

**IME672A**

Data mining and Knowledge Discovery

(PROF - FAIZ HAMID)

Final Report

Group members:

Amrendra Pratap Singh (170097)

Avani Shrivastava (170178)  
Harish Dechiraju (170286)

Himanshu Verma (170303)

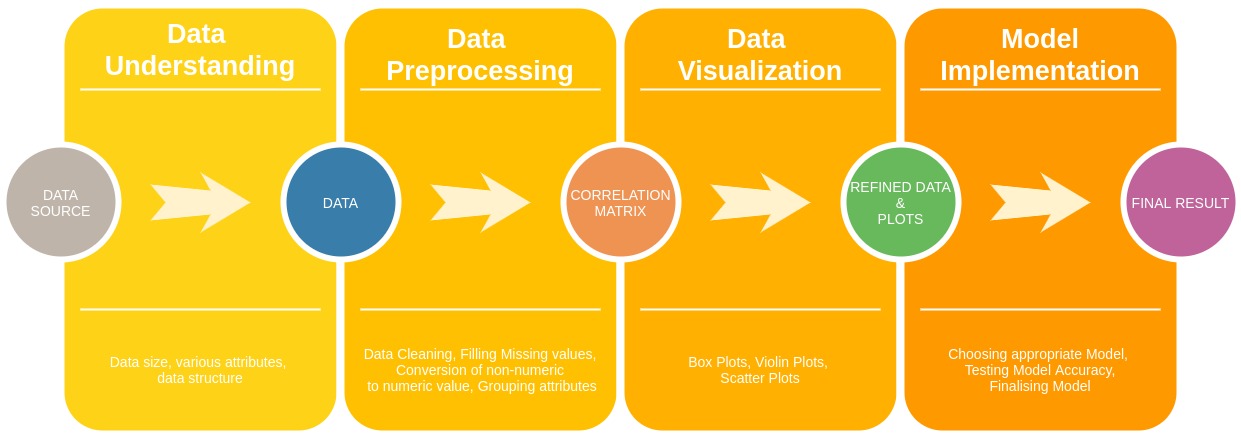
Inshu Namdev (170308)

Pulkit Pandey (170516)

**Problem Statement** - Predict Average Rating of FIFA 19 Player

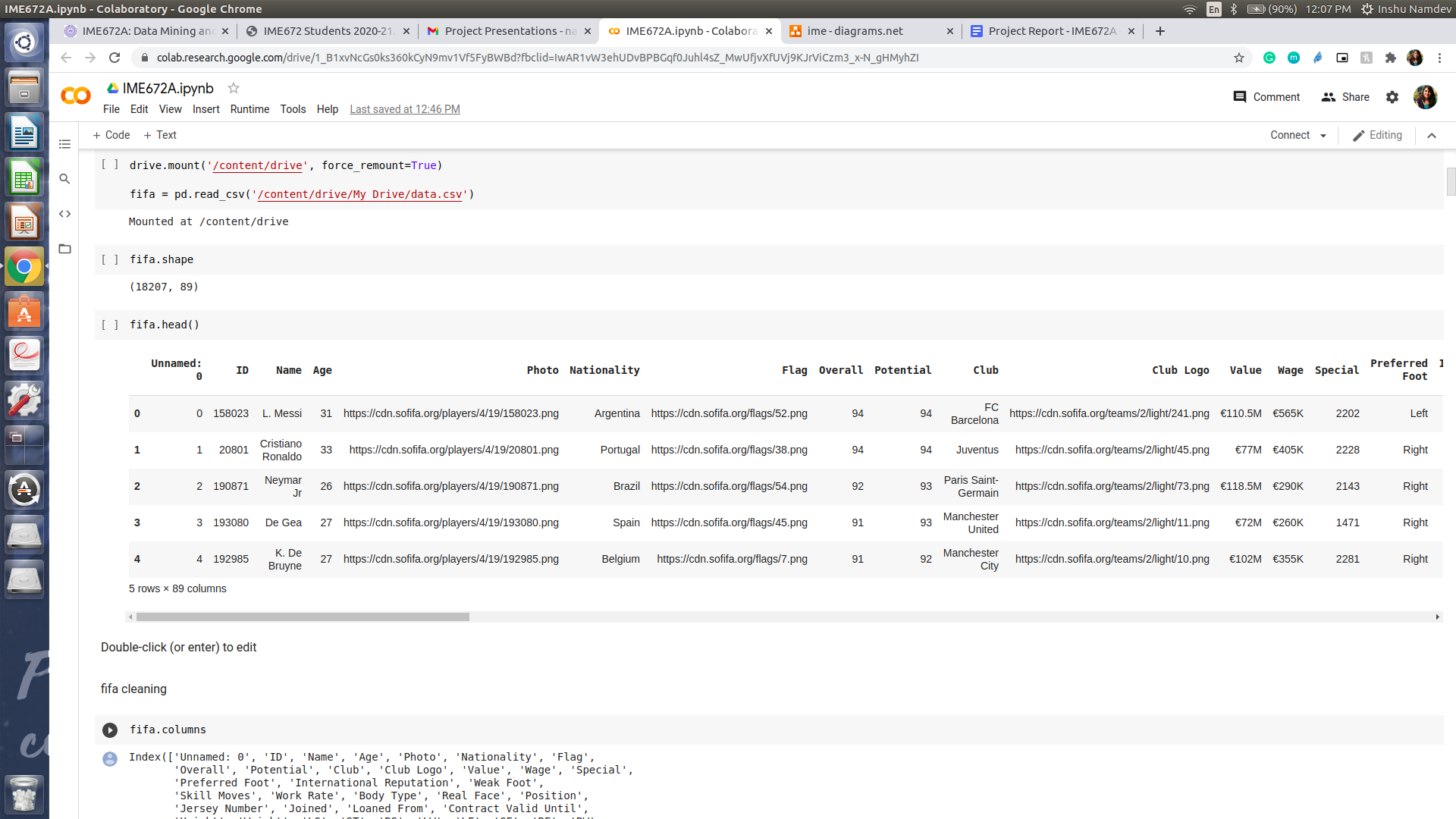
Data Source - <https://www.kaggle.com/karangadiya/fifa19>

