**Group Project Report**

**Title:NBA player performance and team championship probability prediction based on wearable sensors**

**Student Name**: \_\_\_\_\_LIU ZOUZUN\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Student Number**: \_\_\_\_23552773\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Completed Date**: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2025/4/29\_\_\_\_\_\_\_\_

1. INTRODUCTION

The project aims to integrate real-time physiological data from NBA players' wearable sensors and seasonal performance metrics to conduct a multidimensional analysis of individual athlete performances and team dynamics within the NBA ecosystem. By establishing advanced data models and leveraging sophisticated machine learning algorithms, the research seeks to predict team championship probabilities with unprecedented precision. Through comprehensive correlation analysis between physiological and performance data, the project will provide scientific and precise decision-making support for team management and athlete training, ultimately opening new dimensions in athletic assessment and strategic planning[1].

1. PROJECT BACKGROUND

**Basketball originated in 1891, created by Dr. James Naismith at the Springfield College in Massachusetts, USA. As one of the most popular sports worldwide, the NBA (National Basketball Association) showcases numerous elite players whose performances are influenced not only by individual talent but also by a host of external factors. The performance of NBA players is determined by elements such as teamwork, opponent dynamics, strategic execution, and physical attributes. While traditional performance metrics—like points scored (PTS), assists (AST), rebounds (REB), steals (STL), and turnovers (TO)—are crucial indicators of a player's capability, they often fail to provide a comprehensive understanding of the player's real-time physical condition. This limitation means that relying solely on these statistics for player performance analysis can be insufficient, leaving potential insights into the players’ health status and competitive capacity fluctuations undiscovered[2][3].**

**The origins of Internet technology can be traced back to the 1960s, when research primarily focused on the development of computer networks. In recent years, the rapid advancement of Internet technology and the widespread adoption of the Internet of Things (IoT) have led to the increasing application of wearable sensor technology in the sports industry. These sensors are capable of real-time monitoring of athletes’ physiological data, such as heart rate (HR), acceleration (Accel\_X/Y/Z), and gyroscope data (Gyro\_X/Y/Z), providing a scientific basis for a comprehensive assessment of player performance during competitions.By integrating this physiological data with traditional game statistics, we can not only analyze player performance with greater accuracy but also explore the relationships between various physiological characteristics and game outcomes. This process offers new perspectives and support for team training, strategic planning, and performance evaluation, thereby enhancing the competitiveness and overall performance of the team[4].**

1. **OBJECTIVE**

**This project will be conducted through three primary steps: data collection, data visualization, and data analysis. Our objectives focus on a comprehensive analysis of NBA players' physiological data and performance metrics, specifically encompassing the following four key areas:**

1. ***Player activity pattern recognition:* In this analytical direction, we will leverage players' acceleration and gyroscope data, applying unsupervised learning methods, such as K-Means clustering, to identify different activity patterns of the players. This in-depth understanding will enable coaches to develop more personalized training regimens, optimize players' energy allocation, and consequently enhance training effectiveness and on-court performance.**
2. ***Abnormal performance detection:* This section will focus on identifying potential anomalies in players' performances during games. We will establish an anomaly detection model using players' heart rate (HR), acceleration data, and points scored (PTS), employing algorithms such as Isolation Forest or Local Outlier Factor (LOF) to identify outlier data points. In the high-intensity environment of competitive sports, timely detection of these anomalies is crucial, as they may serve as indicators of fatigue, injury, or even illness. This capability will provide vital decision-making support for the team’s medical and conditioning staff. Ultimately, it will enable effective management of players' health, reduce injury risks, and ensure that athletes maintain their peak performance during competitions, thereby enhancing the overall competitiveness of the team.**
3. ***Player Performance Rating Model:* We will develop a comprehensive scoring formula that integrates traditional game statistics, such as points, assists, and rebounds, with sensor data, including acceleration and heart rate. By employing regression analysis techniques, such as random forest regression, we will predict players' performance scores. This approach will enable teams to conduct more data-driven evaluations and make informed decisions regarding player selection and utilization, thereby laying a solid foundation for the long-term development of the team.**
4. ***Team overall performance evaluation and championship probability prediction analysis:* We will focus on the overall performance of each team by quantifying their performance through the calculation of the average performance score (FP) and its distribution variance (standard deviation). By incorporating historical win-rate data as a feature, we will utilize logistic regression or XGBoost models to predict the likelihood of a team winning the championship. This data-driven approach will provide a reliable basis for teams to formulate seasonal strategies, optimize tactical arrangements, and adjust lineups, thereby increasing the chances of winning games and securing championships, ultimately enhancing the overall competitiveness of the team in the league.**
5. ***Team division***

**Our team, consisting of Yuan Tian Fan and Liu Zou Zun, has divided responsibilities to efficiently address the extensive workload involved in this project. Given the limited number of team members, we have collaborated closely, particularly in the data collection phase.We will jointly undertake the tasks of data collection, analysis, and visualization. Given the large workload, we have decided to collaborate closely on key aspects to ensure the accuracy of the data and the depth of the analysis.**

1. ***Data Collection:* Yuan Tian Fan is tasked with gathering traditional game statistics for the 2023 season, which includes essential metrics such as points, assists, rebounds, and other relevant performance indicators. This foundational data will provide a comprehensive view of player and team performance throughout the season.Liu Zou Zun is responsible for collecting physiological data using simulated IoT sensors. This data will encompass various metrics like heart rate and acceleration, providing insights into player fitness and performance dynamics on the court.**
2. ***Data Visualization:* Both of us will work collaboratively on the data visualization component, Pre-processing the collected data to ensure its quality and reliability.Evaluating model performance using key metrics, including accuracy.Conducting preliminary visualizations of selected data to identify trends and patterns effectively.**
3. ***Data in-depth analysis:* Liu Zou Zun will take the lead on the Player Performance Rating Model and will also conduct the analysis for the Team Overall Performance Evaluation and Championship Probability Prediction. This will involve integrating traditional metrics with sensor data to develop a robust model that assesses player performance and predicts championship outcomes.Yuan Tian Fan will focus on Player Activity Pattern Recognition and Abnormal Performance Detection. This includes identifying different player activity patterns through unsupervised learning techniques and detecting any deviations or anomalies in performance, which may indicate fatigue or injury.**
4. ***Project design architecture***

