Project Phase 2: Project Progress Report

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

Team Members:

Prithika Kandasamy
Hemant Kumaar Aruljothi

1. Problem Statement

In Phase 1, our project proposed using Elo ratings to predict international football match outcomes. Following initial data exploration and integration, we refined both the scope and statistical framing of the problem.

Exploratory Data Analysis revealed:

- Substantial heterogeneity in goal distributions across tournament types.
- Measurable attenuation of home advantage effects in neutral venues.
- A statistically strong association between Elo rating differences and match outcomes.
- Significant missingness in environmental variables requiring structured filtering and aggregation.

Based on these findings, the refined research question is:

To what extent can international football match outcomes (home win, draw, away win) be predicted using structured relational data that integrates Elo rating differentials, contextual

match features (venue type, tournament classification), and environmental covariates, within a supervised machine learning framework?

This updated formulation reflects:

- A shift from purely descriptive modeling to predictive inference.
- Explicit consideration of confounding factors such as venue neutrality.
- Evaluation of incremental predictive value contributed by environmental features.
- Integration of relational database design with statistical modeling principles.

The modeling objective is therefore both predictive (classification of accuracy) and inferential (understanding feature importance and explanatory strength).

2. Data Preprocessing

2.1 Data Sources and Structure

We integrated multiple large-scale datasets:

- 49,071 international match records (core outcome dataset)
- 47,555 goal-level observations
- 665 penalty shootout records
- 900 World Cup tournament matches
- 7,907 player records
- 2,548 World Cup goal-level events
- 42,807 environmental observations

All datasets were stored within a structured SQLite relational database (fifa26_prediction.sqlite) to ensure schema consistency, referential integrity, and scalable SQL querying.

2.2 Relational Database Design

A normalized relational schema was constructed to support:

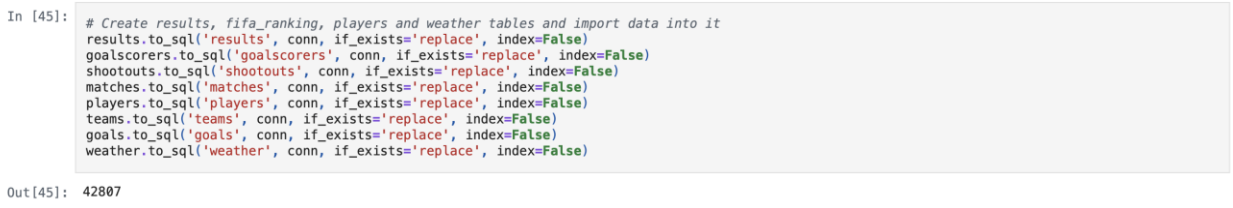
- Efficient joins between match, team, and player tables
- Aggregation of goal-level data into match-level features
- Controlled integration of environmental variables

Tables created:

- results
- goalscorers
- shootouts
- matches
- goals
- players
- teams
- weather

Primary keys and relational joins ensured data consistency across multi-source integration. SQL operations were used extensively for:

- Filtering historical windows for Elo computation
- Group-by aggregations
- Feature extraction prior to modeling
- Integrity checks across foreign keys



```
In [45]:
# Create results, fifa_ranking, players and weather tables and import data into it
results.to_sql('results', conn, if_exists='replace', index=False)
goalscorers.to_sql('goalscorers', conn, if_exists='replace', index=False)
shootouts.to_sql('shootouts', conn, if_exists='replace', index=False)
matches.to_sql('matches', conn, if_exists='replace', index=False)
players.to_sql('players', conn, if_exists='replace', index=False)
teams.to_sql('teams', conn, if_exists='replace', index=False)
goals.to_sql('goals', conn, if_exists='replace', index=False)
weather.to_sql('weather', conn, if_exists='replace', index=False)

Out[45]: 42807
```

Figure 1: SQLite database storage and SQL validation query.

In [155…

```
# Merge results + match_meta (prefer exact join on date+teams)
base_df = results.copy()

if all(c in match_meta.columns for c in ["match_date", "home_team", "away_team"]):
    base_df = base_df.merge(
        match_meta[["match_date", "home_team", "away_team"] + [c for c in ["stage", "stadium", "city"] if c in match_meta.columns]],
        left_on=["date", "home_team", "away_team"],
        right_on=["match_date", "home_team", "away_team"],
        how="left"
    ).drop(columns=["match_date"], errors="ignore")
else:
    # fallback (only if matches lacks team names): merge by date only (low quality)
    if "match_date" in match_meta.columns:
        base_df = base_df.merge(
            match_meta[["match_date"] + [c for c in ["stage", "stadium", "city"] if c in match_meta.columns]],
            left_on="date",
            right_on="match_date",
            how="left"
        ).drop(columns=["match_date"], errors="ignore")
```

Figure 2: Dataset integration using key matching on date and team identifiers.