**Project Structure -**

****

**Data Understanding**

The project problem was to predict the average rating of a FIFA’19 player, using data from <https://www.kaggle.com/karangadiya/fifa19>.

The first step was to read the dataset. For this step, the CSV file was imported from the given data source. There were around **18,207** rows and **89** columns. Original Structure of the dataset consisted of 89 attributes such as Name, age, photo, nationality, flag, etc. which can be seen in the image given below - 

**Data Preprocessing**

Out of the 89 attributes, the value of the **overall** attribute was to be predicted. It can be seen from the dataset that there should be a total 18,207 values for each attribute. However, there were many attributes with missing values. All the missing values needed to be filled which will clean the dataset and help in better understanding of the data. In the next step, the total number of the null values for each attribute was calculated. It can be observed from the data set that there were a few rows that had missing values for 48 rows and had ‘62’ as the overall rating. To avoid duplication, these rows were dropped, instead of trying to fill the missing values.

Among the non-numeric, attributes, namely, **Preferred foot**, **work rate** and **the player position** were retained to plot them against the other attributes to see if they contributed anything significant to the overall rating of the player.

Next, to find out the relation among the numeric variables, a correlation matrix was plotted and based on that, a number of attributes which had high correlation among themselves, were clubbed to form a single attribute, while those that were clubbed were dropped. The following attributes were clubbed by taking the mean of every attribute corresponding to one entry to form a new single attribute, and the correlation matrix with new attributes were plotted