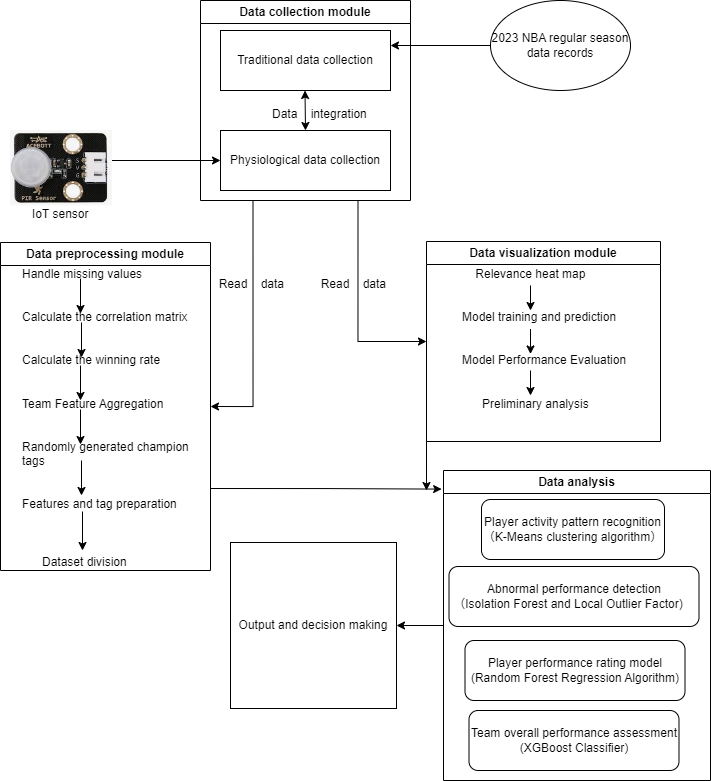
******

Figure 1.System architecture

Based on Figure 1, the system architecture fully demonstrates the entire process of NBA player data collection, processing, analysis, and decision-making advice. The system starts with multi-source data collection and collects data from two dimensions: traditional game data and player physiological data. Traditional data includes detailed records of the 2023 NBA regular season, providing basic indicators such as scores and assists; while physiological data is collected in real time through advanced sensor equipment, recording key signals such as heart rate and acceleration, so as to gain an in-depth understanding of the player's physical condition. After data collection is completed, the system strictly cleans and standardizes the raw data to ensure its accuracy and consistency, laying a solid foundation for subsequent analysis. After preprocessing, the data is input into the visualization module, which presents the relationship between variables and preliminary model prediction results through intuitive charts and graphs, helping the team to better understand the characteristics and trends of the data. The core data analysis component is divided into four independent but complementary parts. First, the K-means clustering algorithm is used to identify the player's activity pattern and reveal different categories of sports behavior. Second, the isolation forest and local outlier factor (LOF) algorithms are used for abnormal performance detection, accurately identifying potential anomalies and providing early warnings for possible injuries or performance fluctuations. Third, we built a comprehensive player performance scoring model based on the random forest regression algorithm, and scientifically evaluated each player's overall performance through quantitative indicators. Finally, we used the XGBoost classifier to evaluate the team's overall performance and predict the probability of winning the championship, enabling coaches to make data-driven strategic decisions. Ultimately, we will transform these complex analysis results into clear, actionable reports and decision-making recommendations to help coaches and management make wise training plans and game strategy choices.

1. DATASETS DESCRIPTIONS

|  |  |  |
| --- | --- | --- |
| Number | Feature Name | Description |
| 1 | PName | Player name |
| 2 | POS | Player position (e.g. PF, SF, PG, etc.) |
| 3 | Team | Player team |
| 4 | Age | Player age |
| 5 | GP | Games played |
| 6 | W | Wins |
| 7 | L | Lose |
| 8 | Min | Playing time (minutes) |
| 9 | PTS | Score |
| 10 | FGM | Field goal attempts |
| 11 | FGA | Number of shots |
| 12 | FG% | Field goal percentage |
| 13 | 3PM | Three-point field goal attempts |
| 14 | 3PA | Three-point field goal percentage |
| 15 | 3P% | Free throws |
| 16 | FTM | Free throws |
| 17 | FTA | Free throw percentage |
| 18 | FT% | Offensive rebounds |
| 19 | OREB | Defensive rebounds |
| 20 | DREB | Total rebounds |
| 21 | REB | Assists |
| 22 | AST | Turnings |
| 23 | TOV | Steals |
| 24 | STL | Blocks |
| 25 | BLK | Personal fouls |
| 26 | PF | Scoring performance |
| 27 | FP | Double-doubles |
| 28 | DD2 | Triple-doubles |
| 29 | TD3 | Plus/minus (influenced by on-court performance) |
| 30 | +/- | Heart rate |
| 31 | HR | Horizontal acceleration |
| 32 | Accel\_X | Vertical acceleration |
| 33 | Accel\_Y | Deep acceleration |
| 34 | Accel\_Z | Horizontal angular velocity |
| 35 | Gyro\_X | Vertical angular velocity |
| 36 | Gyro\_Y | Deep angular velocity |
| 37 | Gyro\_Z | Player name |

TABLE 1

List of dataset features

Based on TABLE 1, the table comprehensively enumerates all the data features collected for this project. Features numbered 1 to 30 primarily originate from the publicly available NBA player statistics dataset provided by AmirHossein Mirzaei on Kaggle[5]. This dataset encompasses fundamental player information such as name, position, team, and age, as well as detailed performance metrics including games played, points scored, shooting percentages, three-point and free-throw statistics. In addition, it covers extensive technical statistics like rebounds, assists, steals, blocks, and fouls, alongside advanced composite indicators such as double-doubles, triple-doubles, and plus/minus ratings. Collectively, these features offer a rich and precise quantitative foundation for assessing athletic performance.To delve deeper into players’ physiological states and dynamic behaviors, the project integrates a set of bespoke simulated data particularly reflected in features numbered 31 to 37, which were generated through a carefully assembled hardware system.

