Seattle University
CPSC 5071 01 26WQ Data Management for Data Sci
Professor: Mr. Yueting Chen

Predicting FIFA World Cup 2026 Outcomes Using SQL and Machine Learning in Python

Team Members

Hemant Kumaar Aruljothi
Prithika Kandasamy

Problem Statement

Predicting outcomes in international football tournaments is a complex analytical challenge due to the interaction of multiple dynamic and context-dependent factors, including team strength, player performance, historical results, tactical strategies, and environmental conditions. The FIFA World Cup 2026 will be hosted across multiple countries, introducing significant variability in climate, travel demands, and playing environments that may influence match dynamics and competitive performance. This project aims to develop a predictive framework capable of estimating winners of individual matches as well as the overall tournament champion for the FIFA World Cup 2026. The model will be constructed using multiple complementary datasets, including historical FIFA match results, world rankings, and player-level performance indicators, to capture both long-term competitive trends and short-term performance variations across teams.

In addition to traditional performance metrics, the project explicitly incorporates environmental conditions as an explanatory factor in predictive modeling. Weather variables such as rainfall, humidity, temperature, and cloud cover may influence physical endurance, tactical decision-making, and match tempo, thereby affecting competitive outcomes. By analyzing how teams historically perform under different environmental scenarios, the model will estimate match results by jointly considering competitive strength and environmental context.

Datasets

1. International Match Results Dataset

Source: Kaggle – International Football Results
This dataset contains historical match information including teams, match dates, scores, tournament type, and match location. It is essential for analyzing team performance trends and historical outcomes.
Preprocessing needs include:

- Handling missing or incomplete match records
- Standardizing team names across datasets
- Filtering relevant matches (e.g., recent or World Cup matches)
- Creating performance indicators such as win rate and goal difference

2. FIFA World Rankings Dataset

Source: Kaggle – FIFA World Rankings
This dataset contains official FIFA rankings and ranking points for national teams over time. Rankings provide an objective measure of team strength.
Preprocessing needs include:

- Matching ranking dates with match dates
- Handling inconsistent country naming formats
- Interpolating missing ranking values

- Normalizing ranking scores for modeling

3. FIFA Player Statistics Dataset
Source: Kaggle – FIFA Player Dataset
This dataset includes player ratings, positions, and performance indicators used to evaluate team strength based on individual performance.
Preprocessing needs include:

- Aggregating player statistics at the team level
- Removing inactive or irrelevant players
- Standardizing performance metrics
- Computing team-level features (average rating, etc.)

4. Weather and Location Dataset
Sources:
Kaggle Football Weather Dataset (Direct Sports Weather Data)
https://www.kaggle.com/datasets/tombliss/weather-data

Football Stadium Location Dataset (Match Location Data)
https://github.com/ThompsonJamesBliss/WeatherData
Weather and location data may influence match outcomes due to environmental and geographical conditions such as temperature, humidity, rainfall, altitude, and stadium location. These factors can affect player performance, match tempo, and team strategies under different playing environments.
Preprocessing needs include:

- Matching weather data with match locations and dates
- Integrating stadium location data with match venue information
- Handling missing or incomplete weather records
- Encoding weather conditions (rainy, cloudy, sunny, etc.)
- Aggregating weather features relevant to match time

If additional datasets are required, alternative sources such as official FIFA records or sports analytics repositories will be evaluated based on data completeness and consistency.

Risks

Several risks may affect predictive reliability and model generalizability. Historical datasets may contain missing, inconsistent, or heterogeneous records requiring extensive preprocessing and validation. Integrating multiple data sources with different temporal resolutions and structures introduces alignment complexity. Player injuries, transfers, and evolving squad compositions may reduce the predictive relevance of historical performance indicators. Additionally, competitive sports are inherently stochastic, and unexpected tactical decisions, officiating outcomes, or match-specific events may influence results beyond what historical data alone can capture.

Challenges

A major modeling challenge is representing team performance when squad composition and tactical strategy are not fixed. National teams may deploy different starting lineups depending on opponent strength, tournament stage, or player availability. To address this, the project incorporates both team-level performance metrics and aggregated player-level indicators derived from historical match participation. However, unexpected lineup adjustments or tactical innovations may alter match dynamics in ways that remain difficult to quantify.

Incorporating environmental conditions introduces additional uncertainty. While weather data provides valuable contextual information, future match conditions cannot be predicted with complete certainty.

Historical weather patterns from prior tournaments may not accurately reflect actual match-day conditions. Consequently, although environmental data enhances contextual realism, reliance on past weather observations introduces unavoidable predictive uncertainty that must be carefully interpreted.

References

https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017
https://www.kaggle.com/datasets/cashncarry/fifaworldranking
https://www.kaggle.com/datasets/stefanoleone992/fifa-23-complete-player-dataset
https://www.kaggle.com/datasets/piterfm/fifa-football-world-cup
https://www.kaggle.com/datasets/secareanualin/football-events
https://github.com/ThompsonJamesBliss/WeatherData
https://www.kaggle.com/datasets/tombliss/weather-data
https://www.ncei.noaa.gov/
https://meteostat.net/