
**CST3990: Undergraduate Individual Project**

# Data Analysis On International Football Results Over The Last 150 Years And Predicting Future Footballing Results Through Machine Learning

**Coursework 1: Project Proposal**
**Submission: Sunday 12th November 2023, 23:50hrs**
**(End of Week 7)**

**Student ID:M00810926**

**Name: Abdullahi Mohamed**

**CAMPUS: HENDON**

**CST3990 Individual Project – Project Proposal**

# Data Analysis On International Football Results Over The Last 150 Years And Predicting Future Footballing Results Through Machine Learning.

## Introduction to subject area

The project goal is to analyse data of football results and find out who were the best teams in each generation and how well do the host country perform. The project also consists of machine learning to do predictive analysis and forecast the winning teams, player performance, goal scorers and the necessary prediction that effects football results together with data visualisation of historical data to support the forecasting.

## Problem within subject matter

There are several problems within the predictive and descriptive analysis of football result, the factors that affect sporting matches can vary, an example would be is the hosting country likelier to achieve success and win more matches because of home advantage rather than the opposing teams. The problem can be solved through data exploration using a query programming language which can identify appropriate findings that helps in making a data driven decision and provides insight to those interested in the topic.

## Reason to investigate the chosen problem.

- Informative to stakeholders such as football fans understanding the best teams in each generation can further their interests within the sport.
- Predictive models based on historical data raises engagement and excitement within the sport, this enhances fans experience.
- Financial benefits for betting agencies through analysing past results and patterns helps future betting for fans.
- Correlation between different gaming scenarios and how it leads to victories.
- Analyse the evolution of football how does results from the 80s compare to present results.

## Aim of the project

The projects goals are to gather an intellectual insight into past footballing results by using methodologies such data analysis and machine learning. An in-depth analysis on the data to find patterns, trends, and other common factors that lead to the success or failure of teams. Precise predictive models allow for prediction of future footballing results, player performance and goal scorers. The completion of this project furthers knowledge of football dynamics scenarios and a new perspective for those interested in sports analysis.

## Objectives

This project is complex and is divided into sections:

- Perform data analysis on past international football outcomes going back more than 150 years.
- Data cleaning, exploring, visualizing data using software is such as Excel, SQL, and Tableau
- Machine learning in Python using Jupyter Notebook forecasting football results.
- Provide a solid understanding of sports dynamics to fans.
- Advanced knowledge of football analytics and insightful information to stakeholders
- Analyse trends and patterns throughout and make a logical decision moving forward.

**Milestone**

- Milestone 1: Data cleaning and preparation in Excel: completed by week 9.
- Milestone 2: Submit Literature review and initial development in week 13.
- Milestone 3: Data explore in SQL: completed by week 13.
- Milestone 4: Data visualization in Tableau: completed by week 17.
- Milestone 5: Machine learning in Python: completed by week 21.
- Milestone 6: Submit Final Report in Week 23

**Motivation**

Interest is global in football and a driving force for this project, using historical data to identify complex patterns helps with engagement and provides fans a great insight to football through this project by using data and predictive analysis on football results.

**Risk Analysis**

- Starting the project too late can lead to a lot of errors further down, therefore following the Gantt chart can help avoid this problem.
- Having minimal rows in dataset can lead to poor analysis and findings and the project will lack purpose, hence finding a dataset that has enough rows is necessary.
- Not having regular meetings with supervisor, the project will lack guidance and can lead to failure, therefore a set appointment with supervisor can help achieve a guided project.
- Delegation to learning the skills necessary for example an hour of learning SQL enhances the data exploration section of project.

**Contingency time**

Problem solving allows to enhance the projects to greater heights, the extra time allows for improvement, dealing with uncertainty gives the project the flexibility and direction.

**Literature Review Section**

Background research guides analytical project, obtaining a literature evaluation allows for a guided use of data analysis and machine learning in the project.

