**FINAL PROJECT REPORT**

**PROBLEM STATEMENT:**

The **FIFA** (Federation International de Football Association) is an organization which describes itself as an international governing body of association football. FIFA is responsible for the organization of football's major international tournaments, notably the World Cup which commenced in 1930 and the Women's World Cup which commenced in 1991.

In this real-world problem, we are predicting various important attributes such as physicality and defense of the FIFA players 2018 and thus estimating the team's overall potential and rank which can be useful to analyze the strongest team. The predictions are done here by building many linear regression models from which one model is selected to predict the expected values. Different classification techniques are used to classify the players based on their performances and characteristics of the players.

**DATA COLLECTION:**

To perform the above tasks, FIFA18 complete player dataset is taken. This dataset contains the following information.

* 180 Fields for every player in FIFA 18.
* Player infomation such as DOB, club, league, nationality, salary and physical attributes
* All playing attributes, such as finishing, crossing, heading\_accuracy and dribbling
* Special attributes like skill moves and international reputation
* Traits and specialties
* Overall, potential, and ratings for each position.

**DATA PREPARATIONS:**

The club names of few players in the data set are null. These missing values are handled by filling the value with unknown club. And these players are excluded from some of the computations which is based on club.

Features such as right defensive mid, center attacking mid, left attacking mid is null for the players who are goal keepers. These attributes are filled with zeros. So that while classifying the work rate attack and defense of each player, they won’t raise exceptions.

**EXPLORATORY DATA ANALYSIS:**

Exploratory data analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often with visual methods. In EDA phase, we have performed the following methods to explore the data set that is selected.

1. Providing statistical summary for some of the important features in the data set.

* The players are grouped according to their clubs and based on the overall scores of players in each group, the ranking of the group is determined to identify the strongest team in FIFA which can have the potential to win the next series of league.



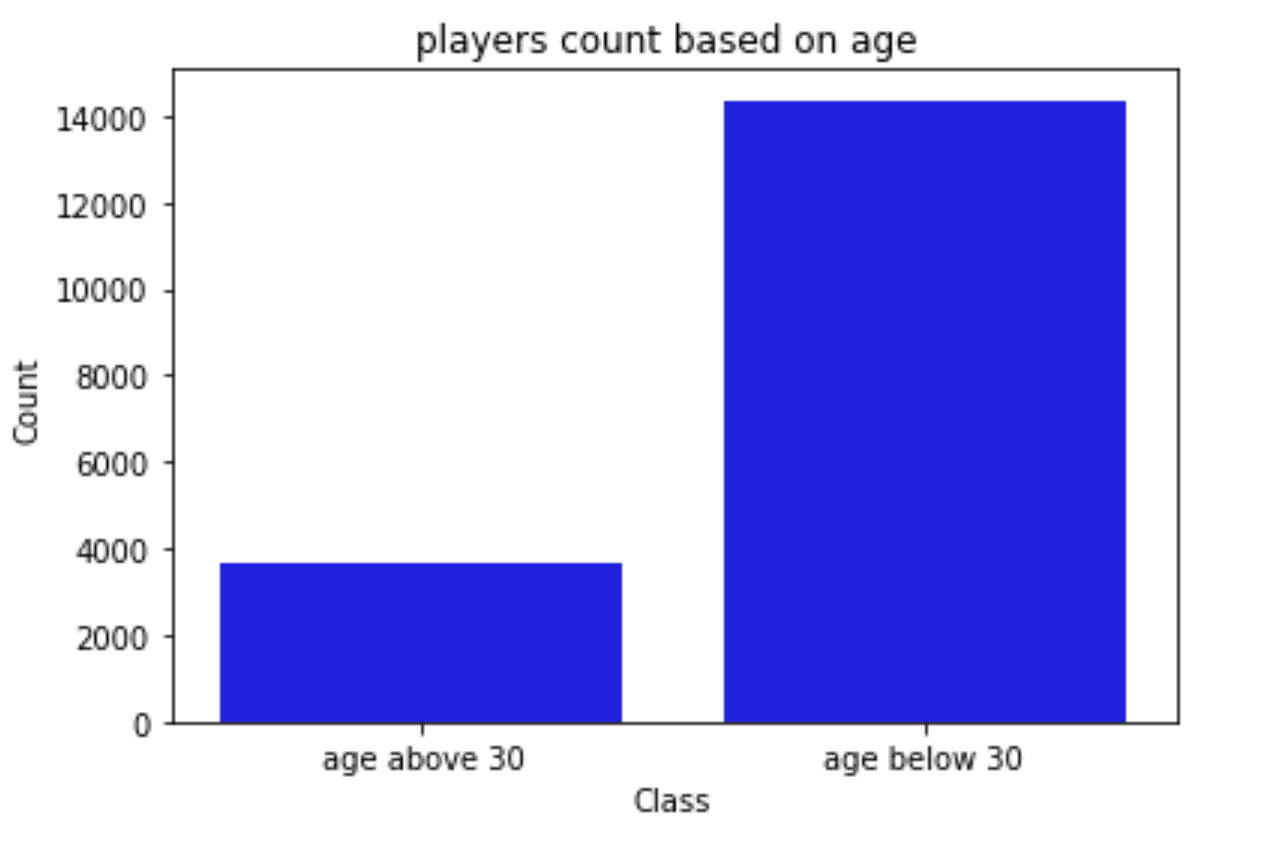
* Average salary of the players in each club is calculated and the list of highest paying clubs can be obtained from the resulting analysis.

