Football Player Pass Completion Prediction Using Decision Trees

---

1. Introduction & Project Aim

This project aims to build a predictive model to determine a football (soccer) player's likelihood of successfully completing a pass based on their in-game attributes and physical characteristics. The core business or scouting question is: **"Can we automatically identify players with strong passing abilities using their FIFA rating data?"**

Using a Decision Tree Classifier, this model learns the patterns in player data to classify them into one of two categories:

- **0 (Not Completed):** Players with a composite passing score ≤ 70.
- **1 (Completed):** Players with a composite passing score > 70.

This approach provides a data-driven tool for talent identification, team selection, or player recruitment by focusing on a fundamental and critical skill in football.

---

2. Dataset & Preprocessing Steps

2.1. Dataset Source and Overview

- **Source:** The dataset is players_15.csv, which contains player data from the FIFA 2015 championship.
- **Size:** 15,465 players and 104 original features.
- **Features:** The dataset includes a wide range of attributes such as player demographics (age, height, weight), overall ratings (overall, potential), and detailed skill ratings (dribbling, shooting, passing, etc.).

2.2. Target Variable Creation
The original dataset did not have a direct "pass completion" label. Therefore, the target variable was engineered as follows:

1. **Identify Passing Columns:** A function was created to automatically detect relevant columns. The primary passing attribute was successfully identified and used.
2. **Define** avg_pass_score**:** Since specific short_passing and long_passing columns were not found, the general passing attribute was used as the proxy for a player's average passing ability.
3. **Set Classification Threshold:** A threshold of **70** was chosen. Players with an avg_pass_score greater than 70 were labeled as "pass completed" (1), and others were labeled as "not completed" (0). This created a clear, binary classification problem.

2.3. Feature Selection
A list of candidate features related to passing, agility, and overall ability was proposed. The final selected features, which were confirmed to exist in the dataset, are:

- dribbling
- passing
- overall
- potential
- age
- height_cm
- weight_kg

These features were chosen because they logically influence a player's passing capability (e.g., technique, vision, composure, physical build).

2.4. Data Cleaning & Splitting

- **Handling Missing Values:** For any missing values (NaN) in the selected feature columns, the median value of the respective column was used for imputation. This ensures the model can be trained without losing valuable data.
- **Train-Test Split:** The dataset was split into a training set (80%) and a testing set (20%). The stratify parameter was used to ensure both sets had the same proportion of pass completers and non-completers, preventing bias.

---

3. Model Used and Justification

Model: Decision Tree Classifier

The following configuration was used:

- DecisionTreeClassifier(max_depth=4, random_state=42)

Why a Decision Tree was Chosen:

1. **Interpretability:** This is the primary reason. Unlike "black box" models, Decision Trees produce a set of clear, human-readable rules (e.g., "IF overall rating > 73.5 AND passing > 68.5, THEN classify as a good passer"). This is invaluable for scouts and coaches who need to understand the *why* behind a prediction.
2. **Non-linearity:** The relationships between player attributes and passing success are unlikely to be simple and linear. Decision Trees can effectively model these complex, non-linear interactions.
3. **Handles Mixed Data Types:** The model works well with the numerical data (ratings, age, physical stats) used in this project.
4. **Fast Training and Prediction:** Decision Trees are computationally efficient, making them suitable for quick analysis and potential integration into larger systems.
5. max_depth=4**:** This parameter was set to limit the depth of the tree, preventing overfitting and ensuring the model remains simple and its decisions are easily explainable.

---

4. Key Findings & Interpretations

4.1. Model Performance

- The model achieved a **Test Accuracy of 1.000 (100%)**.
- **Interpretation:** While a perfect score is exceptional, it warrants scrutiny. This result strongly suggests potential **data leakage** or an overly simplistic dataset. For instance, the target variable pass_completed was derived directly from the passing feature, which is also a key input feature. The model may have simply learned to use the passing value to perfectly predict the label derived from it, rather than learning a deeper relationship from the other attributes. In a real-world scenario, the target should be independent, such as actual pass completion statistics from match data.

4.2. Class Distribution & Business Insight

- The dataset is highly imbalanced:
    - Class 0 (Not Completed): 14,413 players
    - Class 1 (Completed): 1,052 players
- **Interpretation:** Only about **6.8%** of players in the dataset are classified as high-quality passers. This immediately highlights the scarcity of this skill and aligns with the real-world observation that elite passers are a valuable commodity in football. A model that can reliably identify them is therefore highly useful.

4.3. Feature Importance & Player Profiling

Although not explicitly calculated in the provided code, the structure of the Decision Tree (with max_depth=4) inherently performs feature selection. The features used in the top splits of the tree are the most important for making predictions.

- **Expected Key Features:** We can infer that passing, overall, and dribbling are likely the most influential features. This makes logical sense:
    - passing**:** Directly measures passing technique.
    - overall**:** Represents the player's general ability and football intelligence.
    - dribbling**:** Often correlates with ball control and comfort on the ball, which are prerequisites for effective passing.
- **Actionable Insight:** The model suggests that to find a good passer, one should not only look at the passing attribute in isolation but also consider players with high overall competence and good ball control.