**Supervised Learning Data Report**

**Paul Ndirangu**

**Moringa School**

**Table of Contents**

[**Business Understanding**](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.1szv16x3d2n8)[**3**](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.1szv16x3d2n8)

[Problem Statement](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.w4tt6186gpvz) [3](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.w4tt6186gpvz)

[Objective](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.v40uaht3l3p3) [4](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.v40uaht3l3p3)

[**Data Understanding**](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.46jpe3o78zx0)[**4**](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.46jpe3o78zx0)

[**Data Preparation**](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.ylbwipoi2qu8)[**6**](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.ylbwipoi2qu8)

Data Cleaning 6

Label Encoding 6

[Data Exploration](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.bzy0jbk10mux) [7](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.bzy0jbk10mux)

[Univariate Analysis](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.uexq43fir8t8) [7](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.uexq43fir8t8)

[Bivariate Analysis](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.ccdko87dx0d8) [8](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.ccdko87dx0d8)

Supervised Learning [9](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.5abdgxefxj1f)

[Modelling](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.m4pf6nfwn90t) [9](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.m4pf6nfwn90t)

Polynomial Regression [9](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.9umx8ksb7lpi)

Logistic Regression [10](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.wf40uos7ok43)

[**Evaluation**](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.jmxzf54b7pnz) [11](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.jmxzf54b7pnz)

[**Deployment**](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.e5uoayczri8s) [11](https://docs.google.com/document/d/1QDaiLRd1RrCv7c9EbY8Q19pZ-8Ol7wqTRt1IIdNDVd4/edit?ts=5ec39539#heading=h.e5uoayczri8s)

Resources :

***Python Notebook***

(“<https://colab.research.google.com/drive/1t9_H3Y5xvMaubupeiX8gg0FBnLHQv4jq?usp=sharing>”)

***Datasets***

(“<https://drive.google.com/open?id=1BYUqaEEnFtAe5lvzJh9lpVpR2MAvERUc>”)

***Github Repository***

(“<https://github.com/Paul-mwaura/Supervised-Learning-using-Polynormial-Regression-and-Logistic-Regression>”)

**Business Understanding**

Problem Statement

Sport brings the world together. From Africa to Europe, from Asia to South America, and from grassroots to the highest levels, the love of football crosses all borders, cultures, races and ages. This said, it is always a good idea to have the prior knowledge about the outcome of any particular outcome. Even though getting an accurate prediction result is not always assured, having prediction models makes a big difference in the way football companies and sponsors conduct their businesses.

The data available for our objective in question includes the performances of several football teams when playing on their home ground and when playing on away grounds. The data will be analyzed and insights can be acquired from it to enable us to create a predictive model using Polynomial and Logistic regression. These two model approaches will enable us to predict the outcome of a football match given the home/away team rank, the tournament being played and the scores for previous away and home teams.

Objective

The objective of this analysis is to investigate football performances of home and away teams and in turn predict the results of any random teams based on the model we will create. The model will provide a prediction result of a game between team 1 and team 2, based on who's home and who's away, and on whether or not the game is friendly and with rank included in our training.

**Data Understanding**

The data used for this analysis was provided from a Google Drive in the form of a csv file. Two datasets are provided ie; fifa\_ranking.csv and results.csv. These two datasets contain the data required for the model prediction. Merging the datasets we get a DataFrame that has 950 rows and 24 columns. These columns include;

**Fifa\_ranking dataset columns explained**

1. Rank - The position of a country at June 2018

2. Country Full- The country's name

3. total Points - These are the total points a country has accrued after playing against other countries

4. previous\_points - THe points before changes were made.

5. Rank\_change - This is the positive or negative points taken or lost between the time of doing the rank adjustments.

6. cur\_year\_avg - Current Year average- Points accrued in the last before obtaining the dataset(probably 2018)

7. cur\_year\_avg\_weighted - The average of the points per games played accrued in the year last year of obtaining the dataset only

8. last\_year\_avg - Points accrued in the year before obtaining the dataset(probably 2017)

9. last\_year\_avg\_weighted - The average of the points per games played accrued in the year before year of obtaining the dataset only

10. two\_year\_ago\_avg - Points accrued in the second year before obtaining the dataset(probably 2016)

11. two\_year\_ago\_weighted - The average of the points per games played accrued in the second year of obtaining the dataset only

12. three\_year\_ago\_avg - Points accrued in the third year before obtaining the dataset(probably 2015)

13. three\_year\_ago\_weighted - The average of the points per games played accrued in the third year of obtaining the dataset only

14. confederation - The Competition played. This is mostly dependent on the countries location

15. rank\_date - THe specific date the ranking points were updated.