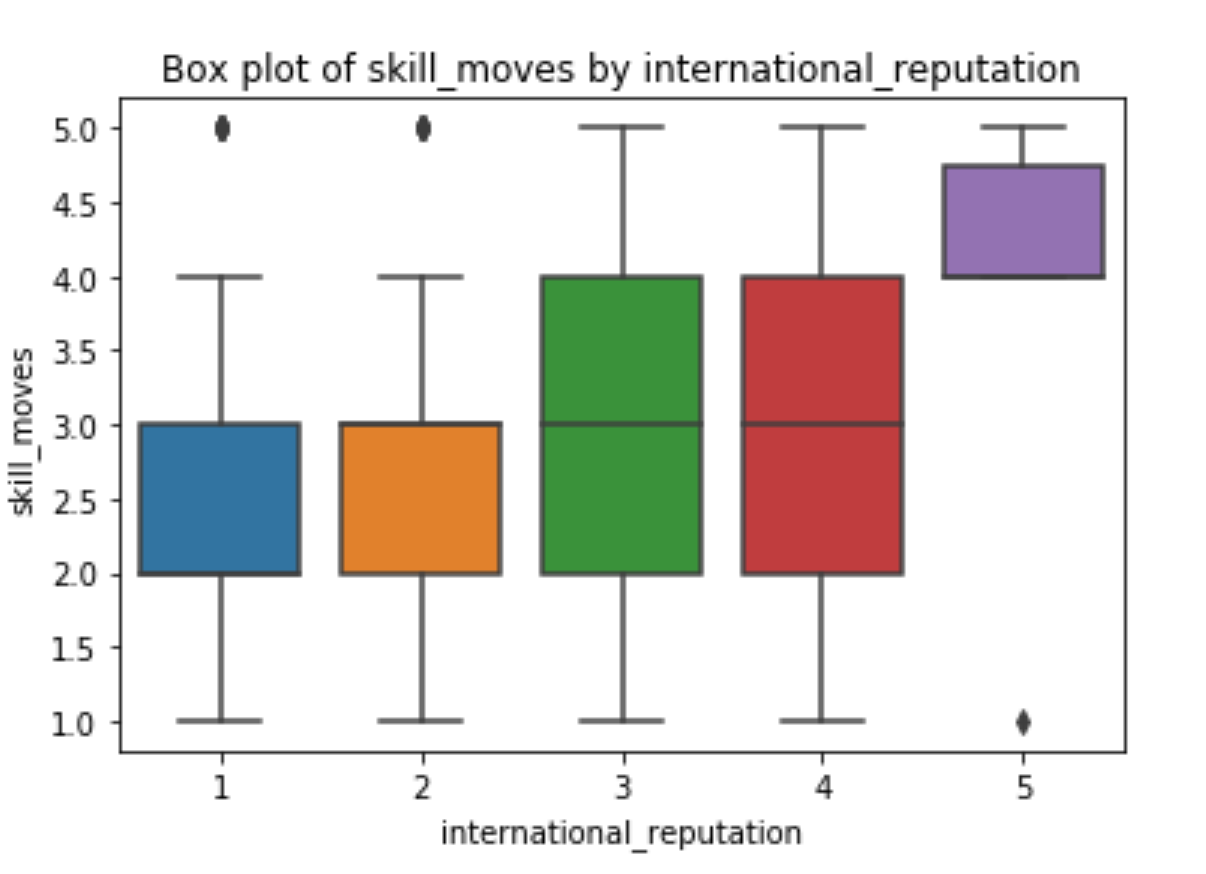
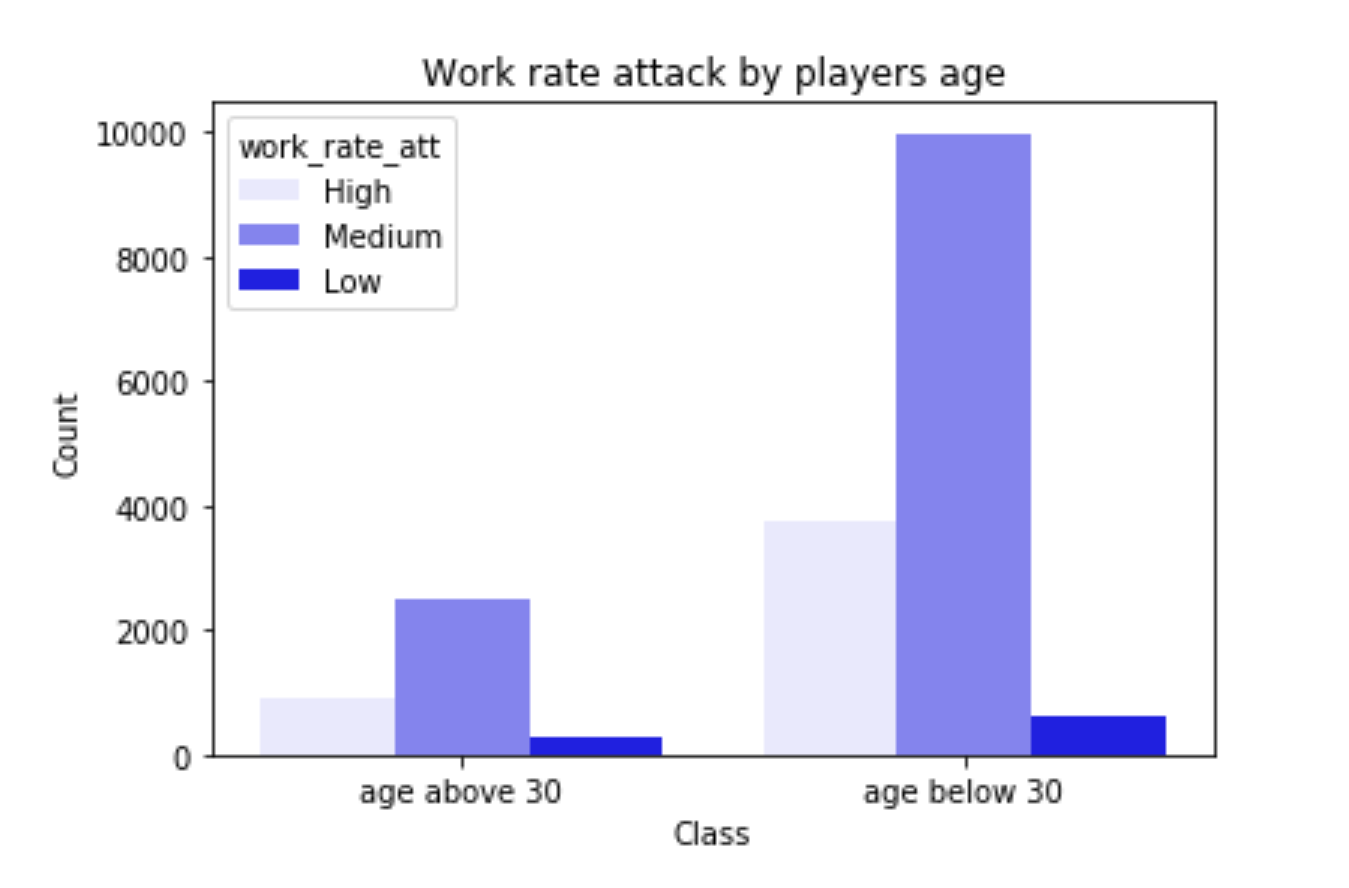


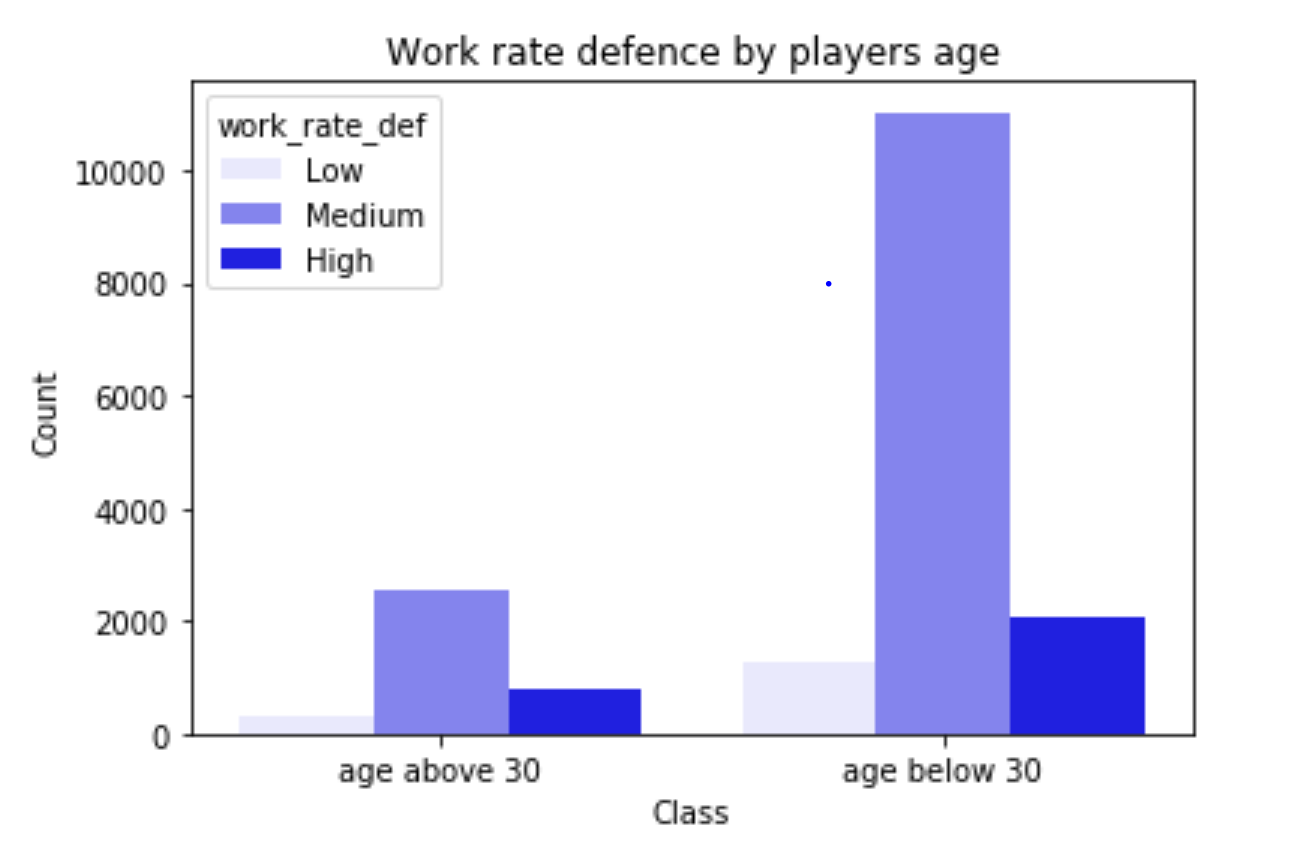
* The count of the players based on their nationality is also listed.

