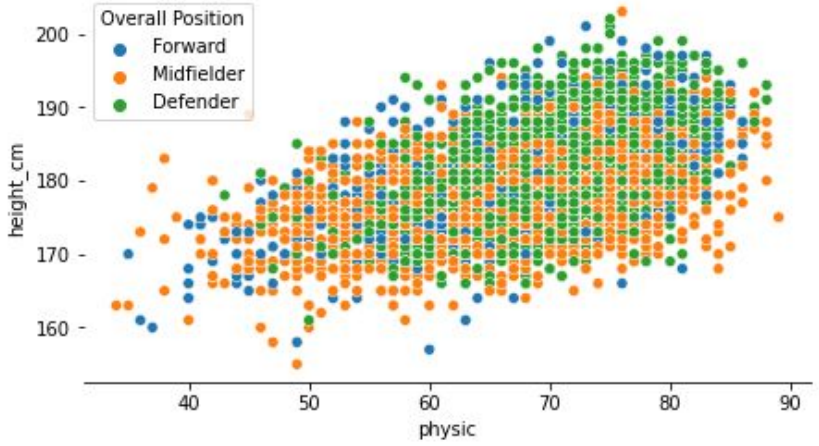
More Insight into Data set

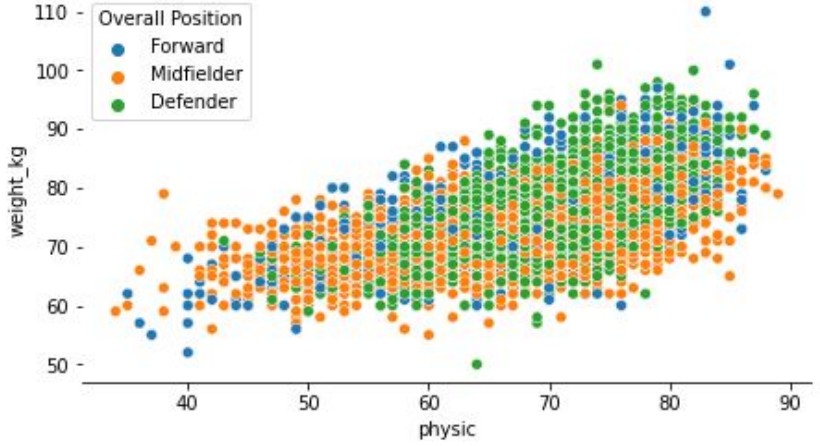




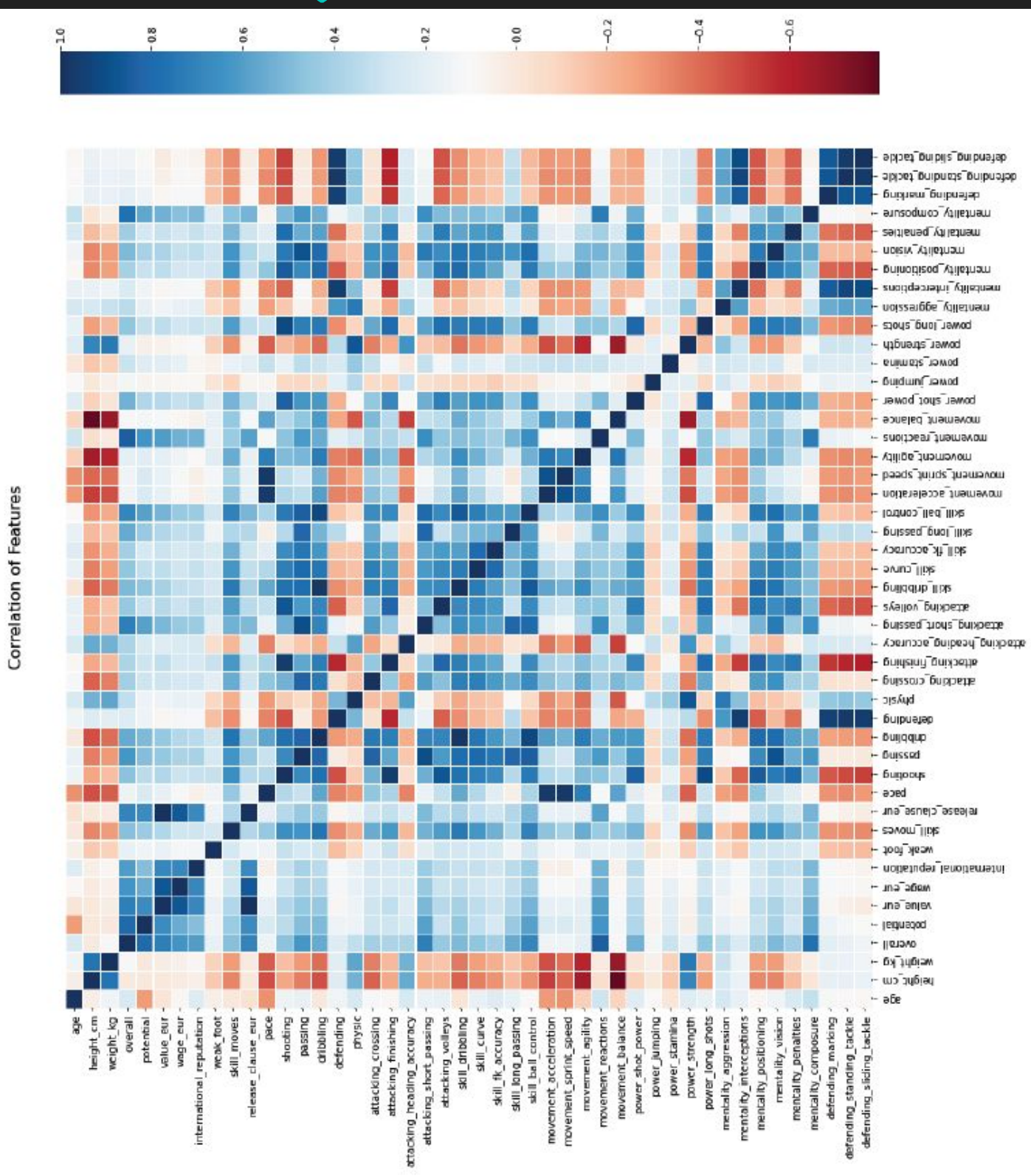






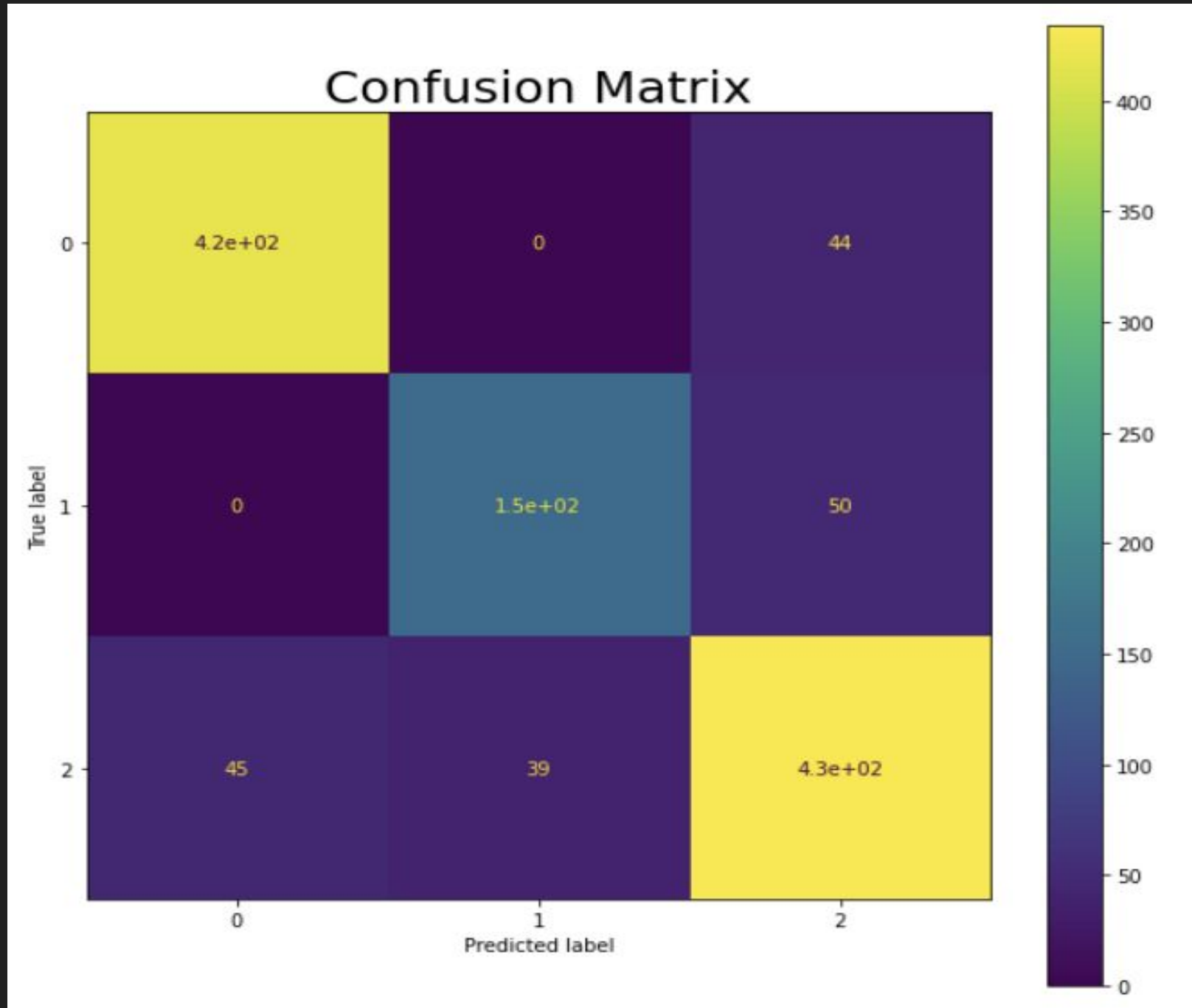


Covariance Matrix Heat Map

