ENSF 444 Project

**Predicting FC Barcelona Match Outcomes using Machine Learning**

**Introduction**

FC Barcelona is a club that heavily invests in data analytics to gain a competitive edge. The Barça Innovation Hub, for example, leverages sports data to enhance team performance and strategy ([Sports Analytics & Data Insights | Barça Innovation Hub](https://barcainnovationhub.fcbarcelona.com/area/sports-analytics/#:~:text=Hub%20barcainnovationhub,Our)). In this context, a viable machine learning problem is to analyze and predict the **outcome of FC Barcelona's matches** (win, draw, or loss) based on pre-match or in-match performance metrics. This is a **classification problem** (predicting categorical outcomes) that falls under football performance analysis. By using historical data and key performance indicators (KPIs) from past games, the club can identify patterns that lead to wins or losses, aiding coaching decisions and tactical planning. Such analysis aligns with Barcelona’s needs, as maintaining high performance and winning matches is a core challenge in both domestic and international competitions.

**Problem Description: Match Outcome Classification**

**Problem:** *Predict FC Barcelona’s match result (win, draw, loss) from performance metrics.*  
Using match-specific features (e.g. possession percentage, shots on goal, pass accuracy, etc.), we can train a classifier to predict the outcome of a game. For example, before or during a match, the model could estimate the likelihood of a win based on these indicators. This classification task directly supports performance analysis: it helps quantify how much various factors contribute to winning or not. Barcelona is known for its dominant playing style – for instance, they often enjoy very high ball possession and typically win many games ([How important is it to have the ball? | by David Sumpter - Medium](https://soccermatics.medium.com/how-important-is-it-to-have-the-ball-47f93b7760fd#:~:text=Medium%20soccermatics,won%20most%20of%20their%20games)). However, a single metric like possession is not a guaranteed predictor of victory; indeed, Barcelona sometimes “out-possesses everyone” yet still drops points ([American Soccer Analysis](https://www.americansocceranalysis.com/home/tag/FC+Barcelona#:~:text=,then%20applying%20that%20to%20tactics)). This highlights the need to consider **multiple KPIs simultaneously**, which machine learning can handle. By analyzing a combination of factors (shooting accuracy, defensive errors, opponent strength, etc.), the model can capture complex patterns that human intuition might miss. In sum, the problem addresses a key sporting challenge: understanding and improving the drivers of match outcomes.

**Dataset Selection**

To tackle this problem, the best publicly available dataset is the **FC Barcelona Match Performance Dataset** on Kaggle ([FC Barcelona Match Performance Dataset - Kaggle](https://www.kaggle.com/datasets/adnanshaikh10/fc-barcelona-statistics#:~:text=This%20dataset%20contains%20200%20match,that%20influence%20match%20outcomes)). This dataset contains **200 match observations of FC Barcelona**, focusing on key performance indicators (KPIs) that influence match outcomes ([FC Barcelona Match Performance Dataset - Kaggle](https://www.kaggle.com/datasets/adnanshaikh10/fc-barcelona-statistics#:~:text=This%20dataset%20contains%20200%20match,that%20influence%20match%20outcomes)). It includes each match’s result (win/draw/loss) along with detailed statistics for Barcelona and their opponents. Typical features cover both offensive and defensive aspects (for example, possession percentage, number of shots, shots on target, passes completed, fouls, etc.), giving a comprehensive view of performance.

**Why this dataset?** It is tailor-made for our problem – it already isolates Barcelona’s matches and relevant metrics, saving time on data gathering and cleaning. The dataset is recent and curated, ensuring data quality and relevance. Using a Barcelona-specific dataset aligns perfectly with the club’s context, as it reflects their style of play and competition history. In contrast, a generic league dataset would include many teams and require filtering and might include factors less relevant to Barcelona. By using this Kaggle dataset (which is publicly available and well-documented), we ensure our analysis directly applies to Barcelona’s matches and KPIs. *(Link to dataset:* [**FC Barcelona Match Performance Dataset on Kaggle**](https://www.kaggle.com/datasets/adnanshaikh10/fc-barcelona-statistics)*)*.

**Machine Learning Models for Comparison**

In a structured **scikit-learn** workflow, we will experiment with at least three different models to compare their performance. We include a mix of linear and non-linear algorithms to ensure we capture any complex relationships in the data. The proposed models are:

1. **Logistic Regression** – a linear classification model that will serve as a baseline. It predicts the probability of each outcome (win/draw/loss) using a linear combination of the input features. While simple, it’s fast and provides interpretable coefficients, letting us see which stats positively or negatively impact the chance of winning. However, a linear model may not capture interactions between features (e.g., the combined effect of high possession *and* high shot accuracy).
2. **Random Forest Classifier** – a non-linear ensemble model consisting of many decision trees. Random forests can capture complex, non-linear patterns in the data by averaging the predictions of multiple trees. This model can handle feature interactions automatically (for example, it might detect that a combination of certain defensive and offensive stats together is a strong win indicator). It also provides feature importance scores, helping Barcelona analysts understand which KPIs are most influential in determining outcomes.
3. **Support Vector Machine (RBF Kernel)** – a non-linear classifier that finds an optimal separating boundary between outcome classes. Using an RBF kernel allows the SVM to create curved decision boundaries in the feature space, which can fit complex relationships. SVMs are powerful for medium-sized datasets like ours (200 samples) and can handle overlapping class distributions by focusing on the most informative data points (support vectors). This helps in a scenario where wins, draws, and losses might overlap in terms of stats but are separable with the right non-linear combination of metrics.

