Basaveshwar Engineering College, Bagalkote

[An Autonomous Government Aided College, AICTE approved, Affiliated to VTU, Belagavi]

REPORT FOR MINI PROJECT (Phase – 1)

Department of Artificial Intelligence & Machine Learning. [2024-25]

SUBJECT:
MINI PROJECT (22UAI506P)

TITLE:
PLAYER ROLE CLASSIFICATION USING ML MODELS

Submitted By		
Sl No.	**Name**	**USN**
01	Keerti Nandi	2BA22AI013
02	Ibrahim Indikar	2BA22AI010
03	Prashant R H	2BA22AI025
04	Vivek S Hosur	2BA22AI052

Project Guide	Project co-ordinator	Head of Department
(Dr. Bharati M. Reshmi)	(Prof. Lakshmi P. Kolur)	(Dr. A. D. Devanagavi)

INDEX

INTRODUCTION

- Cricket is a dynamic sport where players contribute in distinct roles such as Batsman , Bowlers , All-rounders. Accurate classification of these roles is vital for team selection, strategy development, and performance analysis.
- With advancements in data analytics, machine learning (ML) offers a powerful approach to automate this classification process. By analyzing player statistics like batting averages, bowling economies, ML models can identify patterns and assign roles effectively.
- Project focuses on developing a player role classification system using ML, aiming to assist in team optimization and data-driven decision-making in cricket.
- Project also helps to make a balanced and perfect teams for the tournaments.
- Accurate classification of these roles is vital in forming balanced teams, optimizing match strategies, and identifying areas for player improvement.
- Traditionally, player roles are determined based on qualitative insights, which may be subjective. This project aims to use machine learning to automate and standardize the player role classification process by analyzing player performance metrics.
- By leveraging key statistical data such as batting averages, strike rates, economy rates, and wickets taken, this project will develop a model that can accurately classify players into Batsman, Bowlers, and All-Rounders.

OBJECTIVES

- Automate Role Identification:
  - Develop a machine learning model to automatically detect and classify player roles based on performance metrics, game statistics,and positional data.
- Improve Team Strategy:
  - Provide accurate insights into player roles to help coaches and analysts optimize team composition and strategies.
- Enhance Talent Scouting:
  - Identify and categorize players skills and strengths for better recruitment and role assignment during talent scouting.
- Facilitate Performance Analysis:
  - Enable in-depth analysis of players contributions to the game, allowing for role-based performance reviews and improvements.

LITERATURE SURVEYS

1) Cricket Player Role Prediction Using Random Forest

- Authors :- A . Kumar , S . Sharma
- Publication Year :- 2021
- Description :-

  Used Random Forest to predict cricket player roles (batsman, bowler, all-rounder) based on stats like batting average and bowling economy.Achieved 90% accuracy, highlighting the model's strength in handling high-dimensional data and feature importance ranking.

2) Application of Random Forest for Sports Performance Prediction

- Authors :- T. Singh , R . Patel
- Publication Year :- 2019
- Description :-

  Applied Random Forest to predict football player performance, successfully identifying key features like passing accuracy and tackles.Demonstrated effectiveness in handling missing data and reducing overfitting.

3) Predicting Player Roles in Team Sports Using Random Forest

- Authors :- M. Thomas , J. Collins
- Publication Year :- 2020
- Description :-

  Used Random Forest to predict basketball player roles, leveraging player statistics such as points scored and assists.The model's ability to handle categorical and numerical data led to strong performance and interpretability.

Problem Definition

In cricket, accurately categorizing players into roles such as batsman, bowlers, all-rounders is critical for team strategy and performance optimization. Existing manual or semi-automated approaches are subjective, time-intensive, and struggle with dynamic player performance.

Key Challenges

- Performance Variability:
  - Player contributions differ across match formats and conditions.
  - Performance variability in cricket player role classification is a critical challenge that can be mitigated through careful data handling, feature engineering, model regularization, and contextual awareness.

- Data Imbalance:
  - Underrepresented roles like wicketkeepers lead to biased classifications.

- Hybrid Roles:
  - Overlapping skill sets (e.g., batting all-rounders) complicate the process.
  - Hybrid roles in cricket refer to players who excel in more than one specific role, such as batters who bowl or bowlers who contribute significantly with the bat.

DATASET DESCRIPTION

- The cricket player dataset contains **17,389 records(rows)** and **16 columns**, capturing detailed information about players.

- Key features include personal details like name, date of birth, gender, as well as performance metrics such as runs scored, batting average, strike rate, best score, wickets taken, bowling average, strike rate, and best bowling figures.

- The dataset is a mix of **categorical** and **numerical data** , Numerical fields like **runs** and **wickets** offer quantifiable metrics for player performances , while categorical fields such as **gender** and **batting style** classify players into distinct groups.