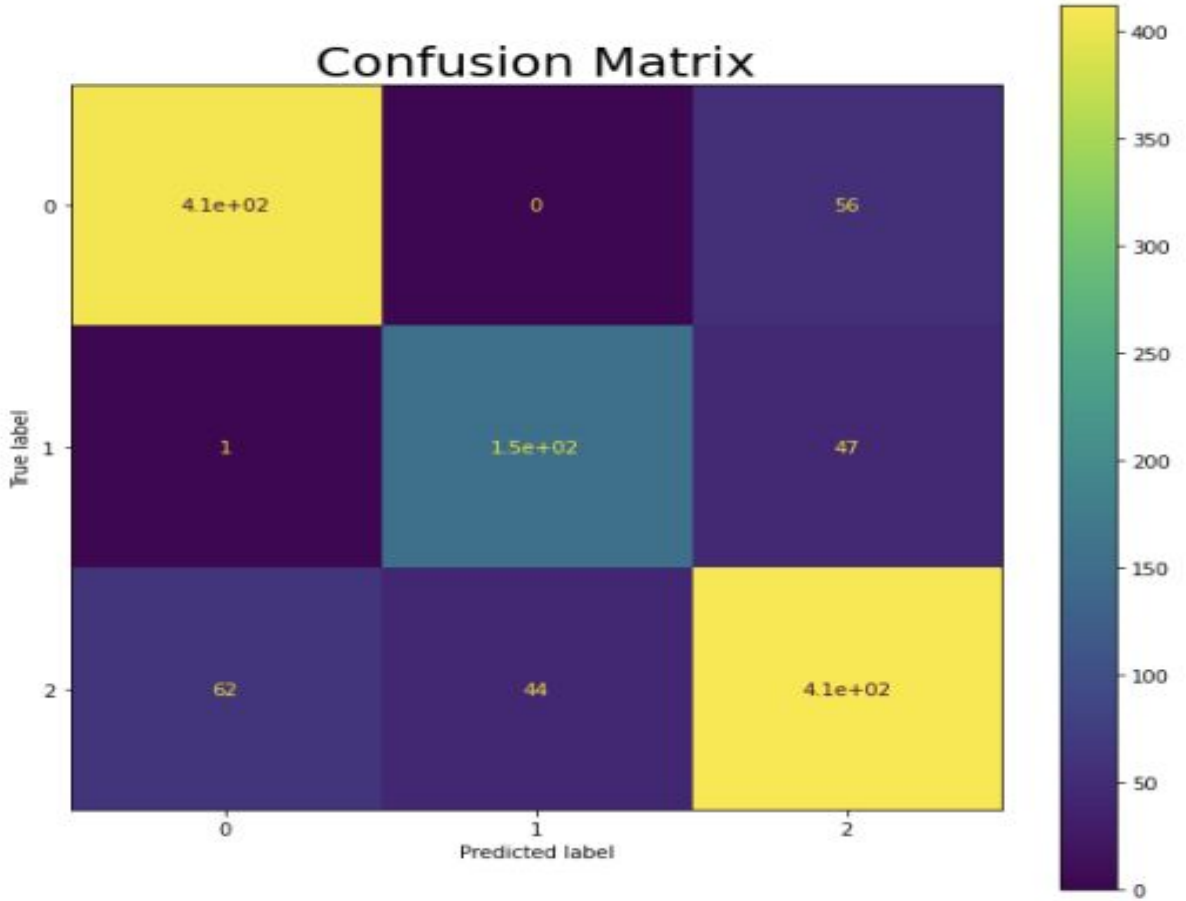


Logistic regression

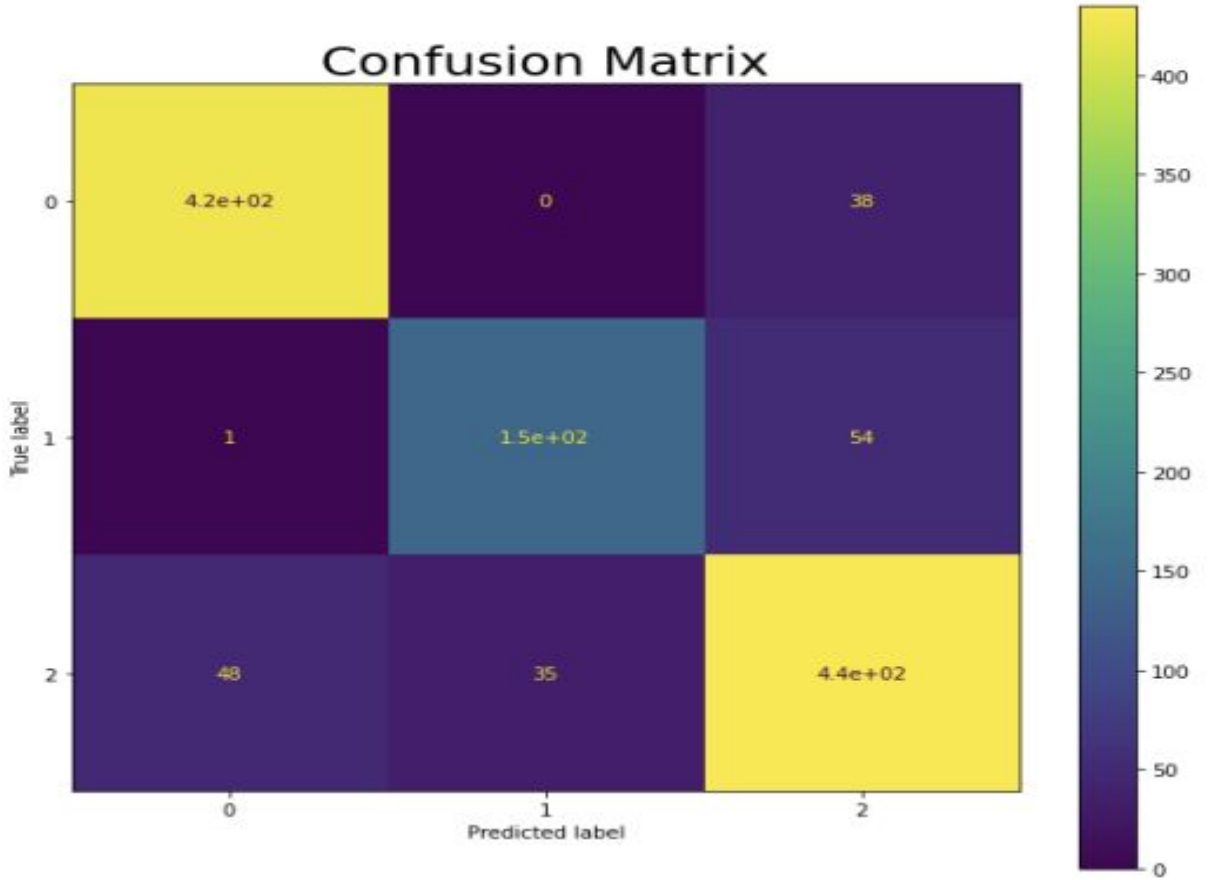
We applied our first model Logistic regression without cross-validation and with cross-validation of 5 folds. Surprisingly we got the same accuracy of 0.85 with both the models. Here we have the confusion matrix:



KNN: Then, we apply the KNN model without cross-validation and with grid search that applies to the K (number of neighbors) in a range from 1 to 25 with cross-validation of 5. KNN with cross-validation has higher accuracy of 0.85 bigger than KNN without cross-validation of 0.82, as expected. Shown below is the confusion matrix of both the models:



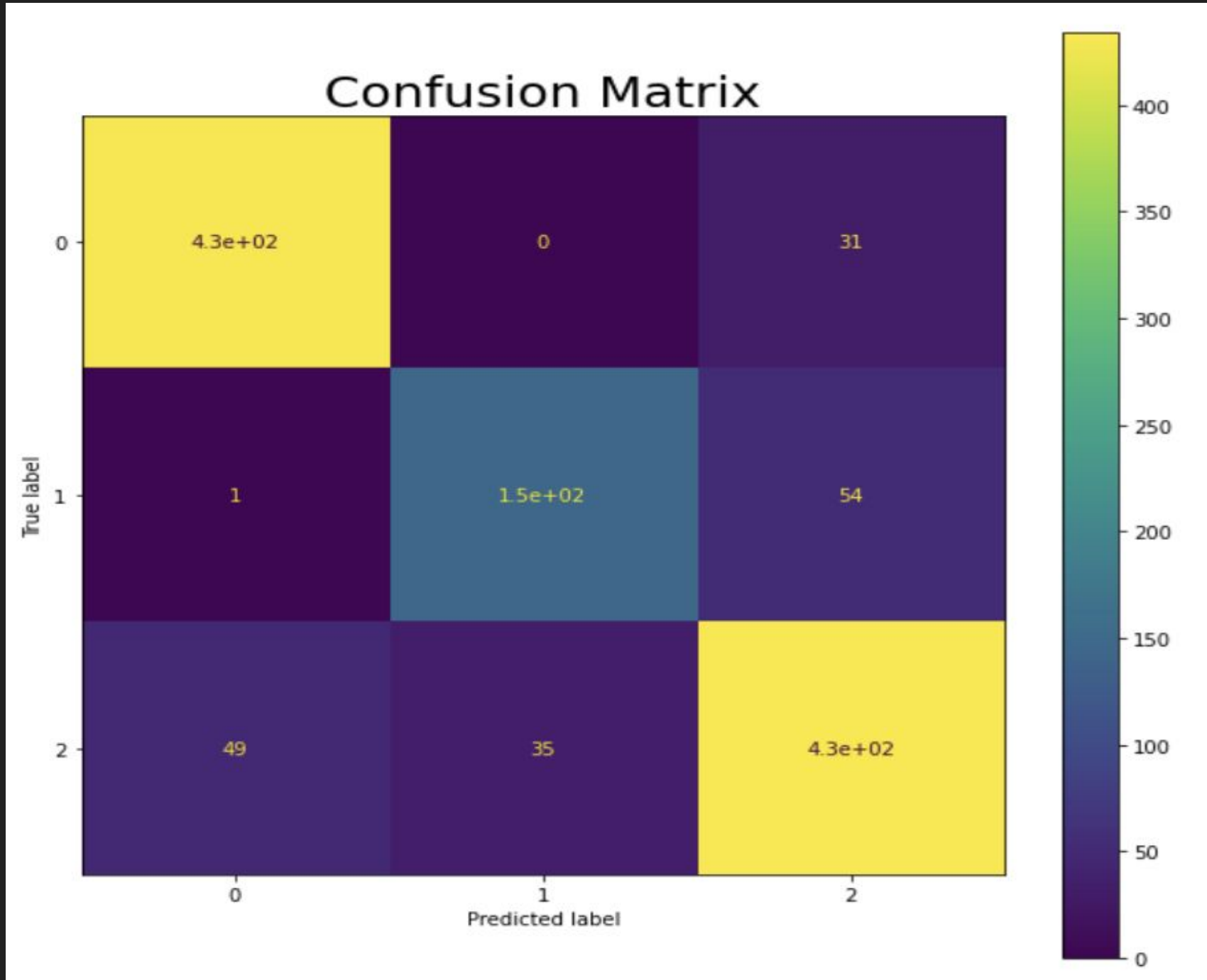
KNN with cross-validation



KNN without cross-validation

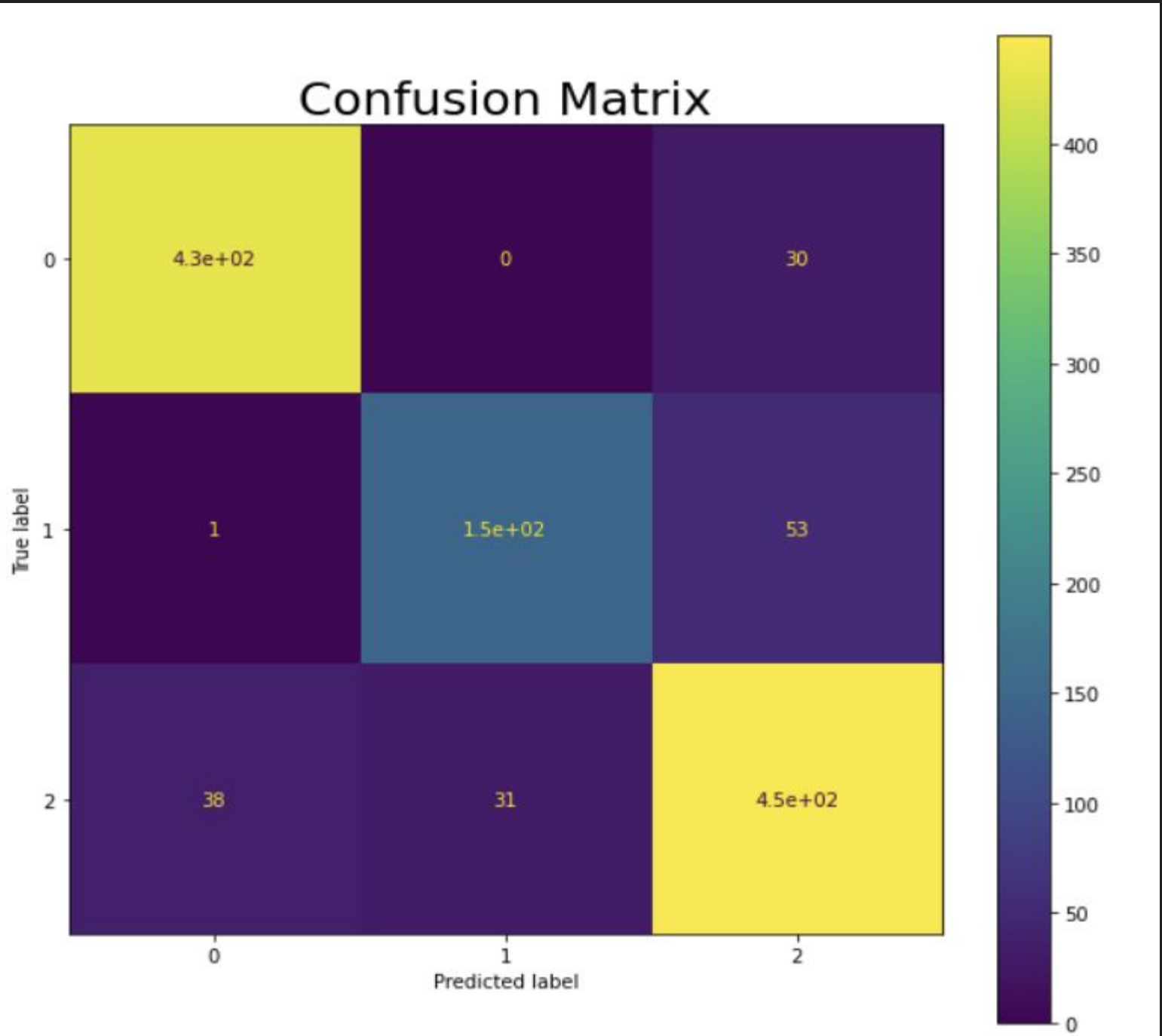
Random Forest Classifier

With random forest classifier, we achieved an accuracy of 85.62%, a little better than the previous two models.



