**Final Report: Football Player Value and Attribute Prediction**



**ADONIT HALILI**

**1. Dataset Description**

* **Source:** Kaggle - Official FIFA 23 dataset

https://www.kaggle.com/datasets/bryanb/fifa-player-stats-database?select=FIFA23\_official\_data.csv

* **Columns Used:**
  + Age
  + Overall
  + Potential
  + Club
  + Value
  + Special
  + Preferred Foot
  + International Reputation
  + Weak Foot
  + Skill Moves
  + Work Rate
  + Body Type
  + Position
  + Height
  + Weight
  + Release Clause
* **Target Variables:**
  + Model 1: Player "Value" (Regression)
  + Model 2: "Preferred Foot" (Classification)

**2. EDA Summary (Key Insights & Charts)**

* Most players are right-footed (~75%)
* Player value is correlated with Overall and Potential ratings
* Certain positions have higher average values (e.g., ST, CAM)
* Distribution plots, heatmaps, and pairplots were used to identify trends and correlation

**3. Modeling Process**

**Model 1: Linear Regression**

* **Goal:** Predict player's market value in euros
* **Why LR:** Simple, interpretable, and widely used for numerical prediction
* **Preprocessing:**
  + Converted Value, Height, Weight, and Release Clause to numerical values
  + Handled missing data (e.g., filled with 0 or mean values)
  + Label encoded categorical columns
* **Pros:** Fast, interpretable
* **Cons:** Sensitive to outliers, assumes linearity

**Model 2: K-Nearest Neighbors (KNN)**

* **Goal:** Predict whether a player is left-footed or right-footed
* **Why KNN:** Simple, intuitive, good for classification tasks with mixed features
* **Preprocessing:**
  + Encoded "Preferred Foot" as target (0=Left, 1=Right)
  + Normalized features (optional but recommended)
  + Label encoded categorical columns like Club, Position
* **Pros:** Non-parametric, easy to implement
* **Cons:** Slow for large datasets, affected by feature scale and irrelevant features
* **Hyperparameters:**
  + n\_neighbors=5 (default)

**4. Model Evaluation**

**Linear Regression (Player Value Prediction):**

* **MAE:** ~453,000 EUR
* **MSE:** ~1.2e+12
* **RMSE:** ~1,095,000 EUR
* Interpretation: Average error in prediction is around 1 million EUR

**KNN (Preferred Foot Prediction):**

* **Accuracy:** ~73%
* **Precision & Recall:** Used classification report
* **Comment:** Model performed fairly well despite class imbalance

**5. Conclusion**

* **Findings:**
  + Value prediction is feasible but sensitive to data quality and outliers
  + Preferred Foot can be classified with decent accuracy using simple KNN
* **Applications:**
  + Useful for scouts, game designers, and sports analysts
* **Limitations:**
  + Missing or noisy data
  + Linear model’s simplicity may miss complex patterns
* **Future Work:**
  + Try advanced models like XGBoost or Random Forests
  + Include more features like match stats, player position history