Assignment 2: Predicting FIFA World Cup 2026 Finalists Using Machine Learning

Weight: 10% of course grade
Submission Date: 08-11-2025
Duration: 4 weeks
Teams: Individual (Full Marks) / Team of 2 (Partial marks deduction by 50%)
Deliverables: Code + design report + live/demo session + Submission in Github with detailed readme and code documentation

Learning Outcomes

- CO1: Understand key machine learning concepts and sports analytics applications
- CO2: Implement classification models on real-world datasets with preprocessing
- CO3: Evaluate model performance using multiple metrics and visualize results
- CO4: Interpret feature importance and connect data insights to domain knowledge

Assignment Tasks

Task 1: Data Collection and Preparation (20 Marks)

- Find and gather historical FIFA World Cup data: team performance, player statistics, match outcomes using available data tools, APIs, or datasets.
- **Develop a custom web scraper** to extract relevant data from at least one new sports-related website not covered by existing tools. Document the scraping code, site targeted, data fields collected, scraping logic, and any challenges faced.
- Clean the data: remove duplicates, impute or remove missing values.
- Engineer relevant features including team average age, FIFA ranking, goal differences, player experience, and team win rate.
- Submit a **data description report** explaining data sources, cleaning steps, feature rationale, and detailed documentation of the custom scraper including code snippets and usage instructions.

Expected Deliverables:

- Cleaned dataset CSV with documented columns explaining meaning.

- Written report (2-3 pages) summarizing data collection, cleaning, feature engineering, and scraper documentation.

Task 2: Model Building and Training (25 Marks)

- Implement at least two classification models (e.g., logistic regression, random forest) to predict finalists.
- Describe preprocessing steps (scaling, encoding), feature selection, and hyperparameter tuning approaches.
- Use train-test split or k-fold cross-validation to ensure fair evaluation.
- Provide fully commented code notebooks or scripts.

Expected Deliverables:

- Complete, documented code implementing models with clear explanations.
- Summary of model training, tuning, and validation approaches.

Task 3: Model Evaluation (15 Marks)

- Evaluate model performance on test data using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
- Generate confusion matrices and ROC curves for each model.
- Critically compare models discussing strengths, weaknesses, and practical implications (e.g., impact of false negatives on game strategy).

Expected Deliverables:

- Evaluation report with metric tables and visualizations (1-2 pages).
- Detailed explanation for model choice based on evaluation results.

Task 4: Feature Importance and Interpretation (10 Marks)

- Analyze and rank feature importance from models (coefficients for logistic regression, feature importance for random forest).
- Connect top features to football domain knowledge—e.g., why average age or goal difference matters for success.
- Discuss any surprising insights or data/model biases discovered.

Expected Deliverables:

- Visualizations depicting feature importance.
- Interpretation note (~1 page) linking features to football context.

Task 5: Final Prediction and Reflection (15 Marks)

- Use your best performing model with the latest qualifiers or simulated data to predict the 2026 finalists.
- Reflect on model limitations, uncertainties in sports outcomes, and ethical considerations of applying ML in sports media and fan engagement.

Task 6: Complete Application Development (15 Marks)

- Build a complete application that integrates the entire process:
    - Data extraction using tools and the custom scraper
    - Data cleaning and feature engineering
    - Model training, evaluation, and visualization
- Provide a user interface (command line or simple GUI) or script orchestration that:
    - Displays available data sources and dataset summaries
    - Lists participating teams dynamically based on the latest data
    - Supports easy data refresh and re-extraction through the scraper/tool pipeline
    - Outputs predictions and model evaluation summaries
- Document application design, dependencies, running instructions, and include screenshots if relevant.

Expected Deliverables:

- Application codebase with modular structure, clearly organized.
- README or user manual detailing installation, usage, and output interpretation.

Submission Guidelines

- All code and documents should be organized as a public/private GitHub repo and relevant access to be given to evaluators.

- Submit a detailed PDF report combining all deliverables (Tasks 1-6) with clear explanations and visualizations.
- Include documentation and screenshots for the custom web scraper and the complete application interface.
- **Video Demo:** 5-minute screencast showing system capabilities
- **Submission Deadline:** 08-11-2025, Late Penalty: 50%, maximum 1 day extension

Grading Rubric

Task	Marks	Week	Criteria Highlights
Data Preparation	20	W1	Data cleanliness, feature engineering, scraper documentation
Modeling & Training	25	W2	Correct model implementation, tuning, clear training procedure
Evaluation	15	W3	Accurate metrics, insightful visual and textual comparison
Interpretation	10	W3	Meaningful feature analysis and domain connections
Final Prediction & Reflection	15	W4	Valid prediction and deep reflection on limitations
Complete Application	15	W4	Functional integrated application with documentation
Total	**100**		

Timeline (4 Weeks)

- Week 1: Data sourcing including scraping and cleaning
- Week 2: Model coding, training, and hyperparameter tuning
- Week 3: Model evaluation, interpretation, and reporting
- Week 4: Complete application integration, final report compilation, and submission

Blooms Level of Evaluation

Task	Bloom's Taxonomy Level	Explanation
Task 1: Data Collection & Preparation	Understand, Apply, Analyze	Gather, summarize, and process real-world data; demonstrate comprehension and technique.
Task 2: Model Building & Training	Apply, Analyze, Create	Implement, tune, and code original models and solutions for predictive tasks.
Task 3: Model Evaluation	Analyze, Evaluate	Use analytical and critical reasoning to interpret, compare models, and assess results.
Task 4: Feature Importance & Interpretation	Analyze, Evaluate	Analyze internal model workings; justify findings using both data and domain knowledge.
Task 5: Final Prediction & Reflection	Evaluate, Create	Make justified predictions and reflect, synthesize insights and limitations.
Task 6: Complete Application	Apply, Create	Build an integrated solution that delivers end-to-end reproducible results.