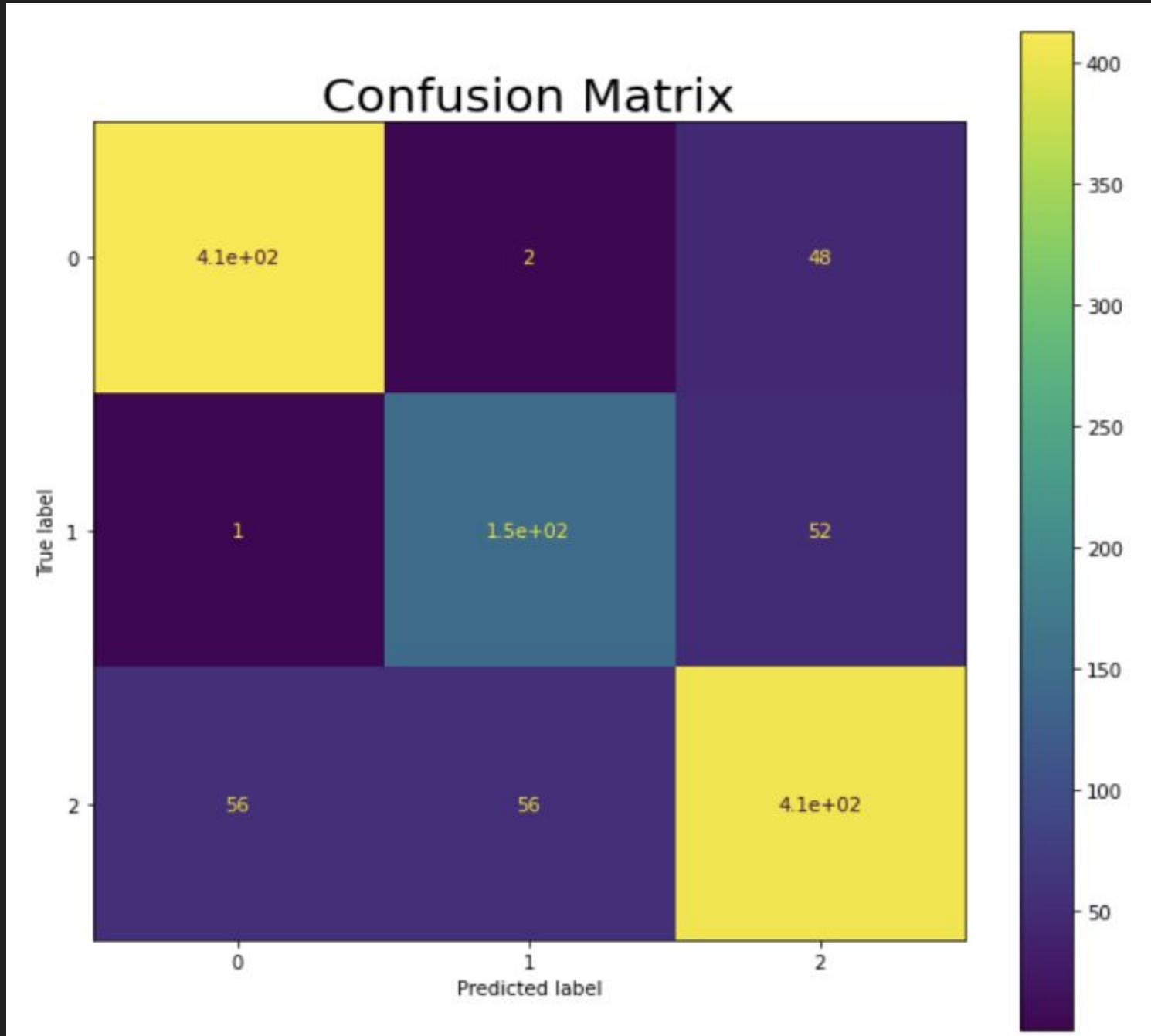
SVM Classifier

The accuracy of the model improved further to 87.06% using SVM Classifier. Its confusion matrix is as follows:



Neural Network

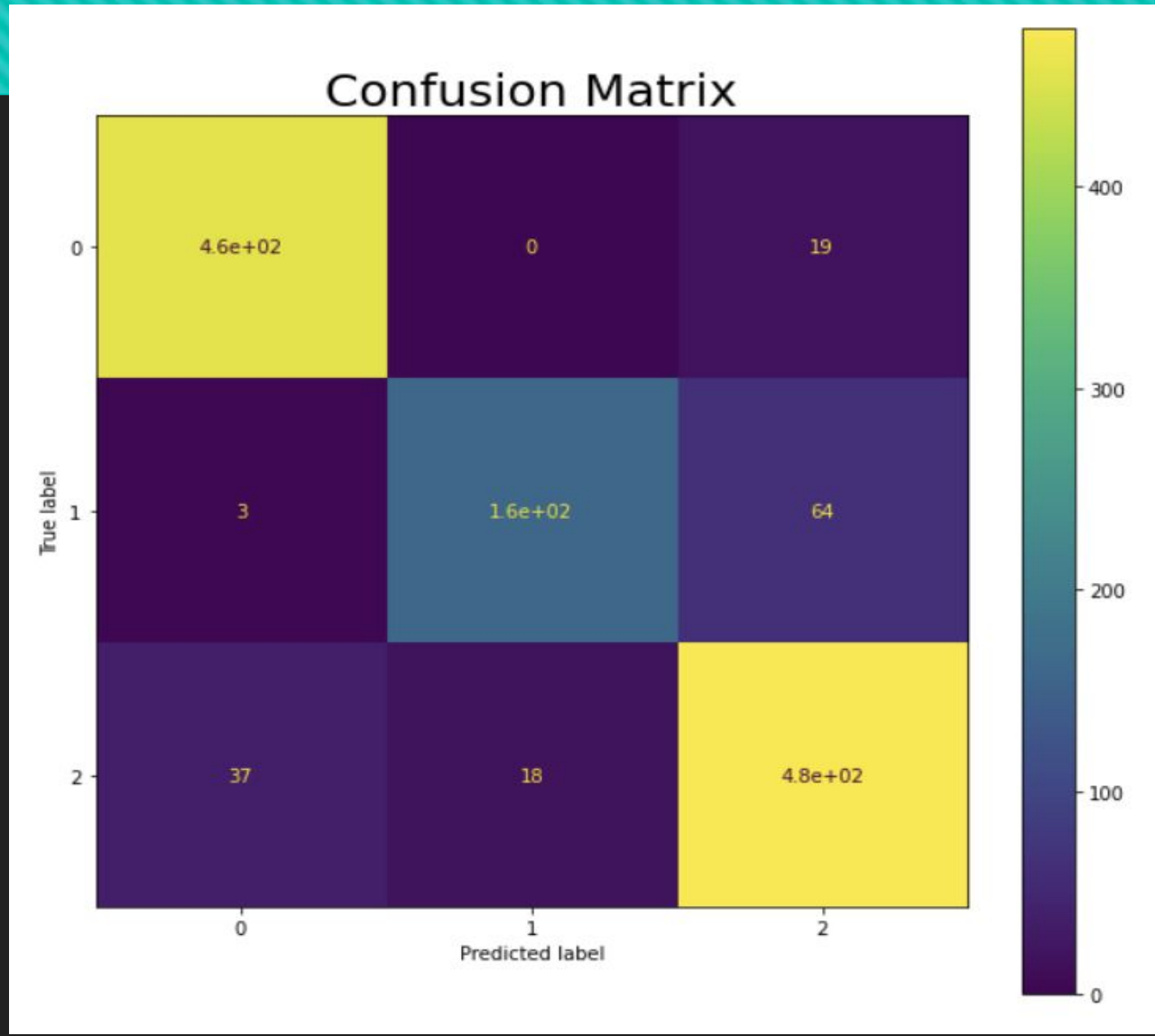
We used a neural network with different parameters of the number of hidden layers and the size. After tuning the parameters we find that the best option is 43,43,43 NN, i.e. 3 hidden layers of size 43 each. We got an accuracy of 81.82 %. Confusion Matrix is shown here:



Model Evaluation

According to the accuracy, we can see that almost all of the models get about the same score.

The best model is the SVM classifier with an accuracy of 87.06%. Therefore,we used this model to predict positions of player on FIFA 20 datasets. We found an accuracy of 88.62 % which is very good and close to our training accuracy. Below is the confusion matrix on this test data (FIFA20)



Practical Significance

If you are an avid football fan, you might be aware of how a coach, The formation and The playing style can transform a player's career. There are countless examples of generational talents who never performed and Seemingly average players who turned into footballing icons just based on whether they had a coach who played them in a right position and systems. The proposed model with almost 85 % accuracy, gives an excellent way to judge a player's utility instead of using Trial and error or depending on the Coach's calibre.

Future Work

Keeping in mind how football clubs and virtuals games are growing, we can see a lot of different quality work that can be arranged out of this type of analysis and predictive models:-

- Probing whether the current year's video game data can provide insight into the future rating of players.
- Historical comparison between Messi and Ronaldo or between other stars. (what skill attributes changed the most during the time - compared to real-life stats).
- Sample analysis of top n% players (e.g. top 5% of the player) to see if some important attributes as Agility or BallControl or Strength have been popular or not across the FIFA versions. An example would be seeing that the top 5% of players of FIFA 20 are faster (higher Acceleration and Agility or pace) compared to FIFA 15. The trend of attributes is also an important indication of how some attributes are necessary for players to win games (a version with more top 5% players with high BallControl stats would indicate that the game is more focused on the technique rather than the physical aspect).

Conclusion

Overall, all models did a very good job in predicting positions. SVM classifier beats all of them by a little margin, but we can use any of the above model to predict players positions. We used python and scikit-learn, starting with just one feature (value) and then adding some new features for training the models (age and finishing features). We noticed that by adding new features to the model, we would not always have better results and we mentioned common approaches to address this problem like feature selection, extraction and dimensionality reduction.

Contributions

Tasks	Avish	Rishika	Kaushikraj	Neeraj
Data Preprocessing	S	R	I	A
Model selection	R	I	S	C
Model Implementation	R	S,A	A,C	S,I
Testing	C	S	I	R
Model evaluation	I	C	R	S
PPT preparation	S	A	R	C

R:Responsible for maximum effort, S: Supported with good effort, C: Consulted for extra points, A: Approved after review, I: Informed about the completion of step

Thank You!