2.3 Data Cleaning & Validation

The primary results dataset contained no missing values in outcome-defining variables, ensuring statistical robustness in the dependent variable.

Cleaning procedures included:

- Conversion of date variables to datetime objects for temporal analysis.
- Validation of score variables as integer types.
- Removal of duplicate records.
- Handling limited missing values in scorer/minute fields.
- Filtering weather dataset to retain only matches with sufficient alignment.
- Exclusion of non-informative high-missing variables in environmental data.

Data validation checks confirmed:

- No inconsistencies in goal totals.
- Referential integrity between team identifiers.
- Logical consistency in neutral venue indicators.

Data cleaning and Feature engineering

In [48]:
```
# Clean results dataset
results["date"] = pd.to_datetime(results["date"], errors="coerce")
results["neutral"] = results["neutral"].astype(int)

# drop rows where date failed to parse
results = results.dropna(subset=["date"])

# Overwrite the existing table
results.to_sql("results", conn, if_exists="replace", index=False)
```

Out[48]: 49071

In [45]:
```
# Create results, fifa_ranking, players and weather tables and import data into it
results.to_sql('results', conn, if_exists='replace', index=False)
goalscorers.to_sql('goalscorers', conn, if_exists='replace', index=False)
shootouts.to_sql('shootouts', conn, if_exists='replace', index=False)
matches.to_sql('matches', conn, if_exists='replace', index=False)
players.to_sql('players', conn, if_exists='replace', index=False)
teams.to_sql('teams', conn, if_exists='replace', index=False)
goals.to_sql('goals', conn, if_exists='replace', index=False)
weather.to_sql('weather', conn, if_exists='replace', index=False)
```

Out[45]: 42807

Figure 3: Dataset structure and missing value assessment

The results dataset contains 49,071 observations with no missing values in outcome-defining variables.

2.4 Feature Engineering & Normalization

A central contribution of preprocessing was dynamic Elo rating computation using historical match sequences.

Elo ratings were updated iteratively based on:

- Match outcomes
- Rating differences
- Home advantage adjustments

Derived modeling features include:

- Elo rating difference (continuous predictor)
- Encoded categorical outcome (3-class classification)
- Neutral venue indicator (binary)
- Tournament category encoding
- Weather aggregates (temperature, humidity, wind speed where available)

Continuous predictors were standardized when required by model assumptions (e.g., logistic regression).

Categorical variables were encoded using appropriate dummy variable representation.

Feature engineering was performed with attention to avoiding data leakage by ensuring chronological separation of training and test information.

3. Preliminary Results of Exploratory Data Analysis

3.1 Descriptive Distribution Analysis

- Mean goals per match ≈ 2.5
- Goal distribution exhibits right skewness
- High-scoring outliers are rare but influential

Distributional plots confirmed that match outcomes follow expected football scoring patterns.