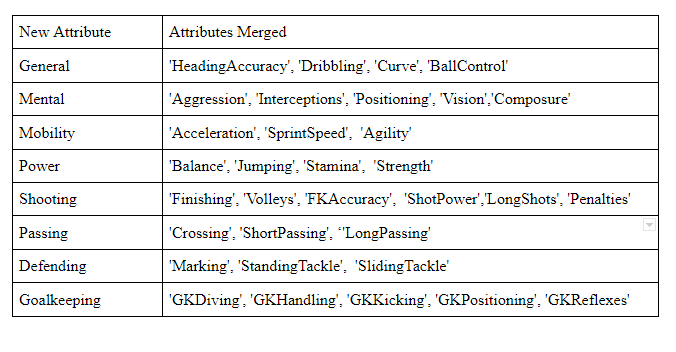
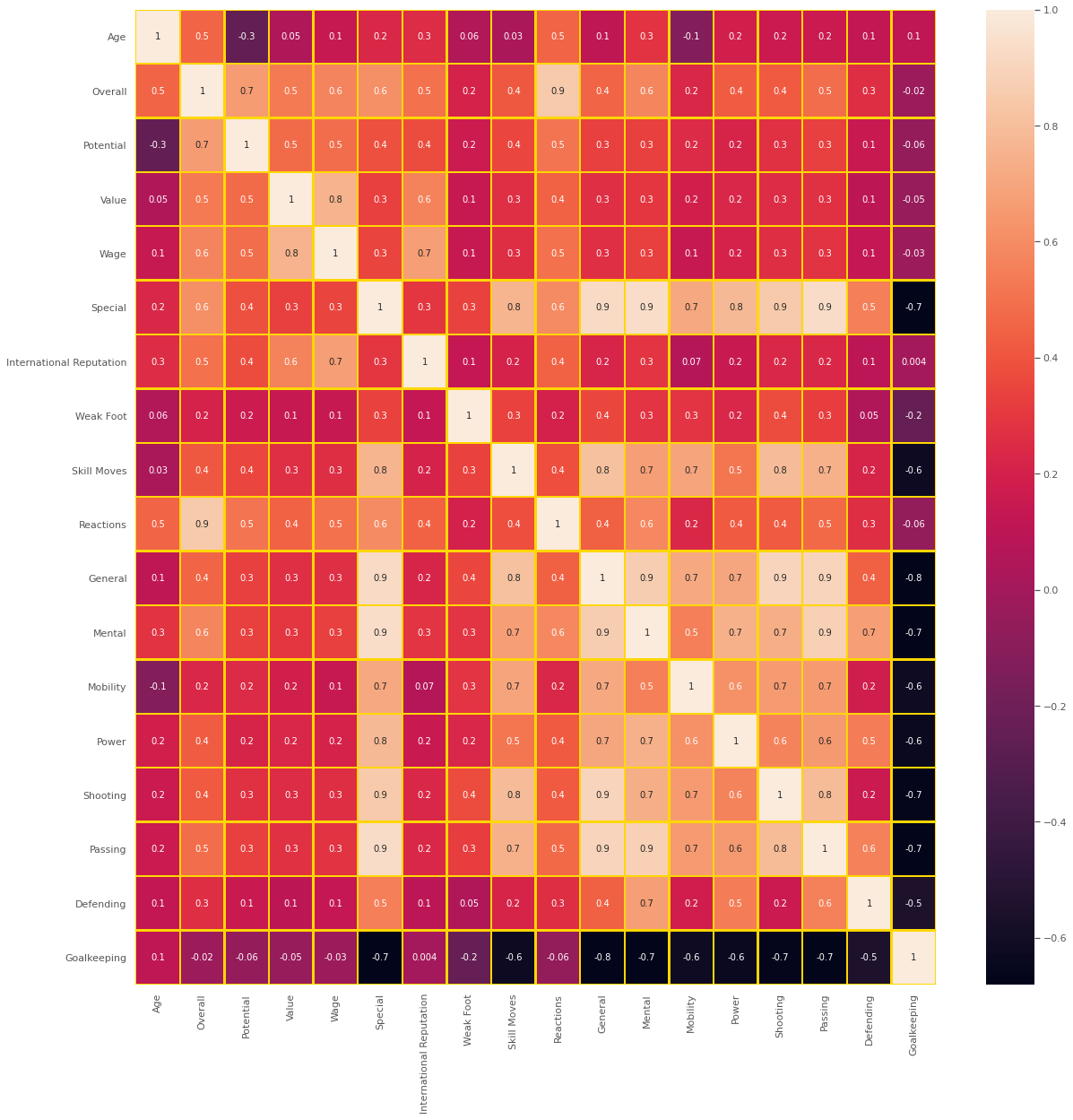


Fig 2: Table showing the new attributes formed by merging attributes with high correlation between themselves