*(All the above models are available in Python’s scikit-learn library. We ensure at least two are non-linear — Random Forest and SVM — to capture the potentially non-linear nature of football performance data.)*

We will follow a **structured machine learning workflow**: first performing data preprocessing (handling missing values, normalizing features if needed), then splitting the data into training and test sets. We’ll train each model on the training set and evaluate on the test set using appropriate metrics (accuracy and possibly F1-score or confusion matrix, since class frequencies might be imbalanced if Barcelona wins most games). We will also use cross-validation on the training data for more robust model tuning, adjusting hyperparameters (e.g., number of trees in the forest, or SVM’s kernel parameters) to avoid overfitting. This systematic approach ensures fair comparison of the models under the same conditions.

**Justification and Alignment with Club Needs**

**Why this problem?** Predicting match outcomes from performance indicators directly addresses FC Barcelona’s competitive goals. Every match’s result is crucial for winning league titles and Champions League trophies. By understanding the data behind wins and losses, the club can identify what factors to improve. For example, if the model finds that a certain threshold of shots on target is critical to winning, coaches can emphasize creating quality chances to reach that threshold. Conversely, if certain patterns tend to precede a loss (e.g., too many turnovers in midfield), the team can work to mitigate those. In essence, this ML problem turns raw match data into actionable insights, informing tactics and training. It’s a form of performance analysis that complements coaches’ expertise with data-driven evidence.

**Why the dataset?** The chosen Kaggle dataset is well-aligned with Barcelona’s needs because it encapsulates the club’s own historical performance. It is granular enough to provide insight (with many KPIs) but also focused on the club (only Barcelona’s matches), ensuring relevance. Being public and curated, it allows us to prototype a solution without requiring proprietary club data. The dataset’s KPIs mirror what Barcelona’s analysts and sports scientists already track for each game. Using this data for modeling can validate the importance of those metrics. The fact that it includes 200 matches provides a substantial sample covering multiple seasons, opponents, and competition contexts (league, cup, etc.), making the model broadly applicable to future matches.

Additionally, exploring multiple models (especially non-linear ones) is justified because football outcomes can depend on complex interdependencies. A linear approach might suggest, for instance, that “more possession is always better,” but as noted, **having extremely high possession doesn’t guarantee a win on its own** ([American Soccer Analysis](https://www.americansocceranalysis.com/home/tag/FC+Barcelona#:~:text=,then%20applying%20that%20to%20tactics)). Non-linear models can uncover that it may require a balance – e.g., high possession *plus* high shot conversion leads to wins, whereas high possession without creating chances might not. By comparing models, we ensure we choose an approach that best captures such realities. This rigorous evaluation aligns with FC Barcelona’s pursuit of excellence; just as the team tests different strategies on the pitch, we test different algorithms to find the one that best predicts and explains performance outcomes.

In summary, the proposed problem and dataset offer a practical way for FC Barcelona to apply machine learning to **performance analysis**. The classification model can serve as a decision-support tool, helping the club’s analysts and coaches focus on the factors that truly impact results. With a proper scikit-learn implementation and comparison of several models, Barcelona can confidently integrate these insights into their strategy, addressing the ongoing challenge of maintaining peak performance and winning consistently.

**Sources:**

* FC Barcelona Innovation Hub – emphasis on sports analytics for performance ([Sports Analytics & Data Insights | Barça Innovation Hub](https://barcainnovationhub.fcbarcelona.com/area/sports-analytics/#:~:text=Hub%20barcainnovationhub,Our))
* Kaggle – *FC Barcelona Match Performance Dataset* (200 matches with performance metrics) ([FC Barcelona Match Performance Dataset - Kaggle](https://www.kaggle.com/datasets/adnanshaikh10/fc-barcelona-statistics#:~:text=This%20dataset%20contains%20200%20match,that%20influence%20match%20outcomes))
* Soccermatics (David Sumpter) – observation that Barcelona often had high possession and many wins ([How important is it to have the ball? | by David Sumpter - Medium](https://soccermatics.medium.com/how-important-is-it-to-have-the-ball-47f93b7760fd#:~:text=Medium%20soccermatics,won%20most%20of%20their%20games))
* American Soccer Analysis – note on Barcelona’s possession vs. results and the need to find which aspects of possession matter