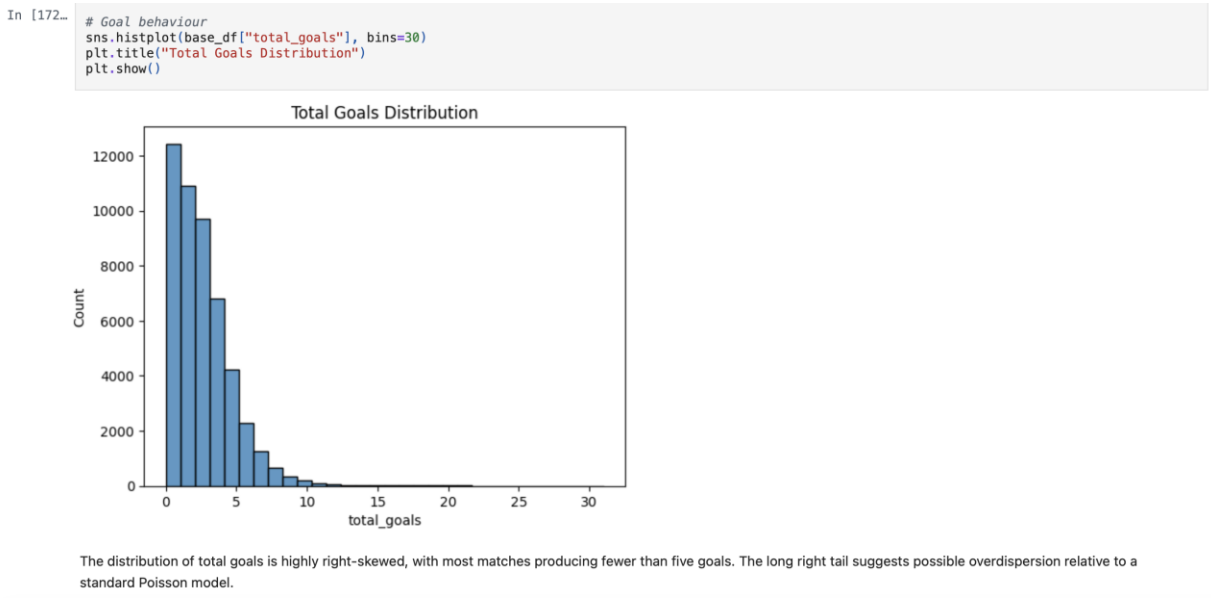


Figure 4: Distribution of total goals per match (right-skewed)

3.2 Elo Differential and Outcome Relationship

Exploratory visualization demonstrated:

- Monotonic increase in home win probability as Elo differential increases.
- Draw probability peaks near zero Elo difference.
- Away win probability increases with large negative Elo differential.

Preliminary correlation analysis indicates a strong directional relationship between Elo difference and match outcome classification.

This validates Elo difference as the primary predictive covariate.

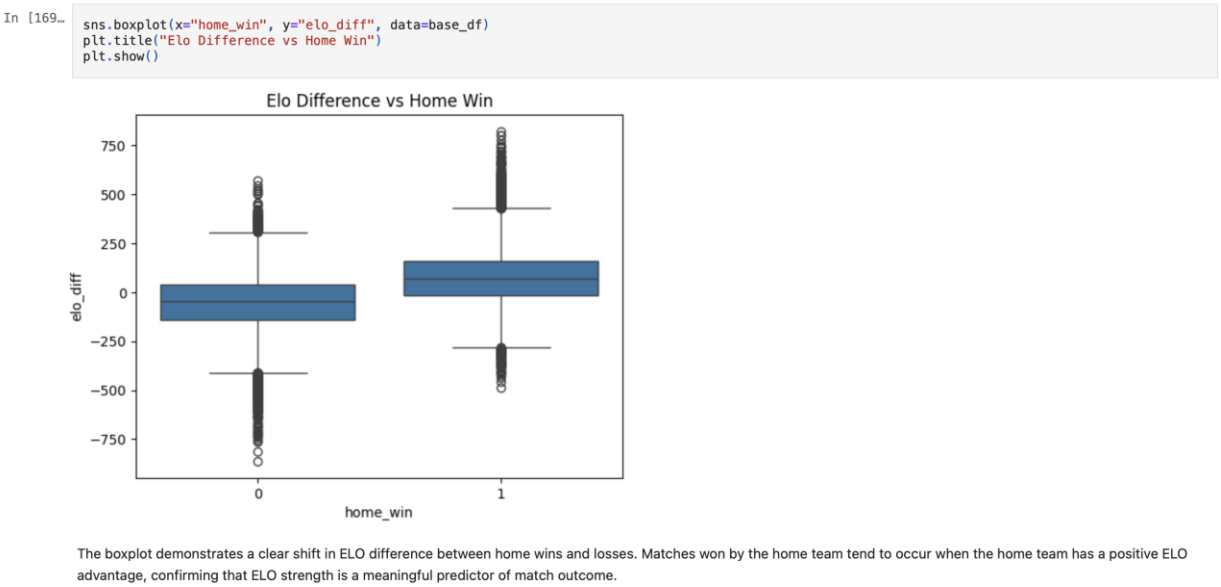


Figure 5: Elo differential distribution across match outcome classes

3.3 Venue Effect

Comparative analysis between neutral and non-neutral matches revealed:

> The observed reduction in home win probability suggests a statistically meaningful venue effect, formal hypothesis testing will be conducted in the modeling phase.

- Reduced variance in outcome probabilities for neutral matches.

This suggests that venue acts as an interaction variable and must be retained in predictive modeling.

