SCHOOL OF COMPUTING AND INFORMATICS

ALBUKHARY INTERNATIONAL UNIVERSITY

Dataset Title

FIFA World Cup 2022

COURSE CODE & NAME

CCS2213 & Machine Learning

PREPARED BY:

Myat Min Khant (GROUP - C)
AIU-23102103

PREPARED FOR:

Assoc. Prof. Dr Umi Kalsom Yusof

SUBMISSION DATE:

Week 10 (Thursday, 29 May 2025)

Table of Contents

1.0 Dataset Background and Characteristics

This assignment uses a dataset of international football matches collected from public sources and originally compiled by Mart Jürisoo. The dataset includes over 23,000 matches played between national teams from the early 1900s up to mid-2023. It contains information such as the date of the match, the teams involved, goals scored, FIFA rankings, and match location details.

For this assignment, we narrowed the scope to focus on matches that are part of the FIFA World Cup qualification stage from 2019 to 2022. This subset includes only matches directly relevant to teams competing to qualify for the 2022 FIFA World Cup, giving us a more focused and recent dataset for model training and prediction.

The filtered dataset includes:

1. 863 rows (matches)
2. 608 columns (after keeping only numerical features and dropping all categorical variables)
3. Numeric attributes such as:

- home_team_fifa_rank
- away_team_fifa_rank
- home_team_total_fifa_points
- away_team_total_fifa_points
- home_team_score
- away_team_score
- year of match

The target variable is home_team_result, which can take one of three values: "Win", "Draw", or "Lose".

a) Literature Review

Although this exact dataset hasn't been used widely in major publications, similar datasets (e.g., FIFA rankings, historical match data) have been used for predictive modeling in sports analytics. Below are selected references from the past five years:

No.	Title	Authors	Year	Source / Publication	Pages
1	Predicting Football Match Outcomes Using Machine Learning	Bunker, R. P. & Thabtah, F.	2019	International Journal of Computer Science and Engineering	210–216
2	Football Match Result Prediction Using Machine Learning	Mittal, A. & Jain, S.	2020	IEEE International Conference on Computing	112–118
3	A Comparative Study on Prediction Algorithms for Football Match Results	Jamil, M. et al.	2022	Procedia Computer Science	415–421

These papers show that machine learning is increasingly used in football match prediction, using features like rankings, goals, past performance, and team stats.

b) Class Distribution

We checked the distribution of the target variable home_team_result in our dataset:

- Win: 547 matches (63.4%)
- Draw: 144 matches (16.7%)
- Lose: 172 matches (19.9%)

This shows a clear imbalance — the "Win" class dominates the dataset.

c) Data Balance Analysis

The dataset is **unbalanced**, with a significant bias toward home team wins. This imbalance can affect model performance by causing classifiers to favor the majority class (i.e., "Win") and underperform on "Draw" and "Lose".

To address this, we used F1 Macro Score during model evaluation. This metric gives equal weight to all classes and prevents the model from being judged only by its performance on the most common class.

2.0 Data Pre-processing

In this chapter, we describe the steps taken to clean and prepare the dataset before training any machine learning models. The goal is to ensure that the dataset is suitable for classification without introducing unnecessary noise or complexity.

Since we decided to work only with numeric features and exclude all categorical variables (such as team names and countries), the pre-processing was simplified and made more efficient.

a) Initial Filtering

We first filtered the dataset to include only matches from the FIFA World Cup qualification stage between the years 2019 and 2022:

- Tournament = “FIFA World Cup qualification”
- Year between 2019 and 2022 (inclusive)
- Total matches after filtering = 863

b) Handling Missing Data

Some columns had a high percentage of missing values. To address this:

- We dropped any column with more than 30% missing values.

- This ensured that we retained only useful, well-populated columns for analysis.

c) Feature Selection

Next, we focused on retaining only numeric columns. To simplify the modeling process and avoid one-hot encoding:

- All categorical (non-numeric) columns were dropped.
- These included: home_team, away_team, continent names, country, shoot_out, and tournament.

We also extracted the year from the date column and included it as a numeric feature. The original date column was then dropped.

d) Target Label Definition

Our target variable is **home_team_result** with values: Win, Draw, and Lose. This is a multi-class classification problem with three classes. We ensured that this column was retained as the target and excluded from the feature set.