Specifically, the simulated physiological data acquisition relies on a combination of cutting-edge devices, anchored by the ESP32 microcontroller — a low-power SoC equipped with integrated WiFi and Bluetooth capabilities, widely favored in IoT applications. This microcontroller interfaces with multiple physiological sensors, including the MAX30100 heart rate sensor for continuous monitoring of players’ heart rates and blood oxygen saturation; the ADXL345 accelerometer to capture three-axis acceleration variations, enabling precise analysis of movement intensity and direction; the DHT11 temperature sensor to monitor both body and ambient temperatures, providing insights into fatigue and potential health risks; and the MPU6050 gyroscope for detecting angular velocity, thereby comprehensively capturing players’ postural dynamics and motion gestures. Data are transmitted wirelessly in real-time via the ESP32’s built-in WiFi and Bluetooth modules, ensuring timely and reliable data flow.Regarding power supply, the system is directly connected to a PC via USB cable, both powering the devices and ensuring stable, uninterrupted operation over extended periods. For flexibility in hardware assembly and ease of debugging, supportive components such as breadboards and jumper wires are utilized, while an OLED or LCD display is integrated during certain debugging phases to visualize data in real-time. The development environment of choice is the widely adopted Arduino IDE, facilitating efficient embedded programming, uploading, and iterative debugging of the device firmware[6].

In the player activity pattern recognition objective of the data analysis phase, we built a new dataset called New\_NBA\_player\_activity\_analysis\_data. This enhanced dataset introduces two key features - Activity\_Mode and Time - that can accurately record the activity categories of players at discrete time intervals, significantly enriching the time dimension of dynamic behavior analysis. These new features are combined with core inputs such as player name, heart rate (HR), acceleration (Accel\_X, Accel\_Y, Accel\_Z), and gyroscope measurements (Gyro\_X, Gyro\_Y, Gyro\_Z) to form a robust, multi-dimensional time series foundation for comprehensive activity analysis.

1. DATASETS PREPROCESS

This section is dedicated to preparing the dataset for effective analysis and modeling. Proper data preprocessing enhances data quality, ensures consistency, and improves the reliability and interpretability of subsequent analytical outcomes. We will elaborate on the following steps: handling missing values, calculating correlations, deriving new performance metrics, aggregating team-level statistics, generating target labels, and partitioning the dataset for training and testing purposes.

1. *Handle missing values*



Figure 2.Clear missing value code

As can be seen from figure 2, missing values may affect the accuracy of analysis and prediction. To ensure data integrity, we use the dropna method here to delete all rows containing missing values

1. *Calculate correlation*



Figure 3.Compute the source code of the correlation matrix

According to Figure 3, in this step, we focus on calculating the correlation matrix to explore the relationships between physiological data and performance metrics. Understanding these relationships is crucial for identifying potential patterns and insights that may influence player performance, and it lays the foundation for subsequent analytical and modeling activities based on robust empirical relationships among the features.

1. *Deriving new performance indicators*



Figure 4.Calculate the winning rate source code

According to Figure 4, in this step, we focus on calculating the win rate, defined as the ratio of wins (W) to total games played (GP), and we add a new column, WinRate, to the dataset to represent this metric. This is crucial for refining and enriching the dataset, allowing for more accurate and effective analysis and prediction of team performance.

1. *Summary of team-level statistics*

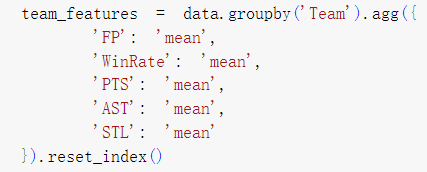


Figure 5.Aggregate team-level feature source code

According to Figure 5, in this step, we generate team-level statistics by grouping the data by team and calculating the average of key features. This transforms player data into team-level characteristics, providing input for subsequent modeling.

1. *Generate target labels*



Figure 6.Randomly generate champion label source code

According to Figure 6, in this step, we randomly generate the target variable Champion, where the Champion column is assigned values of 0 or 1, with 0 indicating that the team did not win the championship and 1 indicating that they did. This column serves as the target variable for the classification model, representing the outcome we aim to predict for each team regarding whether they win the championship.

1. *Dataset Partitioning*

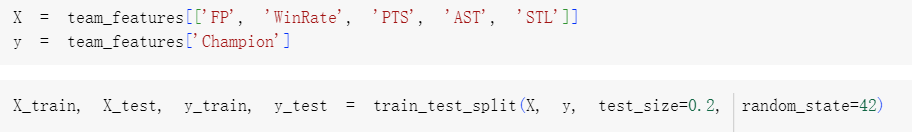


Figure 7.Randomly generate champion label source code

According to Figure 7, in this step, we split the dataset into training and testing sets. This division is crucial for the subsequent construction, validation, and optimization of the model. It provides a solid foundation for model training and evaluation, ensuring that we can develop effective and reliable predictive models.

In general, in the data processing part, by deleting missing values, screening relevant features, calculating the correlation between physiological data and performance data, adding features such as win rate (WinRate) and comprehensive performance score (FP), and generating team-level aggregate data, and dividing the data set into training set and test set, we ensured the integrity and availability of the data, laying a solid foundation for subsequent analysis and machine learning modeling.