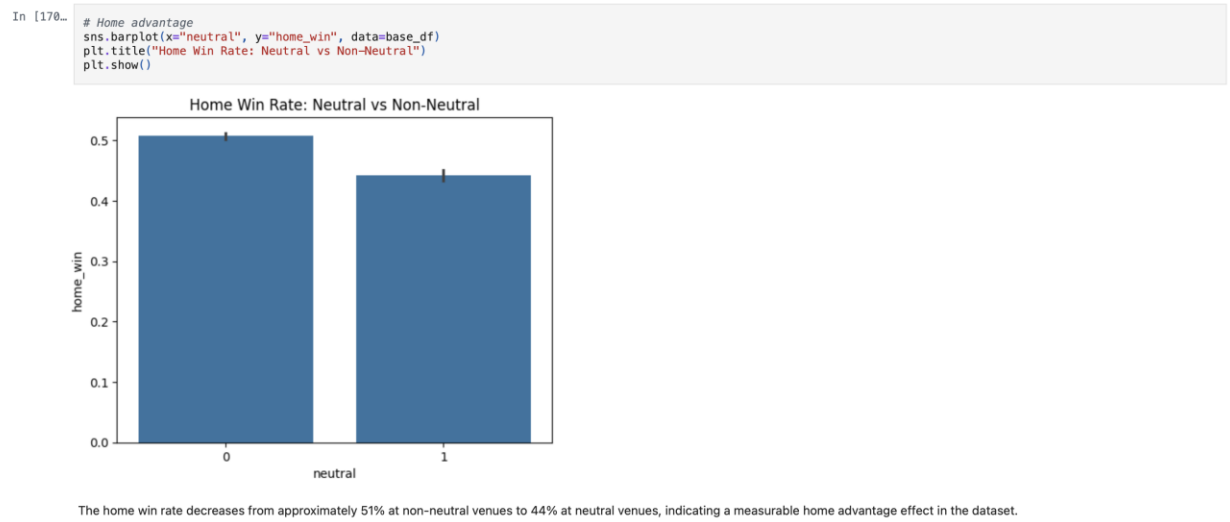


Figure 6: Home win probability decreases from approximately 50.7% in non-neutral venues to 44.2% in neutral venues, providing empirical evidence of a measurable home-field advantage effect.

3.4 Tournament-Level Effects

Tournament category influences scoring intensity and outcome predictability.

Friendly matches show greater variance and unpredictability relative to structured tournaments.

This supports inclusion of tournament encoding in the feature set.

3.5 Environmental Variables

Weather variables show limited linear association with match outcomes.

However, due to partial coverage and aggregation constraints, their predictive contribution will be evaluated empirically through model comparison.

3.6 Multivariate Insight

Correlation matrices and bivariate visualizations suggest:

- Elo difference explains substantial variance in outcome class.
- Neutral venue moderates outcome distribution.
- Tournament category contributes contextual stratification.

The correlation matrix indicates a strong positive association between Elo differential and match outcome encoding, reinforcing its explanatory strength.