1. Obtained relationship between any two features and plotted them in a graph to get better visualization of the dataset.

* Various other aspect of the players like Age is estimated from the given date of birth and it is plotted in graph with respect to their skill moves, international reputations and the number of players in that age category.



1. Understanding the interactions between multiple fields in the data set to make assumptions and predictions on the predictor variables that can help us to find the main target or to classify and cluster the data based on the selected variables.

* The data set consist of 185 features overall. However, not all the attributes would be required for regression and classification techniques that are used here.
* Initially all the relevant features that are required for making the predictions or classification is picked based on the knowledge we possess. Much more relevant features are selected by calculating co relation coefficient between the predictors and the response variables.
* With the attributes that are selected, the data is partitioned into training, validation and test data sets with which regression and classification techniques are applied.

**DATA MODELLING:**

**1) REGRESSION:**

The target variables are predicted using Regression. After analyzing the dataset, relevant attributes are selected and using those attributes the data is partitioned into training, testing and validation using hold-out method.

Different regression models are applied on different combinations of predictor variables on the training dataset and test dataset. The scores of these keeps fluctuating when different combinations of the predictor variables are used. In order to obtain high accuracy, one such technique used is calculating Correlation coefficient between the predictors and responses.

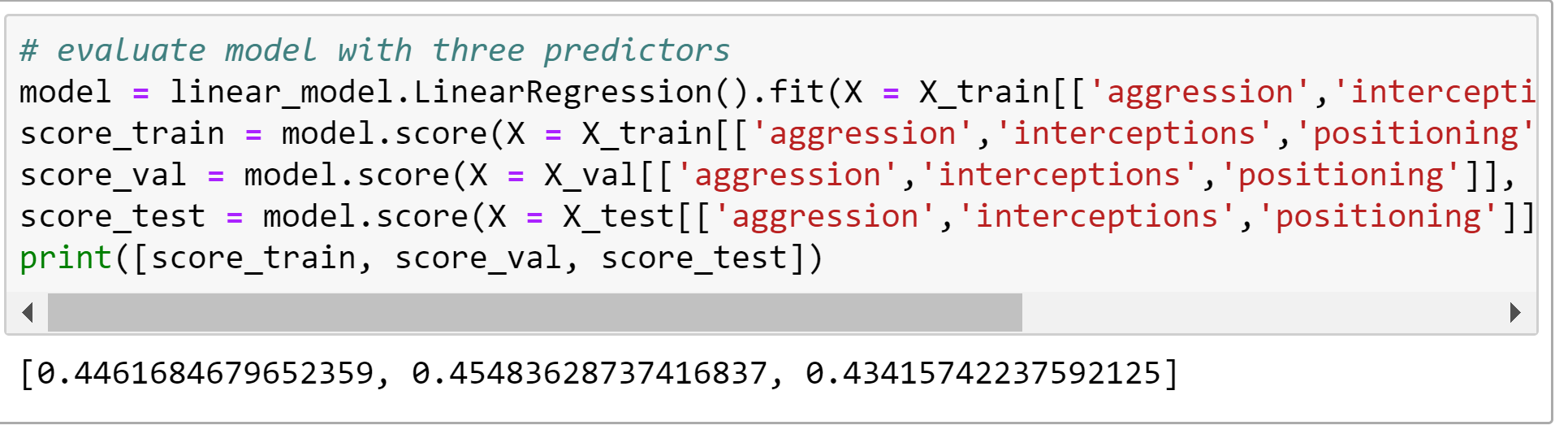
Correlation coefficient between target variable and relevant attributes is calculated to figure out the features that possess better relationship with the target variable than others. The predictors variables that has Correlation coefficient closer to one can predict the target variable with high accuracy.

Based on these results, different models are applied on the predictors and the one which gives the maximum accuracy is used for predicting the values of the target.

**Physicality :**

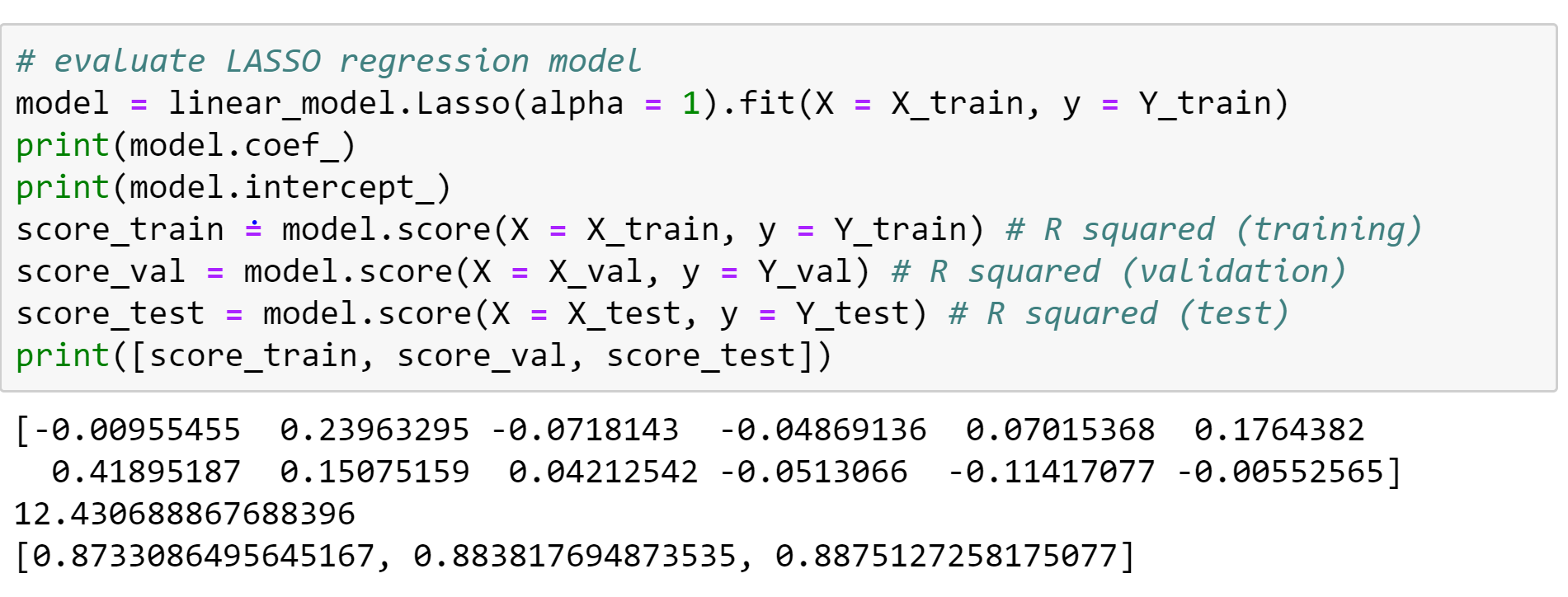
One of the Target variable predicted is ‘physicality’ of the player.