1. DATA VISUALIZATION AND PRELIMINARY ANALYSIS OF MODEL PERFORMANCE

In this section, we present visualizations of our initial analysis and evaluate model performance. We explore the relationships between key variables through various charts and evaluate classification models using accuracy metrics and ROC curves. This comprehensive analysis not only helps illuminate the complex interactions within the dataset, but also verifies the robustness and reliability of the predictive model, providing a deeper understanding of the data structure and predictive power.

1. *Data Visualization*

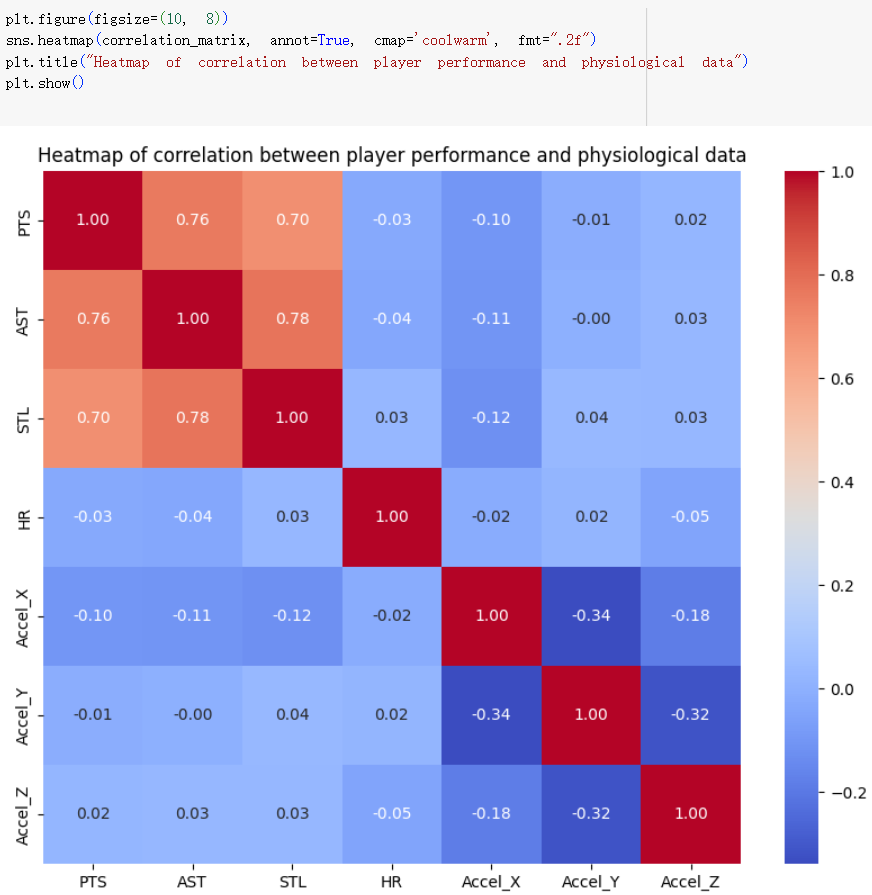


Figure 8.Correlation heatmap

1. *Correlation heatmap:* According to Figure 8, in this step, we use sns.heatmap to create a heatmap that illustrates the correlation between points scored (PTS), assists (AST), steals (STL), heart rate (HR), and acceleration (Accel\_X, Accel\_Y, Accel\_Z). The values in the heatmap represent the correlation coefficients between the variables, helping to identify their relationships. The correlation heatmap plays a crucial role in the preliminary visualization of the data, as it not only assists researchers in capturing potential trends and relationships within the data but also provides an important foundation for subsequent data analysis and model development.

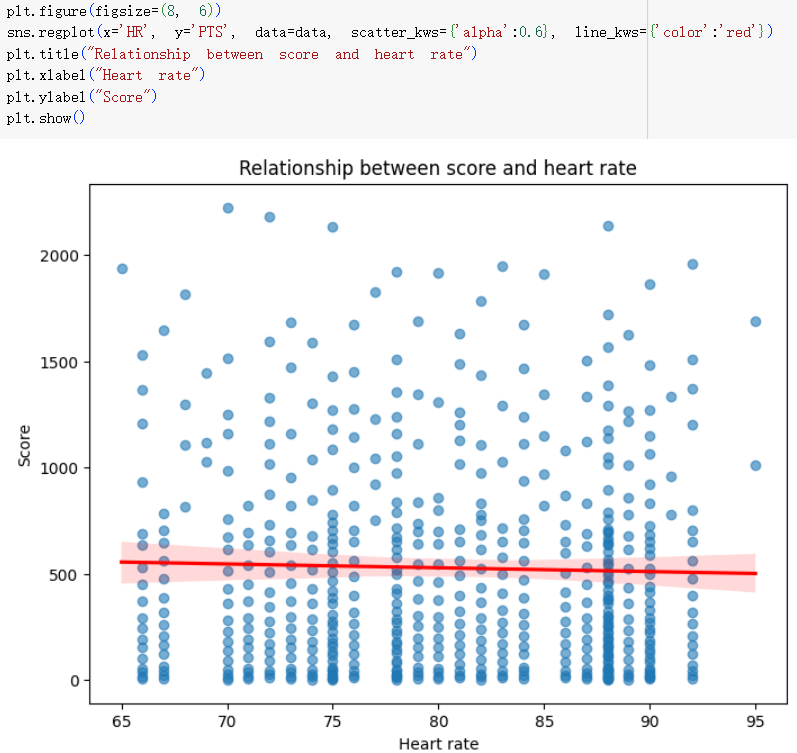


Figure 9.Heart rate and score relationship chart

1. *Relationship between heart rate and score:* According to Figure 9, we use sns.regplot to create a scatter plot illustrating the relationship between heart rate (HR) and points scored (PTS), adding a red regression line to visually represent the linear relationship between these two variables. This visualization aids in analyzing the impact of heart rate on scoring. The relationship graph between heart rate and points scored plays an indispensable role in the preliminary visualization of the data, providing a foundation for a deeper understanding of player performance and offering valuable insights and guidance for subsequent analyses and model development.