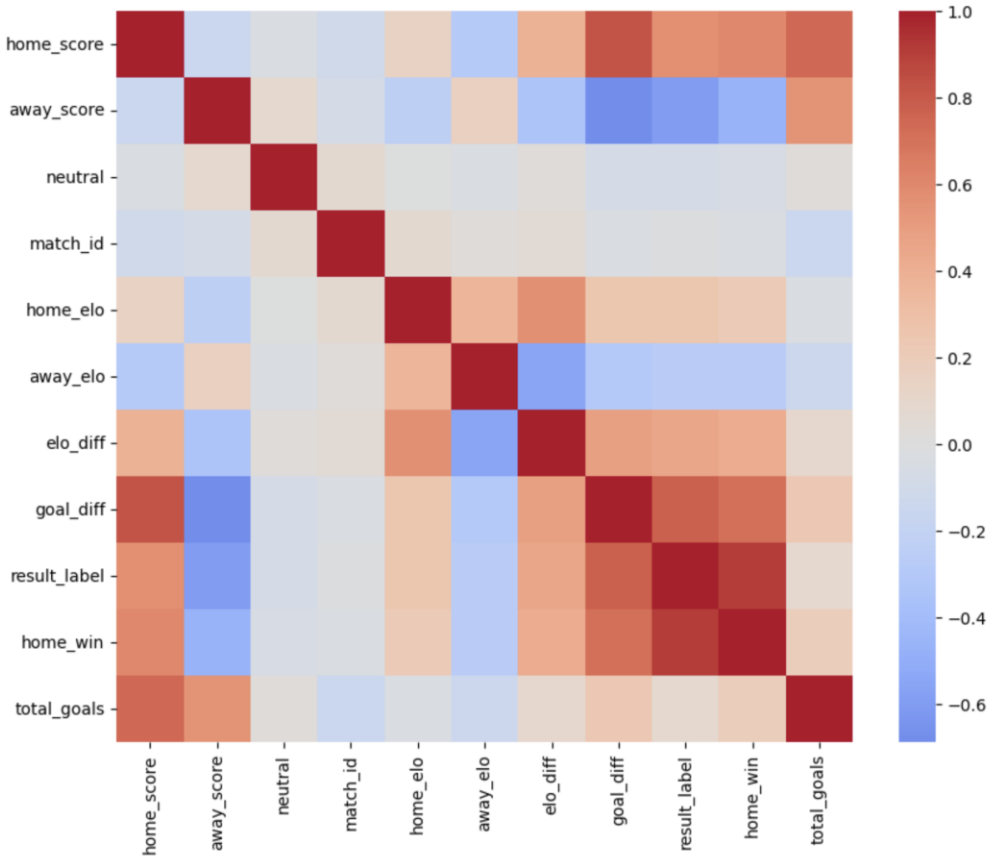


Figure 7: Correlation matrix of engineered predictors

4. Plan for Completion

4.1 Remaining Preprocessing Tasks

- Refine environmental feature aggregation
- Conduct feature importance sensitivity checks
- Ensure chronological integrity in training/testing splits
- Simulation for our prediction

4.2 Modeling Strategy

Planned supervised models:

1. Multinomial Logistic Regression (baseline probabilistic classifier)
2. Random Forest Classifier (nonlinear ensemble model)
3. Gradient Boosted Trees (if time permits)

Evaluation metrics:

- Overall accuracy
- Precision, Recall, F1-score
- Confusion matrix analysis
- Cross-validation performance
- Feature importance ranking

Comparative modeling will quantify the marginal benefit of adding environmental variables beyond Elo-only models.

4.3 Validation Framework

- Train/test split based on chronological separation
- K-fold cross-validation where appropriate
- Sensitivity analysis on hyperparameters
- Overfitting diagnostics with train/test gap

4.4 Team Responsibilities

Hemant Kumaar:

- Database schema design
- SQL data integration

- Elo computation implementation
- Feature engineering and preprocessing pipeline
- Simulation of the prediction

Prithika Kandasamy:

- EDA visualization and statistical analysis
- Model training and validation
- Performance metric evaluation
- Documentation and report formatting

5. Related Work and Citations

Data Sources:

- Kaggle: International Football Results Dataset
- Kaggle: World Cup Database
- Kaggle: Weather Dataset

Methodological References:

- Elo, A. (1978). The Rating of Chessplayers.
- Hastie, Tibshirani & Friedman (2009). The Elements of Statistical Learning.
- Breiman, L. (2001). Random Forests.

These references informed both rating system design and predictive modeling methodology.

This project integrates multiple publicly available datasets to construct a comprehensive database for international football match prediction.

1. International Football Results (1872–2017) Source: Kaggle – International Football Results from 1872 to 2017 It forms the core dataset used to model match outcomes and generate Elo ratings.
2. World Cup Database Source: Kaggle – World Cup Database It enables structured relational analysis and demonstrates SQL database integration within the project.
3. Weather Dataset Source: Kaggle – Weather Data Weather data is incorporated to investigate whether environmental conditions influence match outcomes.
4. Elo Ratings (Computed Feature) Rather than using external rankings directly, Elo ratings were calculated using historical match outcomes. Elo ratings provide a dynamic measure of team strength over time and serve as the primary predictive feature in the modeling phase.

6. GitHub Repository

GitHub Repository:
https://github.com/Hemantkumaar/Predicting-FIFA-World-Cup-2026

7. Main Things included for the project

This project exceeds baseline expectations through:

- Multi-dataset integration
- Relational SQL database implementation
- Dynamic Elo rating engineering
- Environmental feature incorporation
- Supervised machine learning classification
- End-to-end data science workflow

This demonstrates advanced integration of database systems, statistical modeling, and predictive analytics.

Final Statement

This progress report demonstrates:

- Clear refinement of research question
- Structured preprocessing methodology
- Statistically grounded exploratory analysis
- Thoughtful modeling strategy

- Defined team responsibilities
- Clear roadmap toward completion

The project is positioned to deliver a rigorous, data-driven predictive model for international football match outcomes aligned with graduate-level data science standards.