When the regression model is build using this combination of predictors (aggression, interceptions and positioning), the accuracy of the model is less than 50%.



When the regression model is build using the combination jumping, stamina, strength, aggression, agility which gave higher correlation coefficient with the target variable, the accuracy level got increased to 80%.

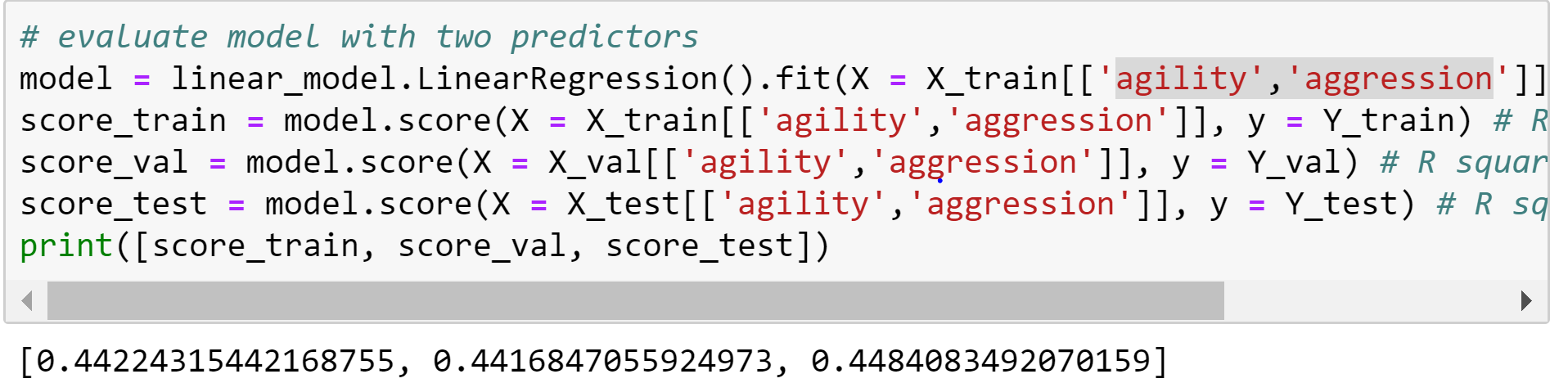
Similarly, when all the predictors are used to build a LASSO Regression model, the accuracy thus obtained is close to 85%.



**Defense :**

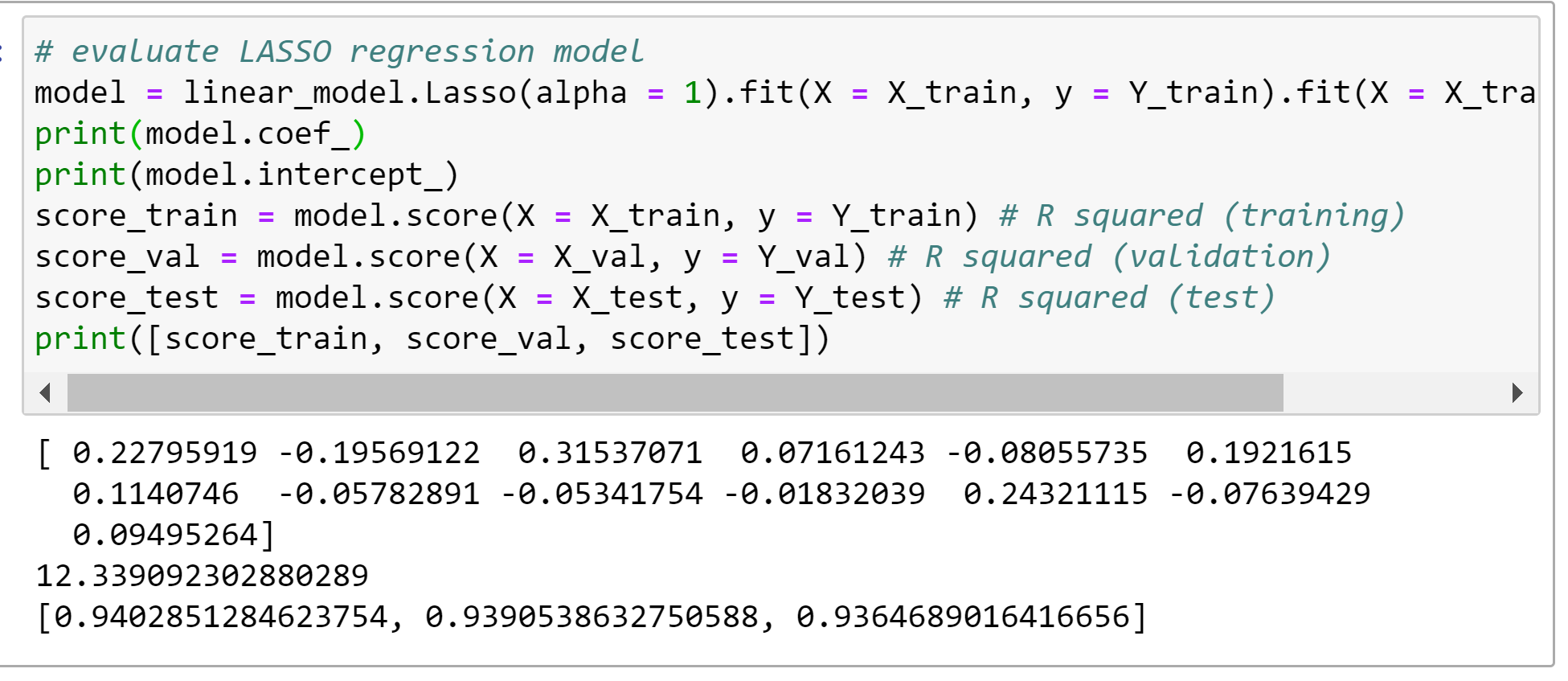
The other target variable that is predicted is Defense skill of the player.

When the regression model is build using this combination of predictors (agility ,aggression) the accuracy of the model is less than 50%.



When the regression model is build using the combinations aggression, standing\_tackle , sliding\_tackle, marking, heading\_accuracy which gave higher correlation coefficient with the target variable, the accuracy level got increased to 80%.

Similarly, when all the predictors are used to build a LASSO Regression model, the accuracy thus obtained is obtained to 92%.



Lasso **regression** method performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. Hence it gives good accuracy and thus it is used in predicting both the targets.

**2) CLASSIFICATION :**

The work rates of all the players are rated High, Medium or Low based on their ability and skills.