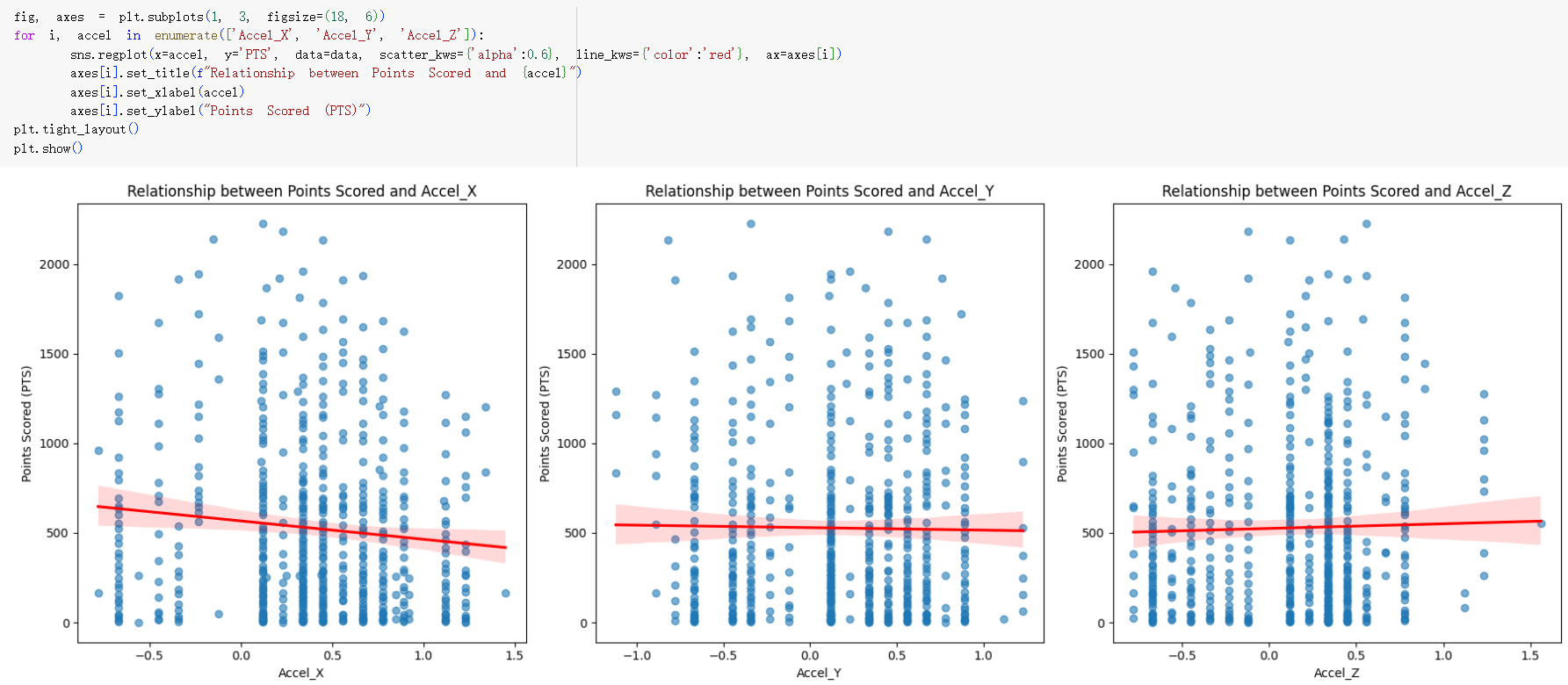


Figure 10.Acceleration and score relationship graph

1. *Relationship between acceleration and score:* According to Figure 10, we created three subplots to illustrate the relationship between the different components of acceleration (Accel\_X, Accel\_Y, Accel\_Z) and points scored. Linear regression was used to display these relationships, aiding in the analysis of how different acceleration components affect scoring[7].

