AI Project Report: Football Foul Classifier

Dataset Preprocessing Steps

Our project used the **Kaggle Football Events dataset**, which contains over **900,000 match events** such as passes, fouls, goals, corners, and cards. Each row represents one in-game action with details like `event_type`, `situation`, `side`, `bodypart`, `time`, and `location`.

Since our goal was to predict fouls, we created a target column called **is_foul**, labeled as **1** for foul-related actions (fouls, handballs, red cards, free kicks) and **0** for others. To handle missing values, we replaced them with **'Unknown'** instead of deleting them, allowing the model to learn from all available data.

We then grouped the detailed pitch locations into four meaningful zones — **Defensive, Midfield, Attack_Wide, and Attack_Central** — under a new column called `location_group`, helping the model understand where fouls often occur. Finally, we applied **Label Encoding** to convert categorical features like `location_group` and `situation` into numeric values and split the data into **80% training** and **20% testing** using **stratified sampling** to balance fouls and non-fouls.

Models Used and Why

We used three machine learning models to predict whether a football event was a foul or not:

1. Random Forest (RF1):
Our baseline model trained with the main features like `event_type2`, `side`, `location_group`, `bodypart`, `situation`, `fast_break`, and `time`.
It works by combining many small decision trees to make stable and accurate predictions.

2. Random Forest (RF2) with Engineered Features:
This version included extra features for better context:

- `match_phase` (early, mid, or late game)
- `is_attack_situation` (during corners or free kicks)
- `is_head_involved` (if the head was used)
- `is_fast_attack_zone` (fast breaks in attacking areas)
  These helped the model better connect fouls to timing, play type, and field position.

3. XGBoost:
A faster, more advanced tree-based model that builds each new tree to fix the previous one's mistakes.
It gave similar results but offered a clear **feature importance chart**, showing which factors most influenced foul predictions.

Key Findings and Interpretations

Across all models, the results were very consistent:

- **F1 Score:** ~0.88
- **Accuracy:** ~0.85
- **ROC-AUC:** ~0.93

The **F1 Score** is a balance between **precision** (how many predicted fouls were actually fouls) and **recall** (how many actual fouls were correctly detected). It's the best metric for this project since fouls are much rarer than other events.
Even though all models achieved similar scores, the **engineered features (RF2)** improved the model's understanding of football patterns without necessarily changing the numbers. This means the model became more interpretable — we could explain *why* certain fouls were predicted.

The **XGBoost feature importance plot** showed that the most influential variables were:

- `situation` (whether it was open play or a set piece),
- `event_type2` (the type of secondary event, like key passes or sending-offs),
- `location_group` (where the action happened on the pitch).

This indicates that **fouls are heavily influenced by context** — for example, they are more likely to occur during set pieces or in attacking areas where pressure is high.

In conclusion, the Foul Classifier project showed how **machine learning can be applied to sports analytics**. The models performed reliably, achieving strong F1 and ROC scores while maintaining interpretability. In the future, this approach could be extended to **real-time match analysis** or improved using **deep learning** for more complex pattern detection.