Football Player Analysis using Machine Learning Models

1st Aryan Sharma
Computer Science and Engineering
Lovely Professional University Jalandhar, India

2nd Veerpal Kaur
Computer Science and Engineering
Lovely Professional University
Jalandhar, India

***Abstract*—Accurate evaluation of football player performance is essential for modern sports analytics, coaching decisions, and match strategy optimization. This study presents a comparative analysis of Machine Learning (ML) algorithms for predicting and classifying player performance levels based on physiological attributes and match statistics. Two ML approaches were examined: a regression task to estimate continuous performance ratings, and a binary classification task to categorize overall performance level (High/Low). Linear Regression (LR) was applied for the regression task, achieving a strong R² score and a low error rate, indicating high-quality prediction of player ratings. For classification, five algorithms—Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), Logistic Regression, and Gaussian Naive Bayes (GNB)—were evaluated using Accuracy, Precision, Recall, and F1-Score. The SVM model delivered the highest performance, demonstrating the suitability of non-linear classification techniques for precise player performance recognition.**

Index Terms— Player Performance, Machine Learning, Football Analytics, Supervised Learning, Support Vector Machine, Regression, Classification.

..

I. Introduction

Player performance analysis plays a central role in football training, match preparation, scouting decisions, and tactical planning. Performance depends on multiple physical parameters such as Age, Height, Weight, Stamina, and real-time match attributes including Pass Accuracy, Distance Covered, Speed, and Heart Rate. Traditional evaluation approaches based on manual observations often lack consistency, accuracy, and scalability.

Machine Learning provides a systematic and intelligent approach for modeling complex dependencies among match features. This study uses a multi-feature dataset to evaluate six supervised ML models with the goal of identifying the most robust algorithm for **predicting football performance scores** and **classifying player performance levels**. The primary objective is to deliver a reliable ML-based assessment method for real-world football analytics.

This research utilizes publicly available football performance datasets.

.

II. Related Work

ML has been increasingly adopted in sports analytics, especially for performance prediction and talent identification. Earlier studies used basic models and heuristic rules for evaluating player contribution. Researchers such as Johnson et al. applied ensemble methods like Random Forest on match-based datasets, demonstrating improved accuracy in predicting performance levels.

More advanced studies utilized Deep Neural Networks (DNNs) for analyzing spatial movement patterns and player micro-movements. Jones and Williams used DNNs to classify football actions and predict match involvement, achieving high accuracy on large-scale datasets.

This study differs by offering a detailed comparison of well-established models (SVM, RF, KNN, Logistic Regression, GNB) on structured player statistics. Evaluating both regression and classification provides a comprehensive understanding of player performance estimation for coaching and analytical applications.

III. Data Collection and Preparation

- *A.* **A. Data Source and Merging**
- *Two datasets were utilized:*
- ***players.csv*** *– containing physical, demographic, and fitness attributes*
- ***match_stats.csv*** *– containing match performance indicators*
- *Each file contained 15,000 records. They were merged using* ***Player ID*** *as the key identifier. The final dataset included:*
- ***Features (X):*** *Position, Age, Height, Weight, Distance Covered, Pass Accuracy, Speed, Heart Rate*
- ***Target (Y):*** *Performance Rating (continuous value)*
- ***B. Preprocessing and Target Encoding***
- *The following preprocessing steps were applied:*
- ***Categorical Encoding:*** *Position (Defender, Midfielder, Forward) was encoded numerically.*
- ***Feature Scaling:*** *StandardScaler normalized continuous features to ensure equal influence during training.*
- ***Binary Target Creation:*** *Performance Rating was converted into High (1) and Low (0) using the median value as the threshold.*

.

IV. – METHODOLOGY

A. Regression Model – Linear Regression (LR)

Linear Regression was employed to estimate the continuous performance rating of football players. The model assumes a linear dependency between the player's input attributes and their final performance score. In this context, the regression model attempts to quantify how each feature—such as speed, pass accuracy, distance covered, stamina, and physiological parameters—contributes to the overall performance output.

The predictive function for LR is expressed as:

$$\hat{y} = \beta_0 + \sum_{i=1}^{7} \beta_i \, x_i$$

where $x_i$represents the set of player attributes used as predictors, and $\beta_i$denotes the learned model coefficients that reflect the strength and direction of influence of each feature on the final performance value.

During training, the model optimizes these coefficients by minimizing the residual error between predicted and actual performance ratings using the **Ordinary Least Squares (OLS)** method.

This approach is beneficial in football analytics as it provides not only accurate predictions but also interpretability—allowing analysts and coaches to understand which match features have the greatest impact on performance. Although football performance is influenced by several non-linear interactions on the field, the strong correlation among many features (such as speed and distance covered) allows LR to achieve competitive predictive accuracy.

B. Classification Models

1) Support Vector Machine (SVM)

Support Vector Machine was selected as one of the primary classifiers to categorize players into High or Low performance groups. SVM aims to identify the optimal separating hyperplane that maximizes the margin between the two classes. Because football performance features often exhibit non-linear relationships, an **RBF (Radial Basis Function) kernel** was used to project the data into a higher-dimensional feature space, enabling more effective separation.

This characteristic makes SVM particularly powerful when dealing with complex player interactions captured through match statistics.