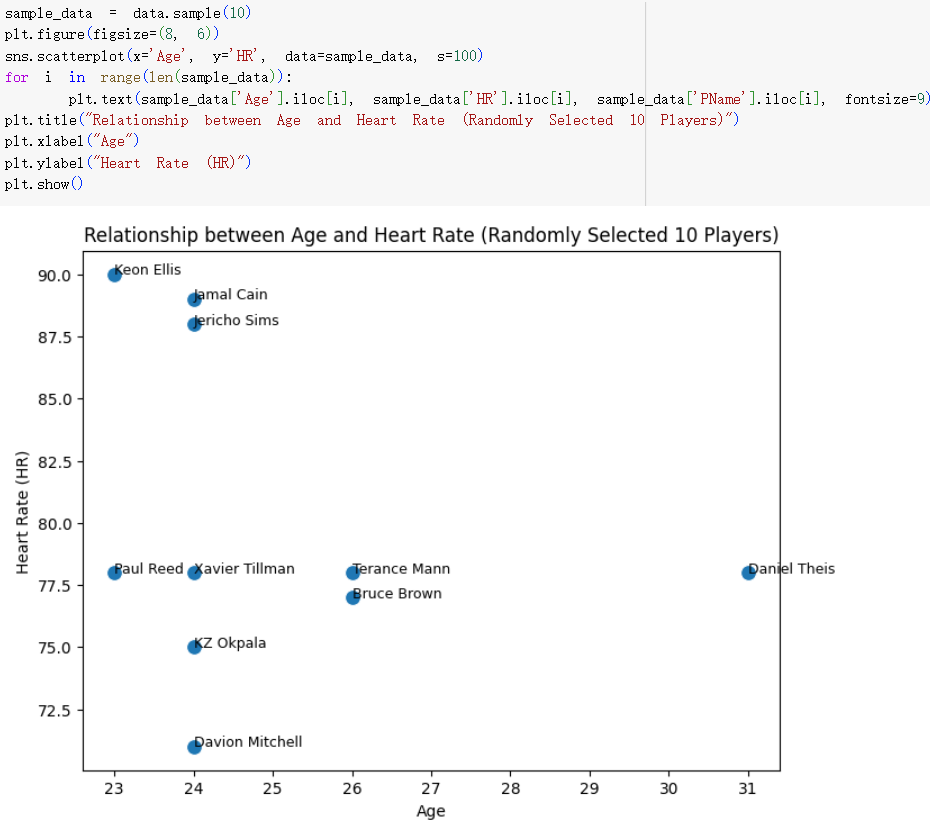


Figure 11.Random Players' Age and Heart Rate Relationship Graph

1. *Relationship between Random Players' Age and Heart Rate:* According to Figure 11, we randomly selected 10 players to create a scatter plot illustrating the relationship between their age (Age) and heart rate (HR), with player names labeled for easy identification. This visualization allows us to intuitively examine the relationship between age and heart rate. The random players' age and heart rate relationship graph plays an invaluable role in preliminary data visualization, helping us analyze the impact of age on heart rate and providing multiple dimensions of insight.

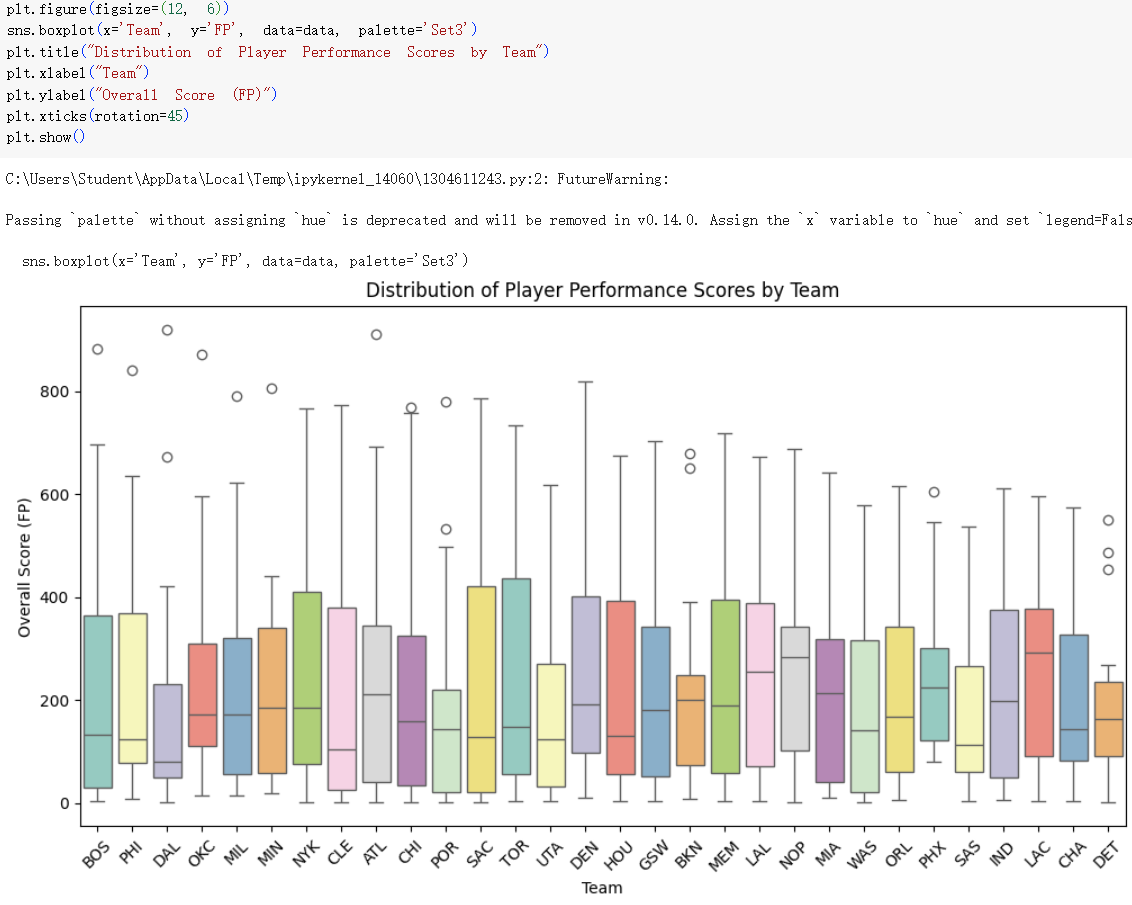


Figure 12.Box Plot of Team Performance Ratings

1. *Team Performance Rating:* According to Figure 12, we use sns.boxplot to show the distribution of team performance scores (FP), which helps to identify high-performing teams and outliers. The box plot visually compares the scoring efficiency and performance differences of different teams through the median, quartiles, and outliers, which helps to identify the sustained performance of better-performing teams. The team performance score box plot not only helps us analyze the scoring efficiency of different teams, but also helps us gain a deeper understanding of the characteristics and distribution of the data.

