Segmenting and Classifying the Best Strikers

***Data Analytics**: Python Project*

Presented by: Anchal Tiwari

Introduction

- Strikers play a crucial role in football, determining match outcomes.
- This project analyzes 500 strikers to identify patterns and classify top performers.
- Utilizes data analytics, machine learning, and clustering techniques.
- **Key focus**: Goals, assists, shot accuracy, dribbling success, and overall contribution.

Project Goals & Objectives

- Utilize data analytics techniques to classify and segment strikers.
- Identify key attributes that contribute to a striker’s success.
- Provide insights for coaches, scouts, and football analysts.
- Develop a predictive model to classify strikers.

visualization and statistical analysis to find key patterns.

- Perform clustering analysis to segment strikers.
- Build a machine learning model to classify striker performance. Slide **4**: Dataset Overview
- 500 football strikers analysed.

Key variables:

- Nationality, Footedness, Marital Status
- Goals Scored, Assists, Shots on Target, Shot Accuracy
- Dribbling Success, Hold-up Play, Aerial Duels Won, Consistency
- Impact on Team Performance, Big Game Performance

Data Cleaning & Preprocessing

Handling Missing Values:

- Used median imputation for numerical data.
- Used mode imputation for categorical data.

Data Type Corrections:

- Converted key performance variables to integer type.

Feature Engineering:

- Created **Total** Contribution Score by summing performance metrics.

Data Visualization

- **Pie Chart:** Distribution of Right vs. Left-footed strikers.
- **Seaborn Count plot:** Footedness distribution across nationalities.
- **Bar Chart:** Top-scoring nationalities.
- **Scatter Plot:** Relationship between hold-up play and consistency.

Clustering Analysis (K-Means)

- **Feature Selection:** Removed Striker ID and selected performance metrics.
- **Elbow Method:** Optimal cluster count = **2**.
- **Cluster Labels:**
  - Cluster 0 → **Best Strikers**
  - Cluster 1 → **Regular Strikers**

Methodology & Techniques

Data Cleaning:

- Missing values handled using SimpleImputer (median & most frequent strategy).
- Data type corrections for performance metrics.

Descriptive Analysis & Visualization:

- Summary statistics, pie charts, count plots.

Statistical Analysis:

- Correlation analysis, significance testing (Shapiro-Wilk, Levene's test).

Feature Engineering:

- Created 'Total Contribution Score' from multiple key metrics.

Clustering (KMeans):

- Segmented strikers into 'Best' and 'Regular' strikers.

Machine Learning (Logistic Regression):

- Predicted striker type based on performance attributes.

Key Insights & Findings

Q. What is the maximum goal scored by an individual striker?

Ans: 34

Q. What is the portion of Right-footed strikers within the dataset?

Ans: 53.4%

Q. Which nationality strikers have the highest average number of goals scored?

Ans: Brazil and Spain

Q. What is the average conversion rate for left-footed player?

Ans: 0.198086

Key Insights & Findings

Q. How many left footed players are from France?

Ans: 42

Q. What is the correlation co-efficient between hold up play and consistency score?

Ans: 0.147

Q. What is the p-value for the Shapiro wilk test of consistency score? Is it normally distributed?

Ans:0.451, Yes, normally distributed ($p > 0.05$)

Q. What is the p-value for the Levene's test of ANOVA analysis? Is the heteroscedasticity assumed?

Ans: 0.808, Yes, the heteroscedasticity is accepted ($p > 0.05$)

Key Insights & Findings

Q. Is there any significant correlation between strikers' Hold-up play and consistency rate?

Ans : Yes, there is a weak positive but significant correlation between strikers' Hold-up play and consistency rate.

Q. Describe the beta value of Hold-up Play you have found in your regression analysis.

Ans: The beta value should be 0.0015. It describes if the Hold-up Play scores increases by 1 score, the Consistency score increases by 0.0015 points.

Q. What is the average Total contribution score you get for the best strikers?

Ans: 123.39

Q. What is the accuracy score of your LGR model? How many regular strikers your model predicted correctly? How many best strikers your model predicted incorrectly?

Ans: 97% accuracy, 42 regular strikers model predicted correctly, 3 best strikers model predicted incorrectly.

Machine Learning Model (Logistic Regression)

Logistic Regression Model:

- Accuracy Score: 97.0%
- Creating confusion matrix.

Feature Scaling: Used StandardScaler.

Train-Test Split: 80% training, 20% testing.

Visualization:

- Confusion Matrix Heatmap
- ROC Curve (if applicable)

Statistical Analysis & Findings

Shapiro-Wilk Test for Consistency Score:

- If p-value> 0.05, the data is normally distributed.
- if p-value<= 0.05, the data is not normally distributed.

Levene's Test for Homogeneity (ANOVA Assumption):

- If p-value>0.05, the assumption of equal variance (homoscedasticity) holds**.**
- If p-value<= 0.05, the assumption is violated (heteroscedasticity).

ANOVA Test on Consistency Among Nationalities:

- If p-value>0.05, no significant difference in consistency scores among nationalities.
- If p_value <=0.05, significant difference exist among nationalities.

Business Impact & Conclusion

Actionable Insights for Coaches & Scouts:

- Identifies top-performing strikers based on key metrics.
- Assists in recruitment, team selection, and tactical planning.

Strategic Decisions:

- Helps teams optimize lineups and focus on player development.

Future Scope:

- Implement Deep Learning for better player performance prediction.
- Expand analysis to other positions (Midfielders, Defenders, Goalkeepers).

Thank You!