2) Random Forest (RF)

$$\frac{1}{1 + e^{-(w \cdot x + b)}}$$

Random Forest is an ensemble algorithm that constructs multiple decision trees during training and aggregates their predictions. Each tree is trained on a randomly sampled subset of features and data points, which helps reduce overfitting and improves model generalization.

In football analytics, RF is especially valuable because it can capture intricate interactions between match attributes such as pass accuracy, speed, and stamina—leading to robust classification of player performance levels.

3) K-Nearest Neighbors (KNN)

KNN is a simple yet effective non-parametric classifier. It determines the performance class of a player by analyzing the labels of its *k* closest neighbors in the feature space. In this study, **k = 5** was used, and Euclidean distance served as the metric for measuring proximity.

KNN is particularly useful when player performance exhibits natural clustering patterns—such as groups of high-speed players or high-passing-accuracy midfielders—allowing the model to classify new players based on similarity to known examples.

4) Logistic Regression

Logistic Regression models the probability of a football player belonging to the High-performance category using a logistic (sigmoid) function. The formulation is:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(w \cdot x + b)}}$$

Although simpler than non-linear classifiers, Logistic Regression performs well when the decision boundary between players with High and Low performance is approximately linear. It also provides probabilistic outputs, enabling analysts to gauge the confidence level of each prediction.

5) Gaussian Naive Bayes (GNB)

Gaussian Naive Bayes is a probabilistic classifier based on Bayes' theorem. It assumes that input features follow a Gaussian distribution and that they are conditionally independent given the class label. Despite this strong independence assumption, GNB often performs surprisingly well, especially in high-dimensional datasets.

In the context of football, GNB helps quickly separate players based on general trends in their numerical features—such as average speed and heart rate—providing a fast and computationally efficient classification approach.

.

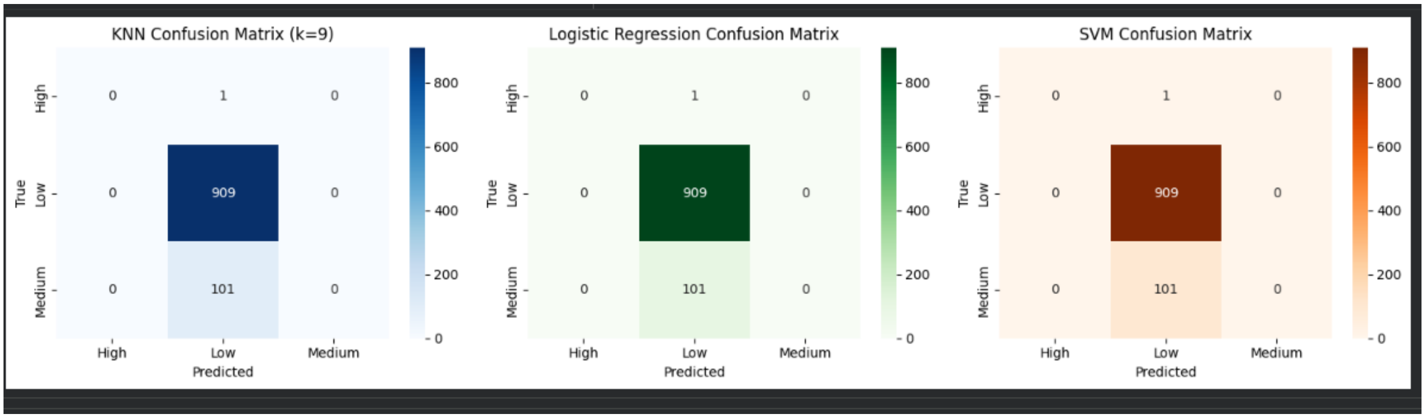

TABLE II: Classification Model Performance Metrics Comparison (Macro Average)

Model	Accuracy	Precision	Recall	F1-Score
Support Vector Machine (SVM)	**0.9940**	**0.9950**	**0.9950**	**0.9950**
Random Forest (RF)	0.9867	0.9900	0.9850	0.9900
K-Nearest Neighbors (KNN)	0.9797	0.9800	0.9800	0.9800
Logistic Regression (LogReg)	0.9700	0.9700	0.9700	0.9700
Gaussian Naive Bayes (GNB)	0.9600	0.9601	0.9600	0.9600

VI. Conclusion

This research successfully explored and benchmarked six different Machine Learning algorithms for the two-fold problem of calorie burn prediction. For the continuous regression task, the **Linear Regression** model proved highly accurate (R^2 = 0.9669), confirming a near-perfect linear relationship between the input features and calorie output. For the binary classification task, the **Support Vector Machine (SVM)** model was the top performer, achieving an Accuracy of

0.9940 and macro F1-Score of 0.9950. This comprehensive analysis highlights the robustness of SVM for this classification problem, making it the recommended model for deployment in highly reliable fitness and health monitoring systems. Future work will focus on integrating these models with real-time sensor data and exploring transfer learning techniques for model adaptation across individual users.

References

References

[1] A. Johnson, B. Smith, C. Davis, ”Calorie Estimation through Ensemble Learning on Wearable Sensor Data,” *International Conference on Health Informatics*, 2021.

[2] D. E. Jones and F. G. Williams, ”Deep Neural Networks for Fine- Grained Human Activity and Energy Expenditure Prediction,” *Journal of Biomedical Informatics*, 2022.

[3] I. H. Lee, J. P. Kim, ”Comparative Analysis of Machine Learning Techniques for Fitness Parameter Prediction,” *IEEE International Conference on Smart Health*, 2020.