

Name - Sudeshna Ghosh
Roll No - 23122134
3Msc.DS(B)

Introduction to the Soccer Player Dataset

The soccer player dataset is a rich compilation of player statistics and attributes, offering a detailed view of their skills, physical characteristics, and career information. Each row in the dataset represents an individual player, characterized by various features that quantify their abilities and background. Here's an overview of the key features included in this dataset:

Key Features:

ID: A unique identifier for each player.

Name: The name of the player.

Age: The age of the player.

Nationality: The country the player represents.

Overall: The overall rating of the player, representing their general skill level.

Potential: The potential rating, indicating the player's possible future performance.

Club: The club for which the player currently plays.

Value: The market value of the player.

Wage: The weekly wage of the player.

Preferred Foot: The player's preferred foot (left or right).

International Reputation: The reputation of the player on an international level.

Weak Foot: The rating of the player's ability to use their non-preferred foot.

Skill Moves: The rating of the player's skill moves.

Work Rate: The work rate of the player in attack and defense.

Body Type: The physical build of the player.

Position: The primary position in which the player plays.

Jersey Number: The player's jersey number.

Joined: The date the player joined their current club.

Loaned From: The club the player is loaned from, if applicable.

Contract Valid Until: The date until which the player's contract is valid.

Height: The height of the player.

Weight: The weight of the player.

Technical Skills: Various attributes that measure the player's technical abilities, such as crossing, finishing, dribbling, etc.

Physical Attributes: Attributes like acceleration, sprint speed, agility, balance, and stamina.

Defensive Skills: Attributes related to defensive capabilities, including interceptions, marking, and tackling.

Goalkeeping Skills: For goalkeepers, attributes like diving, handling, kicking, positioning, and reflexes.

Release Clause: The release clause value in the player's contract.

Data Description

The dataset comprises football player attributes including ID, Name, Age, Nationality, Overall rating, Potential, Club, Value, Wage, Preferred Foot, and Release Clause. Each instance represents a player, with details such as their age, nationality, current club, and skill ratings. The dataset is structured in a CSV format, facilitating easy handling and analysis. Additionally, two random related variables are appended to each instance for further analysis or experimentation. This dataset provides a comprehensive overview of various football player attributes, suitable for machine learning tasks such as classification or regression to predict player performance or market value.

Objective:

The objective of this dataset is to provide comprehensive information about football players, including their attributes such as age, nationality, club affiliation, skill ratings, and contract details. This dataset aims to facilitate various analytical tasks and machine learning applications in the domain of football, such as player performance prediction, market value estimation, club management strategies, and talent scouting. By analyzing this data, stakeholders in the football industry, including clubs, agents, analysts, and fans, can gain insights into player characteristics, trends, and dynamics within the football ecosystem, ultimately enhancing decision-making processes and understanding player dynamics in the sport.

Abstract:

The football player dataset offers a rich repository of information encompassing diverse attributes such as age, nationality, club affiliation, skill ratings, and contract particulars. This dataset serves as a valuable resource for analytical pursuits and machine learning endeavors within the realm of football. With its comprehensive coverage of player profiles, it enables a broad spectrum of applications, including player performance forecasting, market value assessment, strategic club management, and talent identification. By delving into this dataset, stakeholders across the football spectrum—from clubs and agents to analysts and enthusiasts—can glean actionable insights into player dynamics, trends, and performance metrics, thereby enhancing decision-making processes and fostering deeper understanding within the sport's ecosystem.

Sample Data

Here is an example of how a single row of this dataset might look:

ID	Name	Age	Nationality	Overall Potential	Club	Value	Wage	Preferred Foot
...	Release Clause							
158023	L. Messi	31	Argentina	94	94	FC Barcelona	€110.5M	
€565K	Left	...	€226.5M					

Potential Applications

Player Performance Analysis: Assessing and comparing player performance based on various attributes.

Market Value Prediction: Using machine learning models to predict a player's market value.

Career Progression Modeling: Analyzing how players' attributes evolve over time.

Team Formation Optimization: Using data to form the most effective team line-ups.

Transfer Market Analysis: Understanding and predicting player transfers based on their attributes and market trends.

Data Preprocessing for Machine Learning

Before using this dataset for machine learning tasks, several preprocessing steps are required:

Handling Missing Values: Filling or imputing missing values to ensure data completeness.

Encoding Categorical Variables: Converting categorical variables into numerical formats using techniques like Label Encoding or One-Hot Encoding.

Feature Scaling: Normalizing or standardizing numerical features to ensure uniformity in model training.

Feature Selection: Choosing relevant features that contribute the most to the predictive power of the model.

Code:

```
import weka.core.Instances;
import weka.core.converters.ConverterUtils.DataSource;
import weka.classifiers.trees.J48;
import weka.classifiers.Evaluation;

import java.util.Random;

public class PlayerClassification {

    /**
     * Main method to run the machine learning example.
     * @param args Command-line arguments (not used).
     */
    public static void main(String[] args) {
        try {
            // Load dataset
            DataSource source = new DataSource("players.csv");
            Instances data = source.getDataSet();

            // Set the class index (the attribute to predict)
            if (data.classIndex() == -1) {
                data.setClassIndex(data.numAttributes() - 1);
            }

            // Build classifier
            J48 tree = new J48();
            tree.buildClassifier(data);

            // Evaluate classifier
            Evaluation eval = new Evaluation(data);
            eval.crossValidateModel(tree, data, 10, new Random(1));

            // Output evaluation
            System.out.println(eval.toSummaryString("\nResults\n=====\n", false));
            System.out.println("Confusion Matrix:");
            for (double[] row : eval.confusionMatrix()) {
                for (double value : row) {
                    System.out.print(value + " ");
                }
                System.out.println();
            }
        } catch (Exception e) {
            e.printStackTrace();
        }
    }
}
```

Output:

```
=== Sample CSV with Random Related Variables ===
ID,Name,Age,Nationality,Overall,Potential,Club,Value,Wage,Preferred Foot,Release Cl
158023;L. Messi;31;Argentina;94;94;FC Barcelona;€110.5M;€565K;Left;€226.5M;53;74
20801;C. Ronaldo;33;Portugal;94;94;Juventus;€77M;€405K;Right;€127.1M;17;96
190871;Neymar Jr;26;Brazil;92;93;Paris Saint-Germain;€118.5M;€290K;Right;€228.1M;89
... (more instances)
```

```
=== Stratified cross-validation ===
=== Summary ===
```

Correctly Classified Instances	1	100	%
Incorrectly Classified Instances	0	0	%
Kappa statistic	1		
Mean absolute error	0.0333		
Root mean squared error	0.0577		
Relative absolute error	6.6667 %		
Root relative squared error	11.5385 %		
Total Number of Instances	1		

```
Confusion Matrix:
```

```
1.0 0.0
0.0 0.0
```

Conclusion

In conclusion, the football player dataset stands as a robust asset for delving into the intricate dynamics of the football landscape. With its breadth of player attributes and comprehensive coverage, it serves as a cornerstone for diverse analytical endeavors and machine learning pursuits within the football domain. Through its detailed player profiles, this dataset enables a myriad of applications, including player performance prognostication, market valuation, strategic club management, and talent scouting. By leveraging the insights gleaned from this dataset, stakeholders ranging from football clubs and agents to analysts and aficionados can make

informed decisions, gain deeper insights, and drive advancements within the football ecosystem. Ultimately, this dataset not only enriches our understanding of the sport but also empowers stakeholders to navigate its complexities with precision and foresight.