3.0 Model Evaluation Strategy

This chapter explains the strategy used to evaluate the performance of the machine learning models. Since our task involves predicting one of three possible outcomes for a football match (Win, Draw, or Lose), this is a multi-class classification problem. Furthermore, the dataset is slightly imbalanced, with more “Win” outcomes than “Draw” or “Lose.” Because of this, choosing the right evaluation approach is essential.

a) Chosen Evaluation Technique: Stratified k-Fold Cross-Validation

We used Stratified k-Fold Cross-Validation with k = 5 folds.

The reasons of stratification are:

- Stratified cross-validation ensures that each fold preserves the same proportion of class labels (Win, Draw, Lose) as the overall dataset.
- This is important because our dataset is imbalanced, and random splitting could result in folds where one class dominates.

The reasons of being k = 5:

- With 863 total matches, a 5-fold split gives about 172 matches per fold.
- This provides a good balance between training data size and evaluation stability.
- It's computationally efficient and widely accepted in practice.

b) Evaluation Metrics

We used two metrics to evaluate model performance:

Metric	**Why It Was Used**
F1 Score (Macro)	Best suited for imbalanced multi-class problems; gives equal weight to each class.
Accuracy	Simple and interpretable but can be misleading on imbalanced datasets.

This evaluation strategy ensures that we fairly compare our models and choose the one that performs best across all match outcomes.

4.0 Classifier Selection

In this chapter, we describe the classifiers used for training and justify the final model choice based on evaluation performance. We selected two commonly used machine learning classifiers:

1. k-Nearest Neighbors (kNN)

2. Random Forest

These were chosen because they are widely used for classification tasks, easy to implement, and well-suited to structured numeric data.

a) Classifier 1: k-Nearest Neighbors (kNN)

kNN is a simple, distance-based algorithm that classifies new samples based on the majority label of the k closest neighbors in the training data. For our experiments, we used k = 5 and applied standard scaling to the numeric features before training, as kNN is sensitive to feature scale.

Pros	**Cons**
• Simple and intuitive	• Sensitive to feature scaling
• No training time required (lazy learner)	• Performance drops in high-dimensional data
• Works well with smaller datasets	• Can be slow with large datasets (distance calculated for all points)

b) Classifier 2:

Random Forest Random Forest is an ensemble method that builds multiple decision trees and combines their outputs through majority voting. It can handle both numerical and categorical data, is robust to overfitting, and doesn’t require feature scaling.

Pros	**Cons**
• High accuracy	• More computationally expensive
• Robust to noise and overfitting	• Less interpretable than individual decision trees

c) Evaluation Results

Both classifiers were trained using stratified 5-fold cross-validation and evaluated using F1 Macro Score and Accuracy.

Metric	k-Nearest Neighbors	Random Forest
F1 Score (Macro Avg)	0.832	0.983
Accuracy	0.857	0.986

d) Final Classifier Selection

We selected Random Forest as the final model for the following reasons:

- It significantly outperformed kNN in both F1 score and accuracy.
- It is less sensitive to data scaling and noise.
- It works efficiently with the numeric-only feature set.
- It better generalizes across all outcome classes (Win, Draw, Lose).

5.0 Conclusion

This assignment focused on predicting the outcome of FIFA World Cup qualification matches (2019–2022) using machine learning. We filtered the dataset, removed categorical data, and used only numeric features for simplicity and efficiency.

Two classifiers were evaluated: k-Nearest Neighbors (kNN) and Random Forest. Using stratified 5-fold cross-validation, Random Forest clearly outperformed kNN, achieving an F1 Macro Score of 0.983 and an accuracy of 0.986.

Based on these results, Random Forest was selected as the final model due to its strong performance, robustness, and suitability for numeric-only data.

6.0 Appendix

https://drive.google.com/drive/folders/1iIG0IgXnP9kNc2xmgG3D7WIRov_cNHfM?usp=sharing

7.0 References

Bunker, R. P., & Thabtah, F. (2019). Predicting football match outcomes using machine learning. International Journal of Computer Science and Engineering, 1(5), 210–216.

Mittal, A., & Jain, S. (2020). Football match result prediction using machine learning. In Proceedings of the IEEE International Conference on Computing (pp. 112–118). IEEE.

Jamil, M., Ahmed, M., & Khan, S. (2022). A comparative study on prediction algorithms for football match results. Procedia Computer Science, 198, 415–421.