Here, we tried to classify the players based on work\_rate\_attack and work\_rate\_defence.

Only the most relevant features that is used for estimating the work rate attack and defense is taken into consideration for splitting the data set into train and test data set. This is done in order to prevent the results being biased and to reduce the dimensionality.

Then the portioned data is scaled and standardized to make sure that all the attributes will have a mean of zero and standard deviation of one and then smoothened to remove the imbalance using SMOTE which generates new observations by interpolating two minority class observations.

Different classification techniques such as Decision tree, Random forest , Logistic regression and K nearest neighbors are applied to classify the data and accuracy is calculated for every model.

**Work rate attack :**

The attributes used for splitting are :

pac, sho, pas, dri, phy , agility, reactions ,balance, shot\_power, jumping, stamina, strength, aggression, positioning,lw, ls, cf, st, rf, rw, cam, lm , cm , rm , ram ,rcm , lam, vision, penalties, composure, crossing, finishing, heading\_accuracy, short\_passing, volleys, dribbling, curve, free\_kick\_accuracy, long\_passing, ball\_control, acceleration, sprint\_speed.

* **DECISION TREE** classifier:

Setting the maximum depth to 50 and random state to 1, the evaluation metrics thus obtained is

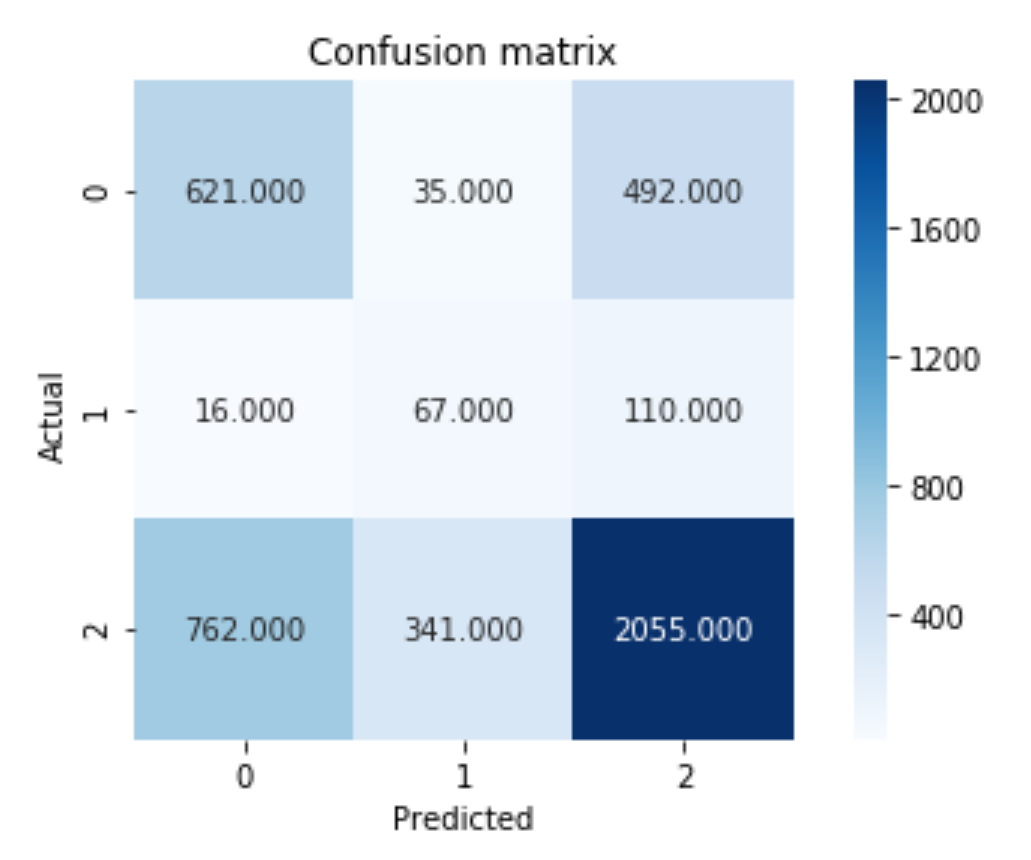
Accuracy : 0.6096910424538786

Error: 0.3903089575461214

precision : [0.44388849 0.15124153 0.77342868]

recall : [0.54094077 0.34715026 0.65072831]

F1 score :[0.48763251 0.21069182 0.70679278]



* Building a **RANDOM FOREST CLASSIFIER** with n\_estimators = 10, criterion = "entropy", random\_state = 0, max\_depth= 50 gave the following results.