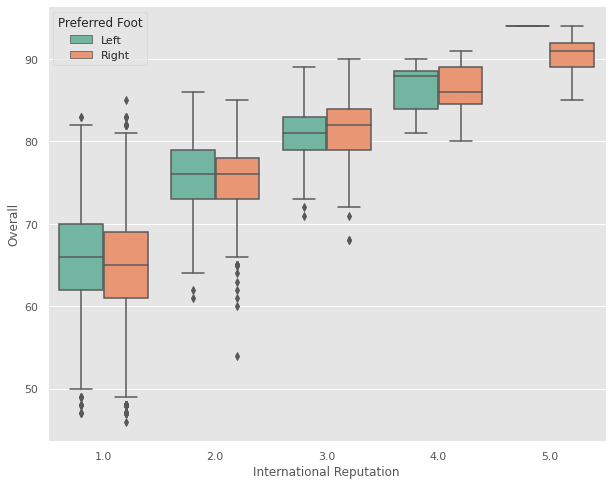
**Data Visualisation**

To begin visualisation on our dataset, we first started with heatmap correlation matrix visualisation. It helped us to quickly grasp and see the impact of different variables on each other.

We can visualise that some of these variables are strongly correlated with each other but the variable ‘goalkeeping’ is showing poor correlation with most of the variables. But we can’t ignore the ‘goalkeeping’ variable in further processes because of the obvious reasons. Goalkeeping ratings are strongly influenced by the goalkeeper skills.

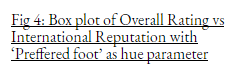
A goalkeeper may perform poorly in other player’s positions like ‘passing’, ‘shooting’ and so on or vice versa. So we finally decided to keep the ‘goalkeeping’ variable and perform further visualisation on the rest of the variables.