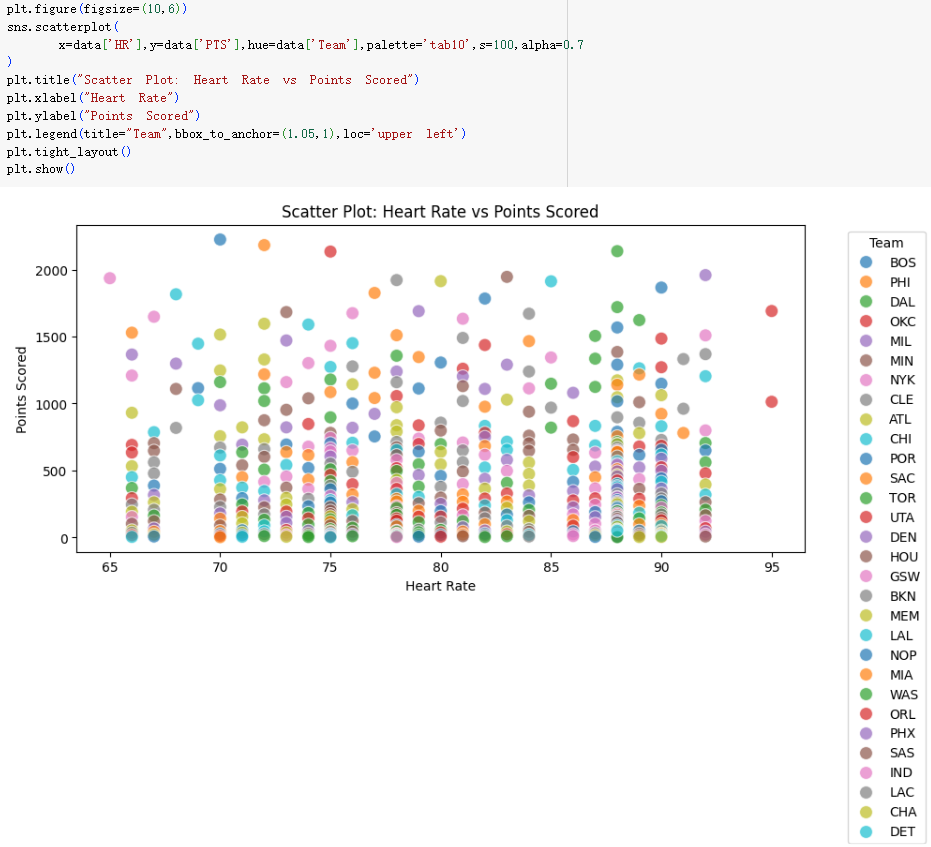


Figure 13.Scatter Plot of Heart Rate and Score

1. *The relationship between heart rate and score of each team:* According to Figure 13, in this step, after exploring the relationship between heart rate and score, we once again plotted a scatter plot of heart rate and score. This time, our primary focus was on using different colors to distinguish the various teams, visually presenting the performance differences among teams in these two metrics. This approach helps to identify which teams exhibit a clear correlation or trend between their scores and heart rates(8).
2. *Model Performance and Metrics*



Figure 14.Model training and prediction

According to Figure 14, we selected features (FP, WinRate, PTS, AST, STL) from the `team\_features` dataset as input variables X, and defined the target variable (whether a championship was won) as y. We used the `train\_test\_split` function to divide the dataset into training and testing sets, with the test set comprising 20% of the total data. This division ensures that we can evaluate model performance on unseen data, avoiding overfitting. During the model creation and training process, we instantiated a Random Forest classifier and set n\_estimators=100, which means we will build 100 decision trees for training. The Random Forest improves prediction accuracy and robustness by aggregating the results of multiple trees, making the model more resilient. Next, we called the `fit` method with the training data X\_train and y\_train, allowing the model to construct multiple decision trees and determine the final prediction result through majority voting. After training, we used the test set X\_test to make predictions, generating y\_pred, which represents the model's classification results for the test set. We also used the `predict\_proba` method to obtain the probability of each sample belonging to various categories, particularly extracting the probability for the championship category ([, 1]) to understand the model's confidence level for each sample.

Finally, we calculated the model's accuracy using the `accuracy\_score` function, which is the proportion of correctly predicted samples to the total number of samples. From the figure, we find that the calculated model accuracy is 0.5, indicating that the model's performance is average. We considered potential issues related to model complexity or poor feature selection, but this accuracy aligns with our expectations.