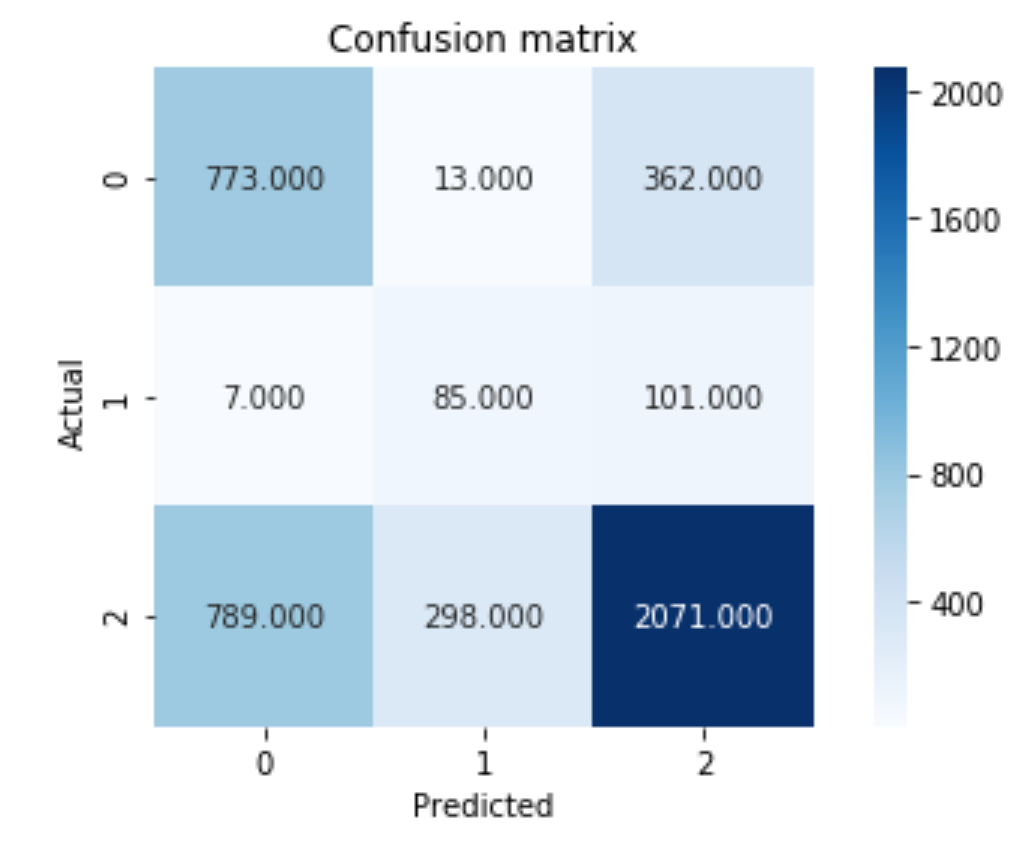
Accuracy : 0.6510335630140031

Error: 0.34896643698599694

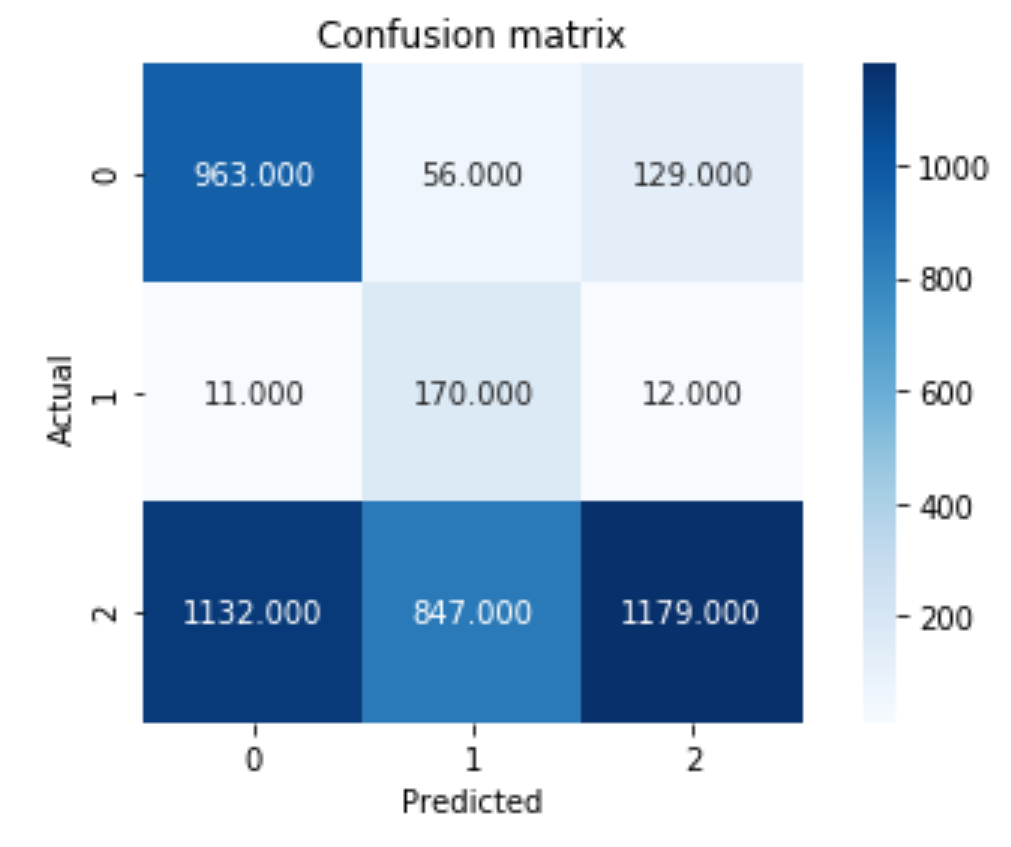
precision : [0.49267049 0.21464646 0.81728493]

recall : [0.67334495 0.44041451 0.65579481]

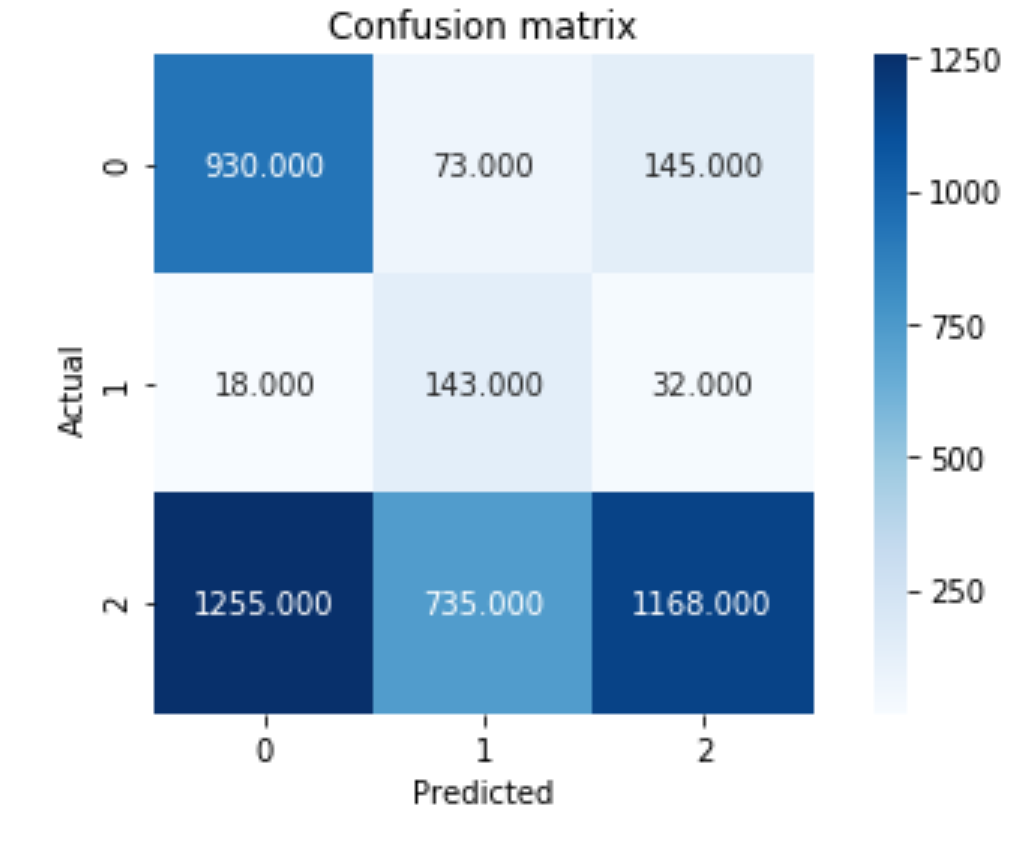
F1 score : [0.56900994 0.28862479 0.72768798]



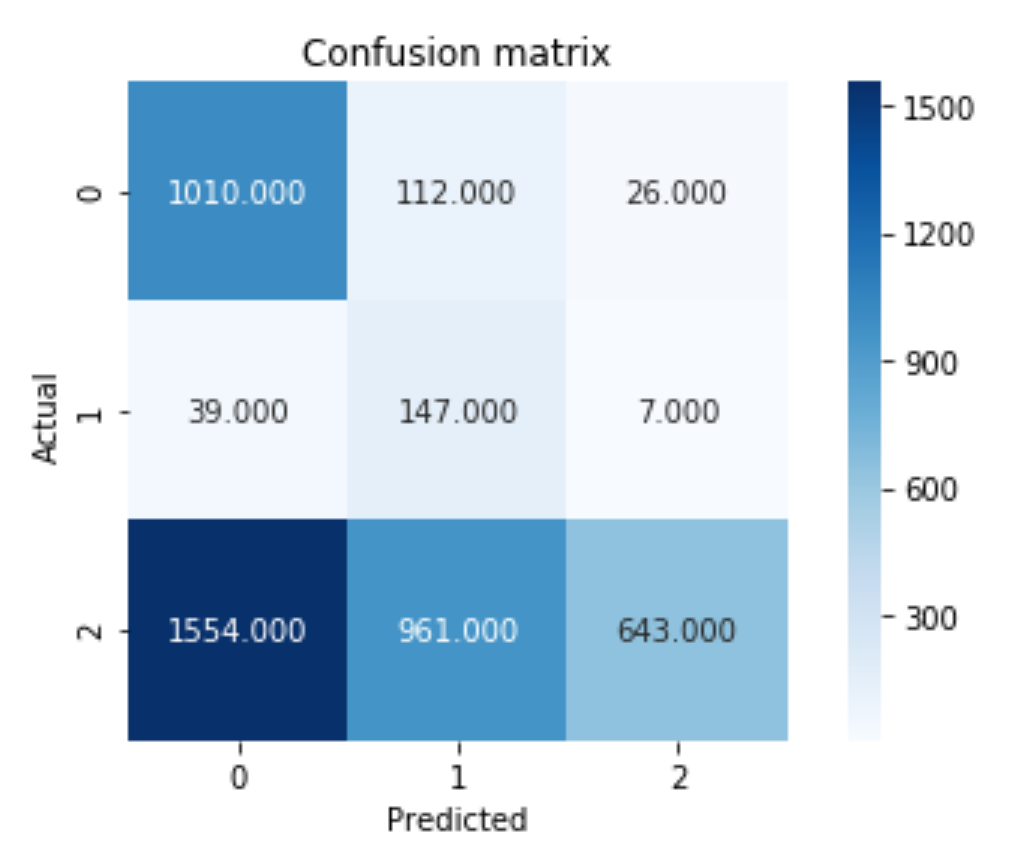
* The **LOGISTIC REGRESSION** gave an imbalance classification with the accuracy 0.51389