- This dataset will serve as the backbone for creating machine learning models capable of accurately classifying cricket players into their respective roles based on historical performance data.

	id	firstname	lastname	fullname	dateofbirth	gender	battingstyle	bowlingstyle	runs	batting_average	strike_rate	best_score	wickets
0	2	Ahmed	Shehzad	Ahmed Shehzad	23-11-1991	m	right-hand-bat	legbreak	10180.0	47.404963	123.97	182.0	NaN
1	3	Anwar	Ali	Anwar Ali	25-11-1987	m	right-hand-bat	right-arm-fast-medium	5011.0	46.026055	104.35	144.0	269.0
2	4	Sarfraz	Ahmed	Sarfraz Ahmed	22-05-1987	m	right-hand-bat	right-arm-offbreak	NaN	NaN	NaN	NaN	244.0
3	5	Azhar	Ali	Azhar Ali	19-02-1985	m	right-hand-bat	legbreak	NaN	NaN	NaN	NaN	189.0
4	6	Fakhar	Zaman	Fakhar Zaman	10-04-1990	m	left-hand-bat	slow-left-arm-orthodox	4758.0	48.844842	144.18	144.0	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...
17384	51878	Tristan	Jungbauer	Tristan Jungbauer	01-01-2000	m	NaN	NaN	5406.0	44.529596	96.86	162.0	NaN
17385	51879	Keerti	Nandi	Keerti Nandi	06-01-2024	f	right-hand-bat	right-arm-spin	10000.0	43.734889	101.50	175.0	305.0
17386	51880	Prashant	Hadapad	Prashant Hadapad	08-06-2003	m	right-hand-bat	right-arm-fast	9000.0	52.450000	105.20	172.0	199.0
17387	51881	Ibrahim	Indikar	Ibrahim Indikar	03-07-2004	m	right-hand-bat	NaN	12000.0	35.000000	195.45	168.0	0.0
17388	51882	Vivek	Hosur	Vivek Hosur	08-11-2004	m	right-hand-bat	right-arm-spin	6000.0	47.230000	200.50	125.0	499.0

17389 rows × 16 columns

- runs:
  - Runs are the points a batsman scores by hitting the ball and running between the wickets, or by hitting the ball to the boundary (4 or 6 runs).

- batting_average:
  - Batting average is the number of runs a batsman has scored divided by the number of times they have been out.

$$\text{Batting Average} = \frac{\text{Total Runs Scored}}{\text{Number of Times Dismissed}}$$

- strike_rate:
  - Strike rate is the average number of runs scored by batsman per 100 balls faced.

$$\text{Strike Rate} = \frac{\text{Total Runs Scored}}{\text{Total Balls Faced}} \times 100$$

- best_score:
  - Best score is the highest number of runs a batsman has scored in a single innings.

- wickets:
  - It is the dismissal of a batsman, either by bowling them out, catching them, or through other forms of dismissal like run outs.

- bowling_average:
  - It is the number of runs a bowler gives up per wicket taken.

$$\text{Bowling Average} = \frac{\text{Total Runs Conceded}}{\text{Total Wickets Taken}}$$

- bowling_strike_rate:
  - Bowling strike rate is the average number of balls a bowler needs to take a wicket.