**Title:** Forecasting football results and the efficiency of fixed-odds betting

**Description:** This research paper focuses on forecasting football results, and it uses ten years of data to produce a regression model. The usefulness to the project is that it can help understand the reasons behind why the teams are favoured to win, it mentions about the common pattern between the money spent per team and wins a season. (John Goddard., & Ioannis Asimakopoulos,2004)

**Title**: Predictive analysis and modelling football results using machine learning approach for English Premier League

**Description:** The focus of this paper is predicting football results using machine learning which is directly useful to the project because it uses Python for the modelling of the project. In addition, not only does to it focus on the results, but it also goes into detail on all factors of football such as it goals per game, minutes played and team budget analysis and how it effects the outcome. (Rahul Baboota & Harleen Kaur, 2019)

**Title:** Data Visualization of Football Performance Preceded to the Goal Scored

**Description:** This research paper focuses on data visualization which is a key section of the project on enhancing team performance, the usefulness to this project is figuring out if match statistics help football teams to win or not. (Zulkifli Mohamad, 2023)

**Title:** Analysis and visualisations of team performances of football games

**Description:** The article is based on analysing team performances of football games including the goals scored and match statistics. Similarly, this project requires analysis of football games to conduct and that the article is a great introductory of data visualisations. (Roberto Gásquez & Vicente Royuela, 2016)

**Title:** The Determinants of International Football Success: A Panel Data Analysis of the Elo Rating

**Description:** The article focuses on the determinants of international football success and has a variety of countries and a thirty-three-year period of matches which is directly relatable to this project because the success of football teams in data analysis is a component of the project. (Tsuneshi Obata & Shizue Izumi, 2022)