* The **K NEAREST NEIGHBOR** with 6 neighbors when applied gave an accuracy of 50%.



* The **NAIVE BAYES** classifier gave an accuracy of 0.40008890864636587



**RANDOM FOREST CLASSIFIER** is the best performing classification model as its accuracy of classification is higher when compared to the other classification models. Random forest aims to improve classification accuracy by reducing the variance of the decision trees, which are assumed to have high variance and low bias.

**Work rate defense:**

The attributes used for splitting are :

pac, pas, def, phy , agility, reactions, balance, shot\_power, jumping, stamina, strength, aggression, interceptions, positioning, rs, rw, rf, rcm, rm, rdm, rcb, rb, rwb, st, lw, cf, cm, lm, cdm, cb, lb, lwb, ls, lf, lcm, ldm, lcb, interceptions, positioning, vision, composure, marking, standing\_tackle, sliding\_tackle, crossing, finishing, heading\_accuracy, short\_passing, volleys, dribbling, long\_passing, ball\_control, acceleration, sprint\_speed.

* **DECISION TREE** classifier:

Setting the maximum depth to 50 and random state to 1, the evaluation metrics thus obtained is

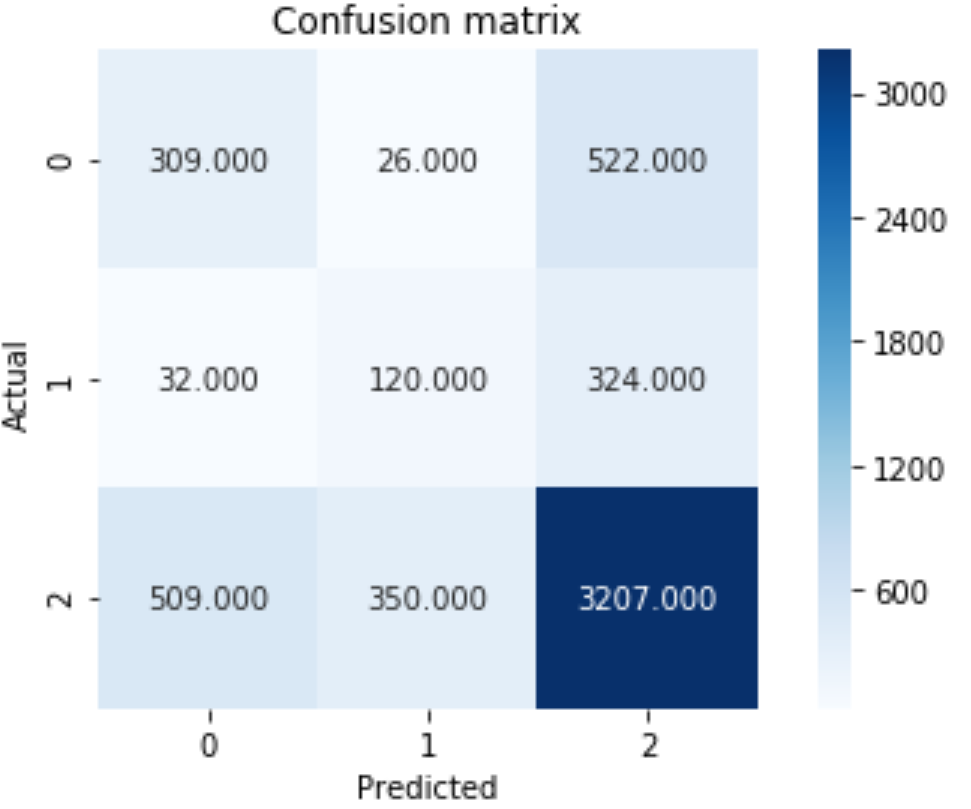
Accuracy: 0.6734580477866271

Error: 0.3265419522133729

Precision: [0.36352941 0.24193548 0.79126573]

Recall: [0.36056009 0.25210084 0.78873586]

F1 score: [0.36203866 0.24691358 0.78999877]



* Building a **RANDOM FOREST CLASSIFIER** with n\_estimators = 2, criterion = "entropy", random\_state = 0, max\_depth= 50 gave the following results.

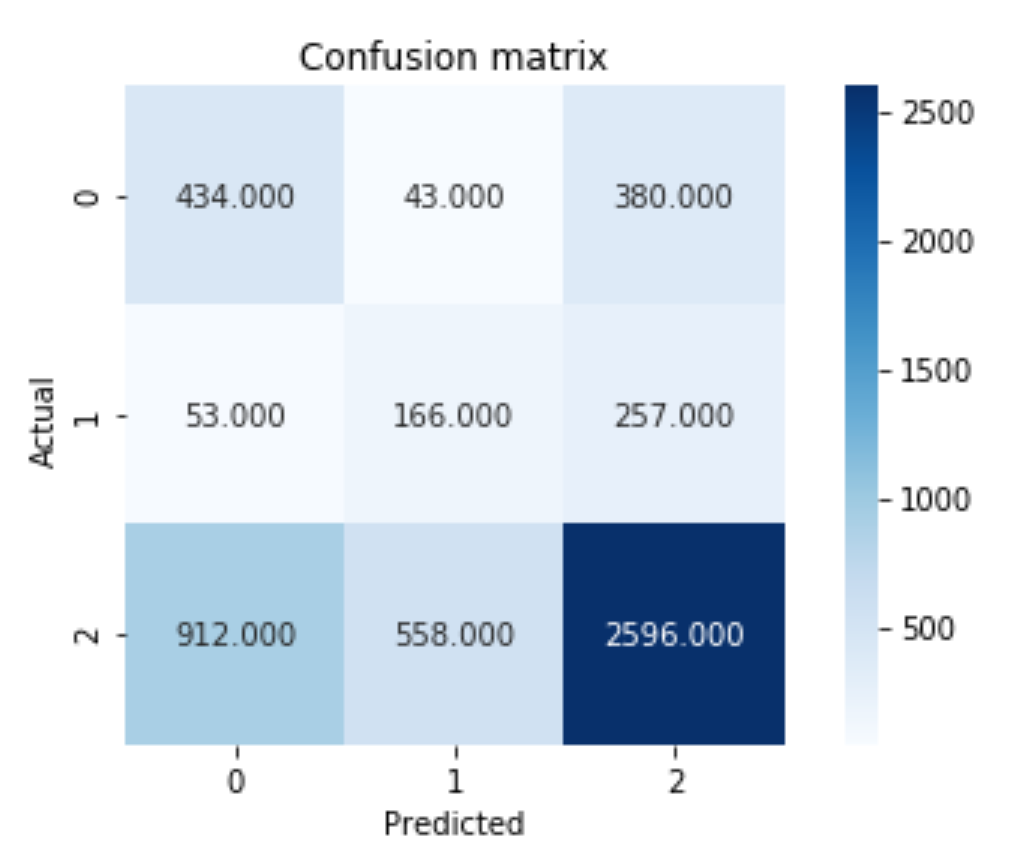
Accuracy : 0.5919614743471013

Error: 0.40803852565289866

precision : [0.31022159 0.21642764 0.80296938]

recall : [0.50641774 0.3487395 0.63846532]