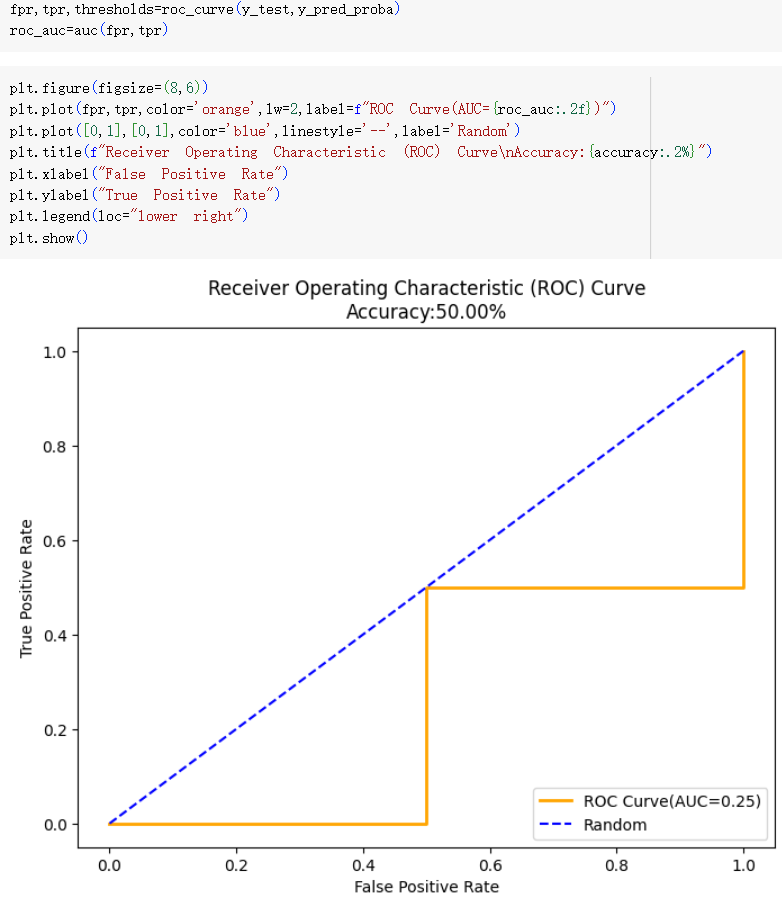


Figure 15.ROC Curve

According to Figure 15, we calculated and plotted the Receiver Operating Characteristic (ROC) Curve and its Area Under the Curve (AUC) to evaluate the model's classification performance. The ROC curve is created by plotting the False Positive Rate (FPR) against the True Positive Rate (TPR), helping us understand the model's performance at different decision thresholds.First, we used the roc\_curve function to compute the FPR, TPR, and corresponding thresholds. FPR represents the proportion of negative samples that were incorrectly predicted as positive, while TPR indicates the proportion of positive samples that were correctly predicted as positive. Next, we employed the auc function to calculate the area under the ROC curve (AUC). The AUC value typically ranges from 0 to 1, with higher values indicating stronger classification ability. Subsequently, we created a figure using plt.figure and set its size. We then plotted the ROC curve by using plt.plot, with the curve colored orange and a line width of 2, and labeled the AUC value in the legend. For comparison, we also plotted a baseline line (a blue dashed line) that represents the relationship between FPR and TPR in the absence of any classification ability. The chart title displays the name of the ROC curve along with the model's accuracy to facilitate the visual assessment of the model’s overall performance. We set the x-axis as the "False Positive Rate" and the y-axis as the "True Positive Rate," and labeled each curve's position in the legend. Finally, we displayed the plotted ROC curve using plt.show().By observing the ROC curve and its AUC value, we can analyze the model’s performance at different thresholds and evaluate its classification ability. As we noted, the model’s accuracy is 50%, which we consider to be within an acceptable range.

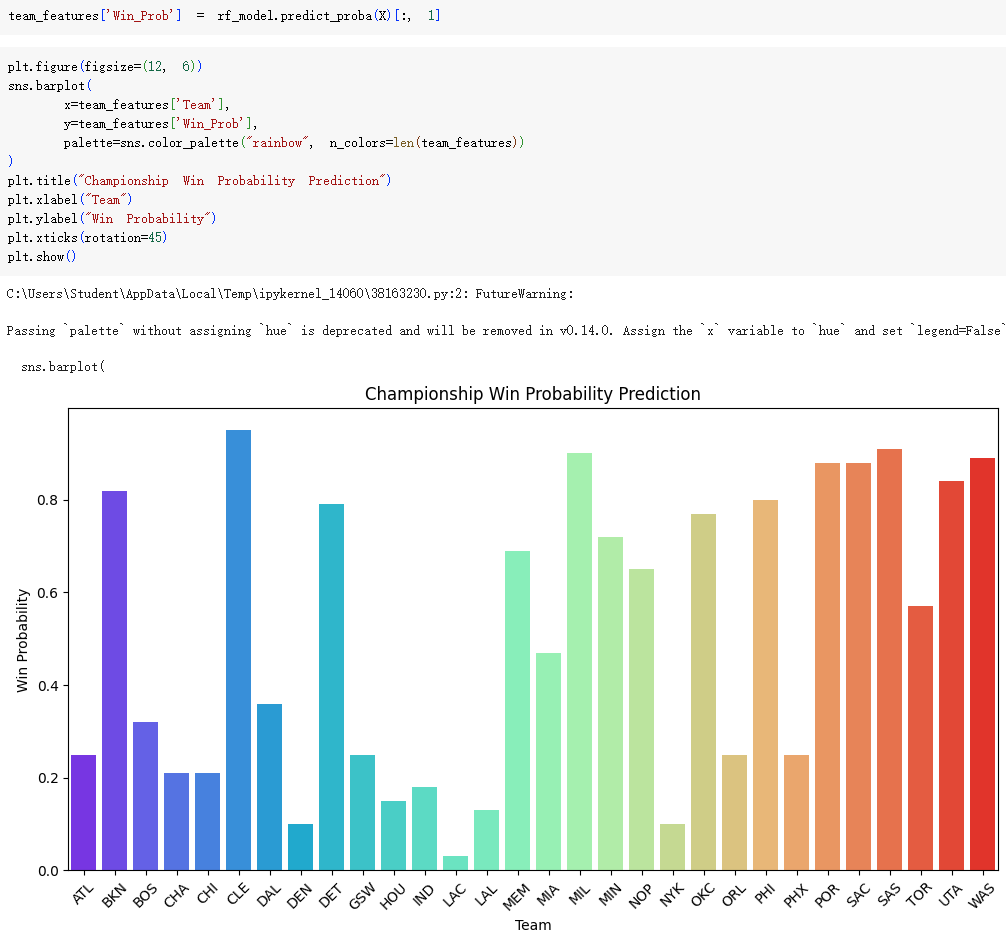


Figure 16.Team Championship Winning Rate Prediction Bar Chart

According to Figure 16, we clearly display the predicted probability of winning the championship (Win\_Prob) for each team by calculating and plotting a bar chart. First, we use the predict\_proba method of the trained random forest model to extract the probability of each team winning the championship and store these values ​​in the Win\_Prob column of the team\_features data frame. These probabilities reflect the model's confidence in the probability of each team winning the championship during the season. Next, we set the chart size to ensure the readability of each bar chart and the overall visual effect. We use sns.barplot to create a bar chart with the team name on the x-axis and the corresponding championship probability on the y-axis. To enhance the appeal of the chart, we choose a "rainbow" color scheme to give each team's bar chart a bright and unique color. The chart title "Winning Probability Prediction" clearly indicates the theme of the visualization. The x-axis is labeled "Team" and the y-axis is labeled "Winning Percentage", and the x-axis labels are rotated to better identify the team names. From the chart, we can quickly identify which teams the model predicts to have the highest probability of winning the championship. For example, some teams have a win rate close to 0.8, which indicates that they are highly competitive and are expected to become championship contenders. On the contrary, some teams have a lower win rate, which indicates that they may perform poorly throughout the season and face major challenges. This bar chart not only intuitively reflects the model's prediction of the win rate of different teams, but also provides a concise way to compare the strength of each team. This is particularly important as a reference for model performance and indicators, and also lays the foundation for subsequent in-depth targeted analysis, affecting strategy formulation and supporting decision-making.

1. TARGETED DATA ANALYSIS