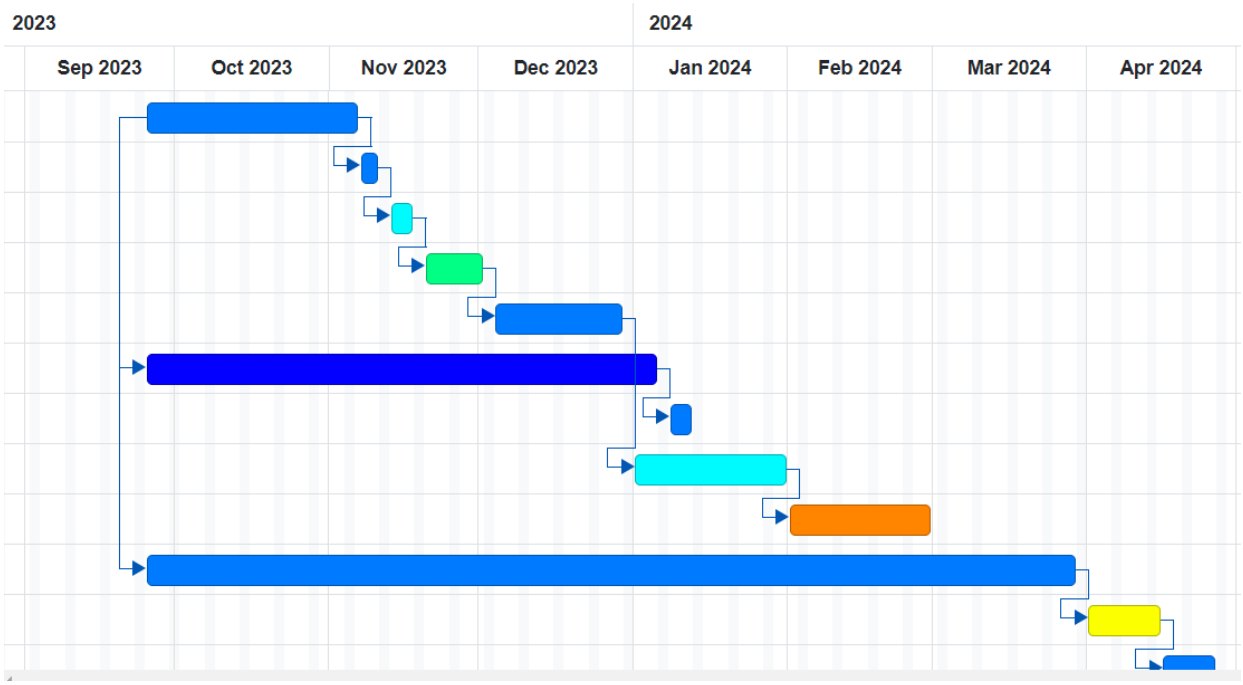
**Gantt Chart**

Gantt chart displays the order and timeline for the completion of the project, this aligns with all deadlines and the various stages needed to complete this project.

**Table**

| | |
|---|---|
| 1 | Write Proposal |
| 2 | Review and submit Proposal |
| 3 | Find Datasets on Kaggle |
| 4 | Data cleaning and preparation on Excel |
| 5 | Data exploration on SQL |
| 6 | Write background and literature review and init… |
| 7 | Review and submit literature review and devel… |
| 8 | Data visualize on Tableau |
| 9 | Machine Learning on Jupyter Notebook |
| 10 | Write Final Report |
| 11 | Finalize and Review all stages |
| 12 | Submit Final Report |

## Graph




## Evaluation

The project's integration of Python, SQL, Tableau, and Excel displays data analysis tools and techniques. By addressing potential risks depicts a responsible and an element of project management, the work distributed technical skill and a logical understanding of football analytics. Around 80% of the data will be trained and 20% will be evaluated to predict football results to evaluate how good the prediction is.

## Resources

Plenty of resources is necessary for this project, including three large datasets with over 40,000 rows with all are obtained from a platform called Kaggle. Excel is to data prepare and clean to ensure its ready for analysis, SQL purpose is to discover patterns in football results, Tableau will be used to produce interactive dashboard for football results, Python based machine learning algorithms provide predictive analysis that transform this project, allows to forecast football results, and provide an evaluation of this project.

## References

## Article

John Goddard., & Ioannis Asimakopoulos (2004). Forecasting football results and the efficiency of fixed odds betting. Wiley online library, 21(1),51-66, https://onlinelibrary.wiley.com/doi/10.1002/for.877

Rahul Baboota & Harleen Kaur (2019), Predictive analysis and modelling football results using machine learning approach for English Premier League, Science Direct, 35(2), 741-755, https://www.sciencedirect.com/science/article/abs/pii/S0169207018300116

Tsuneshi Obata & Shizue Izumi (2022), Analysis and visualisations of team performances of football games, Data science: Present and Future, 5, 885- 898, https://link.springer.com/article/10.1007/s42081-022-00173-z

Roberto Gásquez & Vicente Royuela (2016), The Determinants of International Football Success: A Panel Data Analysis of the Elo Rating, Wiley online library, 97(2), 125-141, https://onlinelibrary.wiley.com/doi/full/10.1111/ssqu.12262#ssqu12262-note-0029

**Conference**

Zulkifli Mohamad, Data Visualization of Football Performance Preceded to the Goal Scored, in K. Imran, innovation and technology in sports, (pp. 57-74). Springer: https://link.springer.com/chapter/10.1007/978-981-99-0297-2_6

**Website**

Kaggle (2023), international football results from 1872 to 2023, available from data card at: https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017 ![date accessed 18th October 2023]

Microsoft. (2023), Microsoft Excel, available from Microsoft: https://www.microsoft.com/en-gb/microsoft-365/excel ![date accessed 28th October 2023]

MySQL. (2023), MySQL, available from MySQL: https://www.thesql.com/ ![date accessed 29th October 2023]

Jupyter. (2023), Jupyter Notebook, available from Jupyter at: https://jupyter.org/ ![date accessed 29th October 2023]

Tableau. (2023), Tableau Public, available from Tableau at: https://www.tableau.com/en-gb ![date accessed 30th October 2023]

ChatGPT. [2023], ChatGPT, available from ChatGPT at: https://chat.openai.com/ ![date accessed 31st October 2023]