**Results Dataset Columns explained**

1. date - The specific Timestamp of the game being played.

2. home\_team - The team hosting the away team.(Though this is subject to neutral being false).

3. away\_team - The visiting team(Though this is subject to neutral being false)

4. home\_score - This is the number of goals the home team has scored/ Away team has conceded.

5. away\_score - This is the number of goals the away team has scored/ Home team has conceded.

6. tournament - This is the name of the tournament being played. This is dependent on the Countries location eg. CAF and the country's football prowess to attend global competitions e.g. World Cup.

7. city - The city where the game was hosted.

8. country- The country where the game was hosted

9. neutral -If a ground is neutral it means none of the teams played in their own country(Mostly for world cup).

# **Data Preparation**

Data preparation included cleaning the dataset and selecting the variables that would be used for univariate analysis, bivariate analysis. Categorical columns(columns with string values) had their values transformed into integers using label encoding and One hot encoding.

Data cleaning was the first data preparation step taken. Data cleaning involved checking for duplicated values and removing these values. The dataset had null values which were dropped. The dataset had outliers in the numerical columns which were also dropped. It is important to note that the outliers in the numerical columns were less than 2% of all the entries thus dropping them did not significantly affect our dataset.

By generating the frequency table for the categorical variable, it was possible to plot frequency distributions and pie charts for specific variables. A frequency distribution table was also developed from the numerical variables.

Only the numeric variables were used in the bivariate analysis. The data frame that was separated for univariate analysis was used for bivariate analysis. Similarly, only the numeric variables were used to plot the correlation matrix as well as compute the Pearson correlation coefficient.

## **Data Exploration**

The data was explored at two levels namely univariate, bivariate, where we were able to identify the relationships between the columns provided.

### **Univariate Analysis**

For univariate analysis, a frequency table, bar graphs and frequency distributions were used to summarize categorical variables. The categorical variables in the dataset include a city, tournament type, country, country abbreviation, neutral, home team and the away team.

### **Bivariate Analysis**

The bivariate analysis was conducted on numeric variables which included

rank

total\_points

Previous\_points

rank\_change

Cur\_year\_avg

Cur\_year\_avg\_weighted

Last\_year\_avg

Last\_year\_avg\_weighted

Two\_year\_ago\_avg

two\_year\_ago\_weighted

Three\_year\_ago\_avg

Three\_year\_ago\_weighted

Home\_score

away\_score

The correlation matrix also shows that columns that are highly correlated with the total home or away scores.

#### **Modelling**

The data was modelled using regression techniques. Regression is a process of predicting the class of given data points. This type of predictive modelling allows data scientists to approximate a mapping function from input variables to output variables. Two regression techniques were used and these are Polynomial regression and Logistics regression. These types of regression cannot deal with non-linear relationships or interactions.

The polynomial regression enables us to predict the game scores for both the home and away teams. The Logistics approach helps us to predict the outcome of any game whether it was a win, draw or a loss for both the home team and away team.

## **Evaluation**

The models built to predict the day type by inputting the independent variables we were able to predict whether a home or away team would win, draw or lose give the teams rank, the tournament type and whether the team was home or away. The conclusion from this is that most football teams tend to win if they played on their home ground and their performance was poor when playing on an away football ground.

## **Deployment**

Analysis conducted will be implemented by providing a summary of the variables analyzed to help Mchezopesa Ltd predict with a good amount of certainty the outcome of any football match played in a particular tournament. The information on the game scores outcome will enable the company to plan accordingly to ensure that they are able to accurately invest in any particular game with the chances of getting losses minimized. This information can be used to create strategies for the specific teams when they play against other teams since they will have a clue on how the game might turn out.