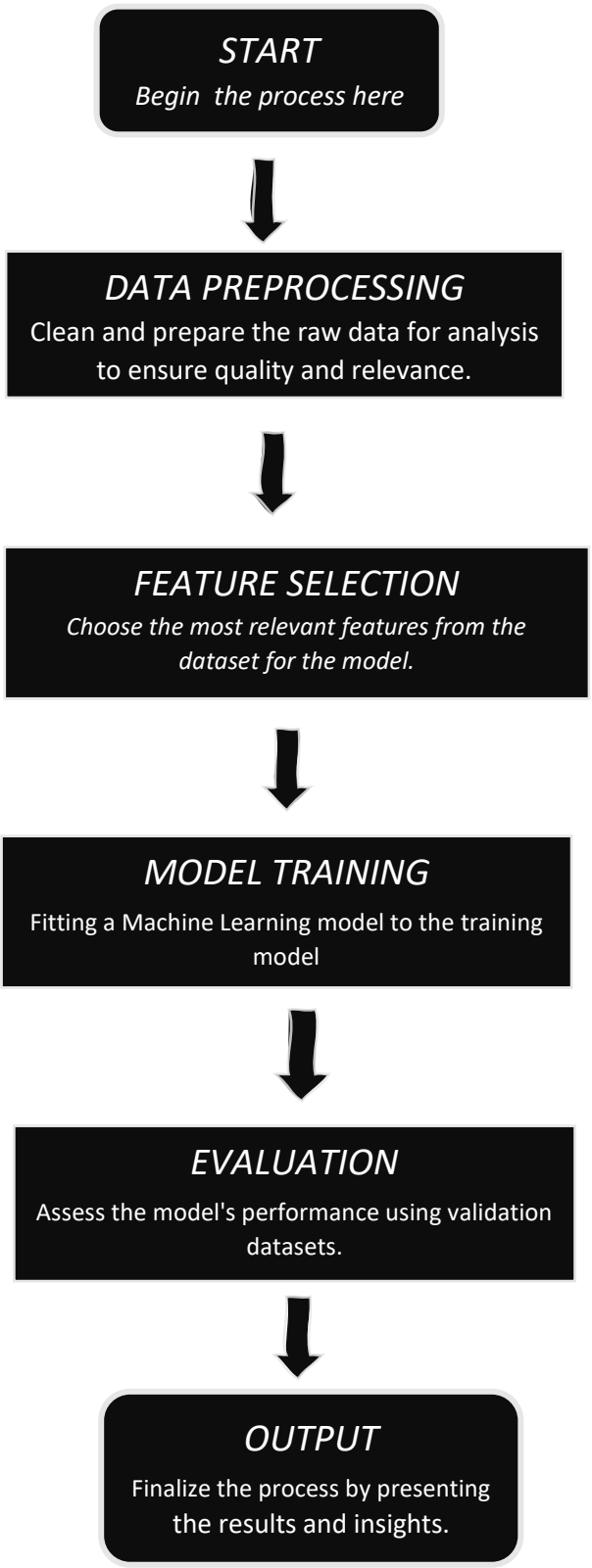
$$\text{Bowling Average} = \frac{\text{Total Balls Bowled}}{\text{Total Wickets Taken}}$$

- best_bowling:
  - Best Bowling is the best performance by a bowler in a single innings , measured by the most wickets taken in the fewest balls.

SYSTEM REQUIREMENT SPECIFICATIONS

- Hardware :
  - Processor: *Intel i5 or above*
  - RAM: *8GB or higher*
  - Disk Space: *Minimum 20GB*
- Software :
  - Python (libraries: scikit-learn, pandas, matplotlib, etc)
  - Jupyter Notebook
  - OS: Windows, MacOS, or Linux
- Functionalities :
  - Player Role Classification
  - Player Performance Analysis
  - Data Visualization
- Non – Functionalities :
  - Scalability
  - User Access Control
  - Dynamic Role Definitions

System Architecture and Proposed Methodology



- Start :
  - Import Required Libraries
  - Collect player statistics like batting average, bowling average, strike rate, wickets, etc., from reliable sources (e.g., Cricbuzz, ESPNcricinfo, Kaggle).
- Data Preprocessing :
  - Clean the data by handling missing values, duplicates, and irrelevant entries.
  - Normalize features (e.g., Min-Max Scaling) to ensure all metrics are on the same scale.
- Feature Selection :
  - Select relevant performance features like runs scored, wickets, and strike rate that influence role classification.
  - Remove redundant or low-impact features using techniques like correlation analysis.
- Model Training :
  - Train machine learning models such as Decision Tree, Random Forest, and SVM on the preprocessed data.
  - Use labelled data (player roles) to supervise learning and enable classification.
- Evaluation :
  - Evaluate models using performance metrics like Accuracy, Precision, Recall, and F1-Score.
  - Compare results to identify the best-performing model for player role classification.
- Output :
  - Classify players into roles: **Batsman**, **Bowler**, or **All-Rounder** based on model predictions.
  - Display results for validation and potential improvements.

Conclusion

- In this project, we used a Random Forest Classifier to predict and classify cricket players into distinct roles (batsman, bowler, all-rounder) based on performance metrics such as batting average, bowling economy rate, and strike rate.

- The Random Forest model demonstrated high accuracy and robustness in handling diverse and large datasets, effectively identifying the most important features contributing to player classification.

- This approach can significantly enhance team strategy and player performance analysis by providing deeper insights into player roles.

- Provides a deeper understanding of players strengths and weaknesses, aiding in performance optimization

REFERENCES

- A Survey of Machine Learning Algorithms in Sports Analytics
  - This paper explores how machine learning techniques are applied to various sports for player classification, performance analysis, and more. (https://www.sciencedirect.com/science/article/pii/S2405452620300299).

- Machine Learning Algorithms for Classification in Sports
  - A review of different machine learning algorithms used in sports analytics, which might be relevant for your project. (https://link.springer.com/chapter/10.1007/978-3-030-32038-1_18).

- Cricket Player Performance Prediction Using Machine Learning Algorithm
  - A Research paper that discusses various machine learning algorithm for cricket performance prediction. (https://www.researchgate.net/publication/332482727_Cricket_Player_Performance_Prediction_Using_Machine_Learning_Algorithms)

*****