## **Appendix: Football Results Analysis and Predictive Modelling**

**Idea: Analysing Football Results and Predicting Future Outcomes**

**Tools and Techniques: **

- **Data Cleaning with Excel: **
  - Remove inconsistencies and manage missing values in the dataset.
  - Ensure data uniformity and accuracy for meaningful analysis.
- **Exploratory Analysis using SQL: **
  - Utilize SQL queries to identify trends in historical match outcomes.
  - Explore factors like home advantage, player performance metrics, and team patterns.
- **Visualization with Tableau: **

- Create interactive dashboards displaying historical match data.

- Utilize heat maps, bar charts, and line graphs for intuitive data representation.

**- **Predictive modelling with Python: ****

- Implement regression models to predict match outcomes based on historical data.

- Use classification algorithms to forecast specific events such as goal scorers or player performance metrics.

**- **Integration and Iterative Analysis: ****

- Integrate cleaned data from Excel and SQL into Tableau for dynamic visualization.

- Iterate through the analysis process, refining visualizations and predictive models based on insights gained.

****Benefits and Objectives: ****

- Gain deep insights into historical football match outcomes and team performances.

- Develop accurate predictive models for forecasting future match results and player performances.

- Enhance understanding of influential factors such as player statistics.

- Contribute valuable insights to the field of football analytics and sports prediction.

Certainly! Adding detailed information and ideas to the appendix of your research can enhance the depth and credibility of your work. Here are some ideas you can consider including in the appendix of your research on data analysis of international football results and predicting future outcomes through machine learning:

1. **Data Collection Methods: **

- Provide a detailed description of the sources from which you collected the historical international football data. This could include websites, databases, or any APIs used for data extraction.

- Mention the specific variables you collected, such as match outcomes, scores, team rankings, player statistics, weather conditions, etc.

2. **Data Preprocessing Techniques: **

- Explain the steps taken to clean and preprocess the raw data. This could involve handling missing values, data normalization, outlier detection, and feature engineering methods applied to the dataset.

- Include any data transformation or encoding techniques used to prepare the data for machine learning algorithms.

3. **Feature Selection and Engineering: **

- Describe the features you selected for your machine learning models and the rationale behind their selection. Discuss how you engineered new features from the existing dataset to improve predictive performance.

4. **Machine Learning Models: **

- Provide a brief overview of the machine learning algorithms you experimented with, such as decision trees, random forests, neural networks, etc.

- Explain the hyperparameters you tuned, and the methods used for model evaluation, such as cross-validation techniques.

5. **Results and Performance Metrics: **

- Include tables or charts displaying the results of your machine learning models. This could involve accuracy, precision, recall, F1-score, or any other relevant metrics.

- Compare the performance of different models and highlight which model performed the best and why.

6. **Challenges Faced: **

- Discuss any challenges encountered during the data analysis and modelling process. This could include data inconsistencies, overfitting issues, or limitations in the dataset.

7. **Future Work and Recommendations: **

- Suggest potential areas for future research related to international football data analysis and machine learning. Provide recommendations on how the study could be expanded or improved.

8. **Code Snippets: **

- If applicable, include relevant code snippets or algorithms used in your analysis. This can be helpful for readers who want to replicate or further explore your work.

9. **Visualizations: **

- Include visualizations that did not make it into the main body of your research paper due to space constraints. Visual representations of data trends and patterns can be insightful.

10. **References: **

- List all the references, datasets, libraries, and research papers you referred to during your study. Proper citation is essential for academic integrity.

Remember, the content in the appendix should complement and support the main body of your research. Make sure to refer to the appendix in your main text when appropriate and explain the relevance of the information provided in the appendix to your study.

**-- **Note: ****

The content in this appendix was generated by ChatGPT, an advanced language model developed by OpenAI, as per the suggestion of the professor. ChatGPT aided in formulating the content and structuring the ideas presented here and was advised by **supervisor**.