Group 21 Proposal: Predicting FC Barcelona Match Outcomes using Machine Learning

a. Who: Proposed Client

The proposed client for this project is FC Barcelona's Data Analytics Department within the Barça Innovation Hub. This department is responsible for leveraging data-driven insights to improve the club's performance through predictive analytics.

---

b. Why: Question/Topic Being Investigated

The objective of this project is to develop a machine learning model that can be used to predict FC Barcelona's match outcomes (win, draw, loss) based on key performance indicators (KPIs) from historical matches. Understanding the factors that influence match results will help the club optimize its strategies and improve on-field decision-making.

---

c. How: Plan of Attack

1. **Data Collection**
   - Use a historical match performance dataset containing relevant KPIs (e.g., possession, shots on goal, pass accuracy)
   - The dataset will be sourced from Kaggle
2. **Data Preprocessing**
   - Clean and normalize the dataset
   - Handle missing values and outliers
   - Feature engineering to extract meaningful insights
3. **Model Selection & Training**
   - Train and compare three different machine learning models:
     - Logistic Regression (linear)
     - Random Forest Classifier (non-linear)
     - Gradient Boosting Classifier (non-linear)
4. **Evaluation & Validation**
   - Compare models based on accuracy, F1-score, and confusion matrix
   - Perform cross-validation to ensure robustness

5. **Insights & Recommendations**
   - Identify the most influential features affecting match outcomes
   - Provide tactical recommendations for game strategies based on model findings

---

d. What: Dataset, Models, Framework, Components

Dataset

- FC Barcelona Match Performance Dataset (from Kaggle)
- Includes 200 match observations with detailed KPIs and corresponding match results
- Dataset link: https://www.kaggle.com/datasets/adnanshaikh10/fc-barcelona-statistics

Machine Learning Models

- **Logistic Regression** – A simple, interpretable model for baseline predictions
- **Random Forest Classifier** – Captures complex feature interactions using an ensemble of decision trees
- **Gradient Boosting Classifier** – Enhances model accuracy by iteratively improving weak classifiers

Frameworks & Tools

- Python (scikit-learn, pandas, numpy, matplotlib, seaborn)
- Jupyter Notebook for implementation
- Model evaluation techniques: Accuracy, F1-score, confusion matrix, cross-validation

---

Summary

This project aims to leverage machine learning techniques to enhance FC Barcelona's competitive edge by predicting match results based on key performance metrics. By comparing different models, we will identify the most effective approach for data-driven decision-making in football.