

# Project Proposal

## Title: Predicting Soccer Player Position from Attributes Using FIFA 23 Data

**Abstract:** This project aims to build a machine learning model that predicts a soccer player's position—Forward, Midfielder, Defender, or Goalkeeper—using their physical and technical attributes from the FIFA 23 dataset. This helps support smarter decisions in scouting, player development, and lineup planning.

**Introduction:** Soccer's growing fast in the U.S., and clubs are putting a lot into developing players. But figuring out a player's best position is still mostly guesswork. This project uses FIFA 23 data to train a model that predicts positions based on skill stats, helping coaches make smarter, data-driven decisions.

**Motivation:** Soccer is personal for me — it's part of my upbringing and family culture. This project connects that passion with my data science training. It matters now because:

- Soccer analytics is becoming mainstream.
- The U.S. is investing more in soccer development.
- Clubs want to make smarter choices early on.

This helps take the guesswork out of assigning positions, especially in youth development. **Previous Work:** Most projects focus on ratings or value. Few use skill stats to predict position—that's what makes this one different.

# Problem Formulation & Plan

## Problem Formulation:

- Task: Classification
- Input: Skill stats (e.g., pace, dribbling, defending)
- Output: Position group (FWD, MID, DEF, GK)
- Tools: Python, pandas, scikit-learn, XGBoost
- Metrics: Accuracy, confusion matrix
- The data will be split into training and test sets (e.g., 80/20), with 5-fold cross-validation for reliability.

**Dataset:** Source: FIFA 23 dataset (Kaggle)

- Size: ~19,000 players, 100+ features
- Preprocessing: Normalize values, group position labels, drop irrelevant fields

**Expected Accuracy:** ~80–90% using Random Forest or XGBoost

## Final Output:

A Python notebook or script that predicts position from input data, with visualizations (charts, confusion matrix, feature importance). The tool will be reusable and easy to modify for future player datasets.

## Timeline:

- Jun 20: Proposal
- Jun 27: Preprocessing
- Jul 4: Train baseline model
- Jul 11: Tune model + build visuals
- Jul 18–Aug 1: Finalize and package