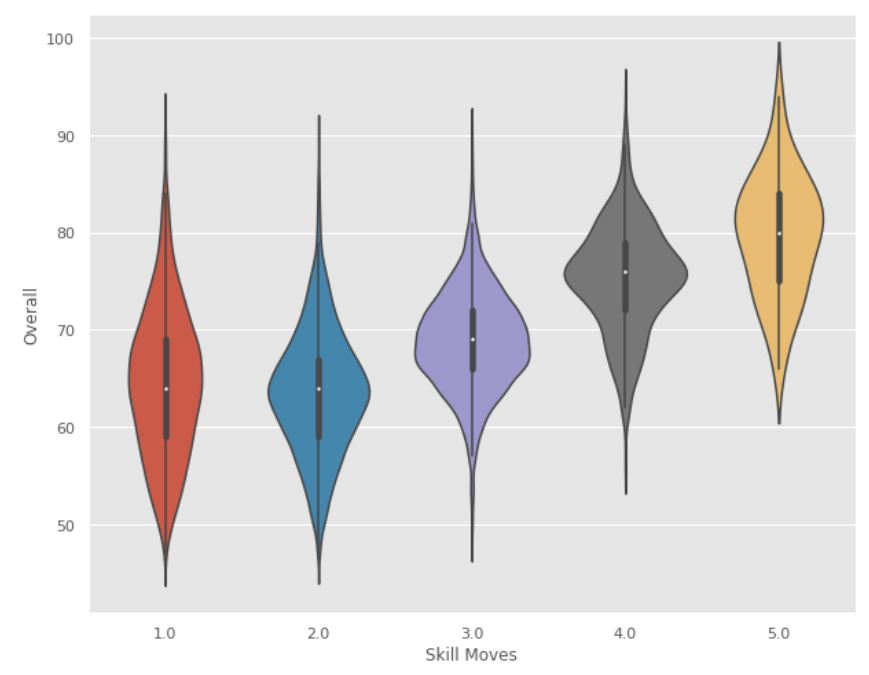
Fig 3: Heatmap of Correlation Matrix



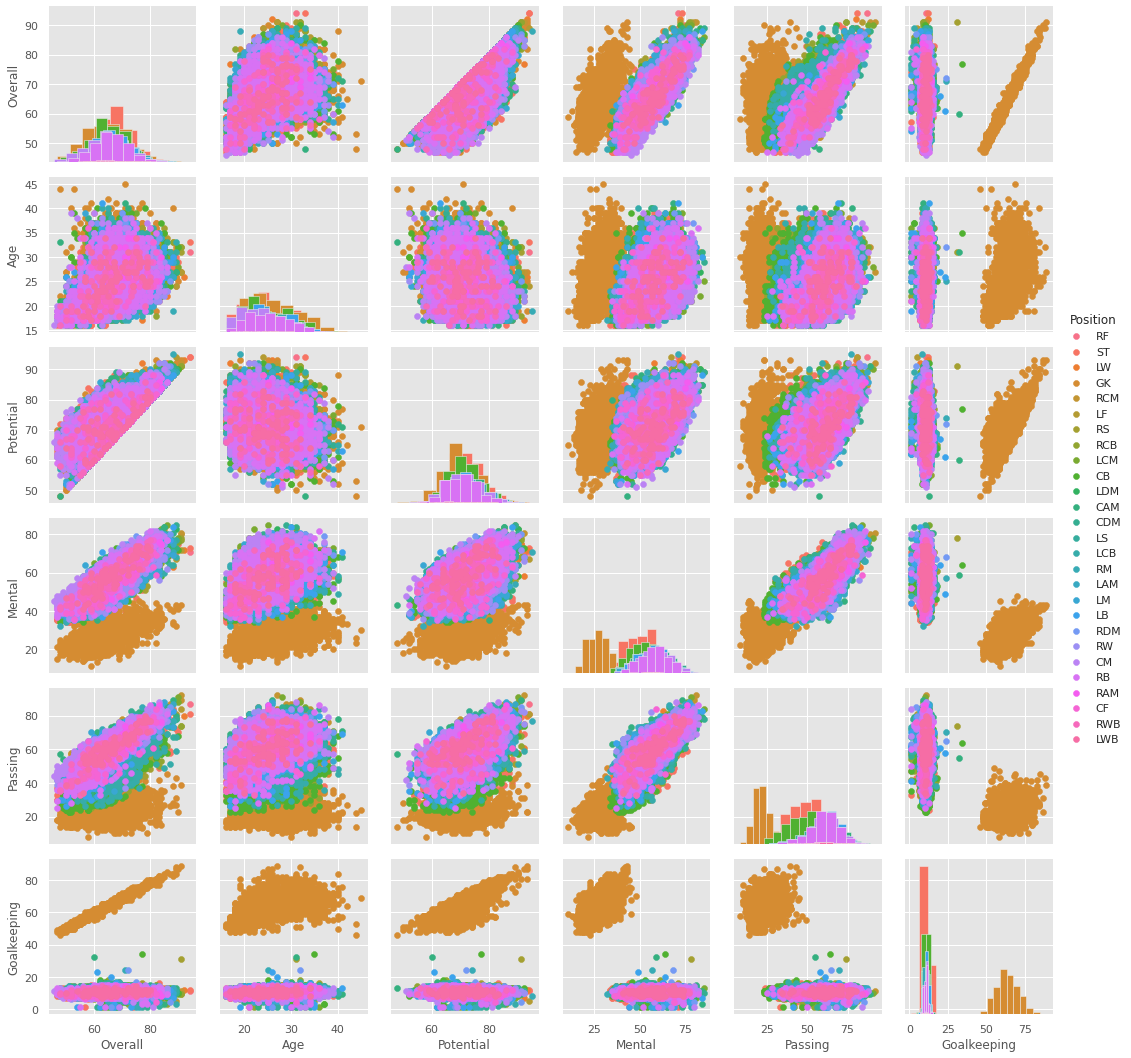
Next, to find the variance of the overall rating vs the non-numeric values we retained, we plotted a box plot for the overall rating vs the international reputation, distinguishing them by their “Preferred Foot”.

As an observation, we can see that an increase in the international reputation of a player also increases his Overall ratings. Almost linear increase in Q1, Q2, Q3, Max and Min can be seen for each and every boxplot.

The skewness in the data for the left and right foot (14k vs 4k) caused a discrepancy as the rating increases, but as far as the variance with the “Preferred foot” is concerned, the plot shows there isn't much variation on the overall rating if we try to distinguish it on the preferred foot. Therefore we can drop this attribute as well.



The adjacent violin plot depicts the overall rating vs skill moves. It is apparent that players with better skill moves have more rating on an average than players with less skill moves. The mean value of the distribution increases as the skill moves parameter of the players’ increases.