F1 score : [[0.38475177 0.26709574 0.71133032]



* The **K NEAREST NEIGHBOR** with 4 neighbors when applied gave an accuracy of 70%.

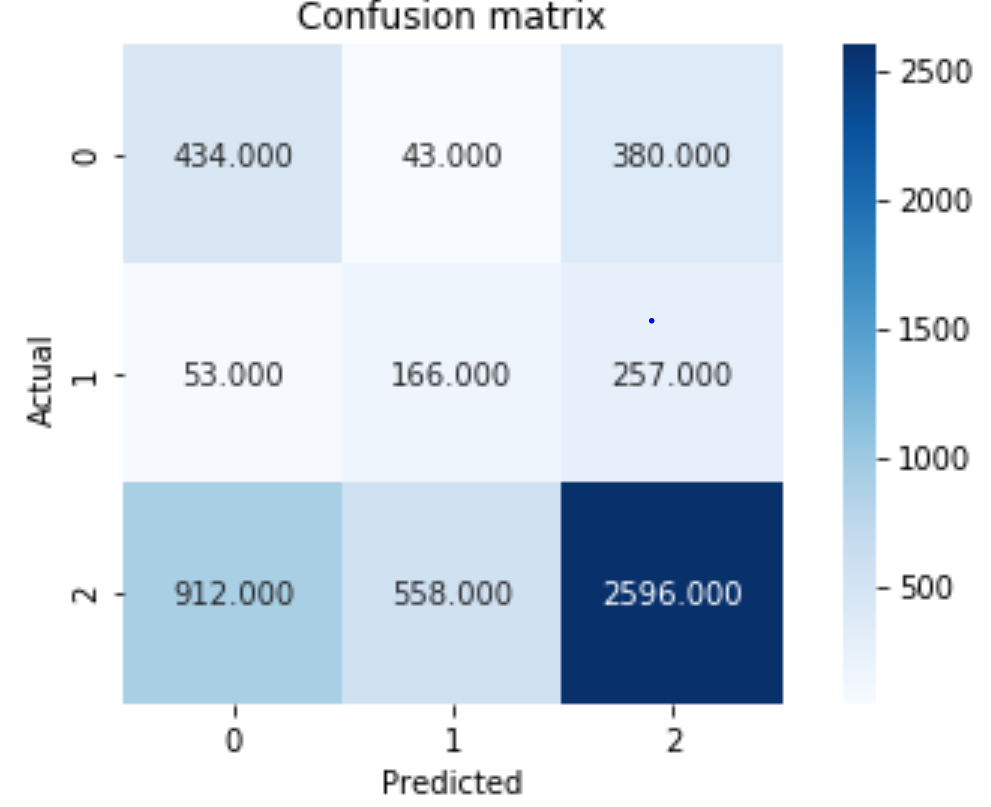
Accuracy : 0.6914243378403409

Error: 0.30857566215965915

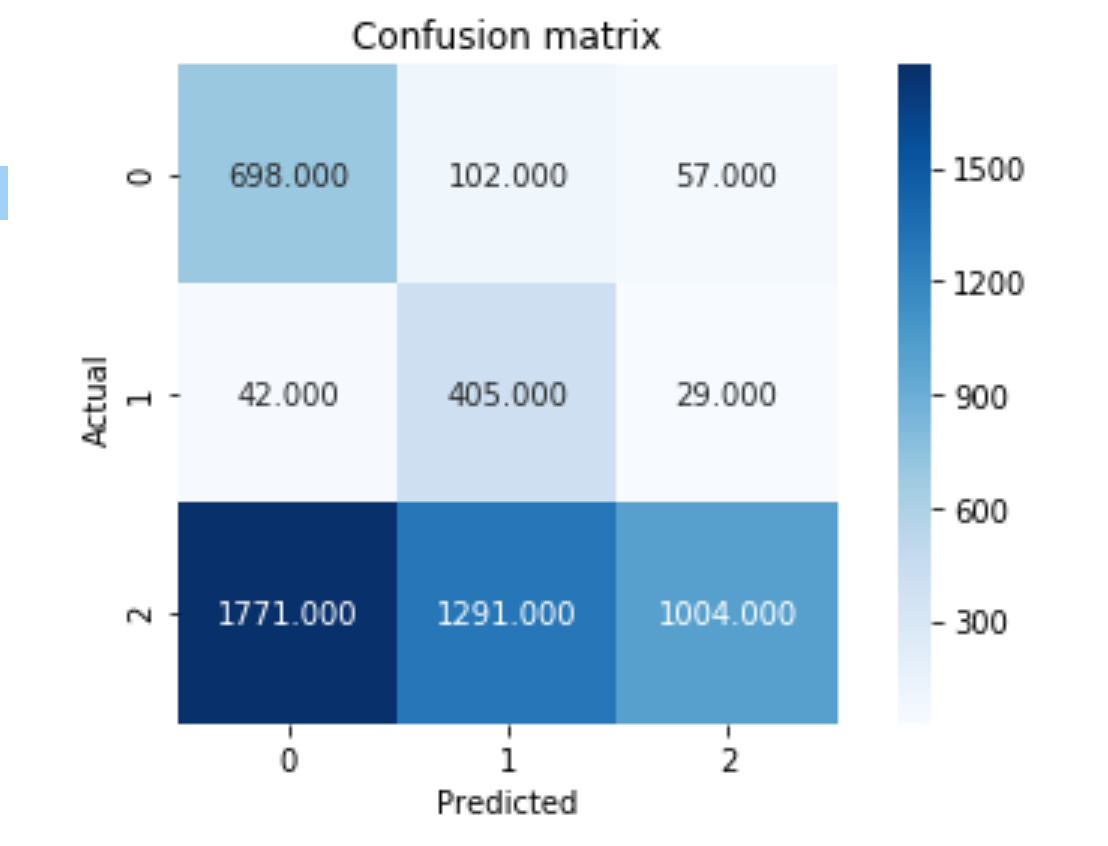
precision : [0.40581542 0.26236559 0.79411055]

recall : [0.37456243 0.25630252 0.80914904]

F1 score : [0.38956311 0.25929862 0.80155926]



* The **NAIVE BAYES** classifier gave an accuracy of 0.39025



**K NEAREST NEIGHBOR** with 4 neighbors is the best performing classification model as its accuracy of classification is higher when compared to the other classification models. 𝒌 -nearest neighbors classifier assigns class labels to observations based on the class labels of the 𝑘 “most similar” observations