A pair plot with histogram diagonal was plotted for the main 6 variables of the data, namely - Overall, Age, Potential, Mental, Passing and Goalkeeping. The categorical variable ‘Position’ was used as the hue, or differentiating factor to get more detailed visualisation of the plot.

Observations -

* The “Goalkeeper” Position(marked in brownish-yellow), showed a distinct separation from the rest positions especially in the ‘Goalkeeping’ attribute.
* This shows that the goalkeeper position attribute can be separated from the rest of the other positions, and gives us a base to frame our model on.
* The rest of the positions were overlapping for almost all positions.
* The goalkeeper position can be separated by taking the goalkeeping attribute value, with goalkeeping value> 40 is a goalkeeper, whereas any value less than 40 is the rest other positions.
* So we can also conclude from this that we can drop the position attribute
* Also it shows that we may have to a separate model for goalkeepers and the rest other positions



**Model Implementation**

To predict the overall rating of the players, we implemented a Neural Network model based on Keras with Adam Optimizer.

The model has one input and one output layer and 4 hidden layers, and the data was split into 30% for testing purposes.

The model also had a validation set which was taken at 10% the training data set and the Mean Standard Error(MSE) was set as the loss, with metrics R2\_coefficient, Mean Absolute Error(MAE) and MSE. The model was trained for 10000 epochs.

Heuristic search was used to find optimal value of parameters, namely, learning rate, number of hidden layers, number of nodes per layer and activation function for each layer, and the model with best performance was chosen.

The model was applied on 3 different data

* **The Original Preprocessed Data**

|  |  |  |  |
| --- | --- | --- | --- |
| DATA | R2\_COEFF | MAE | MSE |
| TRAINING | 0.9980 | 0.2164 | 0.0950 |
| VALIDATION | 0.9947 | 0.2906 | 0.2589 |
| TESTING | 0.9954 | 0.2747 | 0.2160 |

As the Validation had similar results with the training and testing data set, we can confirm that the model has **not overfitted.**

* **The Goalkeeper Data alone**

|  |  |  |  |
| --- | --- | --- | --- |
| DATA | R2\_COEFF | MAE | MSE |
| TRAINING | 0.9990 | 0.1220 | 0.0555 |
| VALIDATION | -2469563.0000 | 0.7387 | 0.9701 |
| TESTING | 0.9678 | 0.8739 | 1.8858 |

As the Validation had a huge difference with the training and testing data set, we can confirm that the model has **overfitted.** Also the performance of the model was poorer than the previous one.

* **The Preprocessed Data minus the goalkeeper data**

|  |  |  |  |
| --- | --- | --- | --- |
| DATA | R2\_COEFF | MAE | MSE |
| TRAINING | 0.9973 | 0.2260 | 0.1243 |
| VALIDATION | -324572.4688 | 0.2925 | 0.3327 |
| TESTING | 0.9956 | 0.2692 | 0.2012 |

As the Validation had a huge difference with the training and testing data set, we can confirm that the model has **overfitted.**

**Therefore we choose the model and data of case 1 as our final model and compared them to other standard regression models as base.**

**Results**

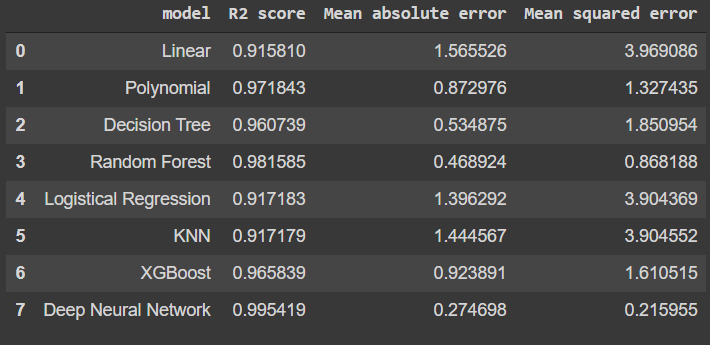


Fig 7: Comparison of Neural Network Model with other standard regression models on Test data

**Final Conclusion**

As the neural network model gave better results compared to other regression models on the original data set, we decided to select that for our data prediction.

**Model Selected - Neural Networks based on Keras**

**Results on Test Data -**

* **R2 Score - 0.9954**
* **MAE - 0.2746**
* **MSE - 0.2159**