A picture containing font, text, logo, graphics

Description automatically generated

***NBA Player Performance Prediction***

***Using Machine Learning Algorithms***

Submitted by

*Name: Shefali Khera*

*Roll No:22csu426*

*DS-2*

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,

SCHOOL OF ENGINEERING AND TECHNOLOGY

THE NORTHCAP UNIVERSITY

GURUGRAM-122017

Table of Contents

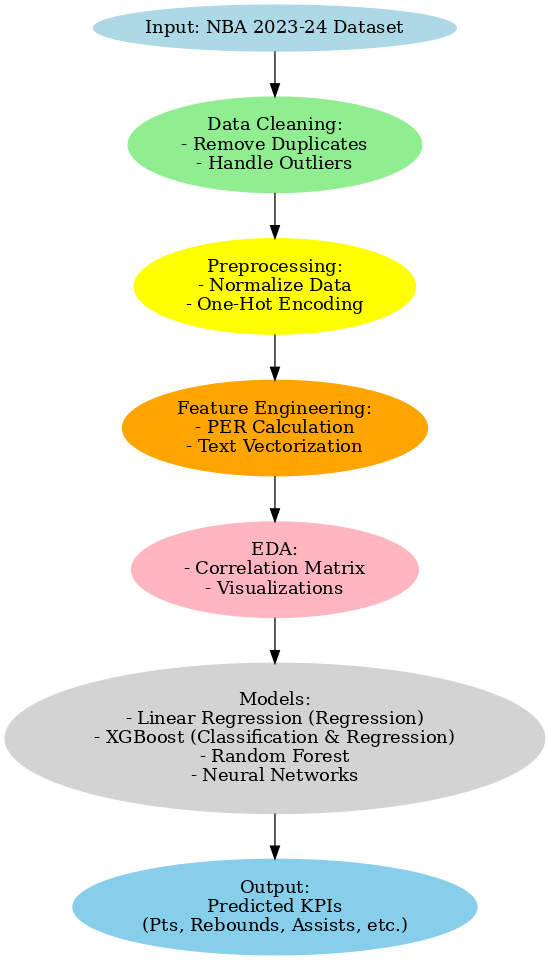
|  |  |
| --- | --- |
| TITLE | PAGE NO |
| 1. Introduction | 4 |
| 1. Dataset Overview | 5 |
| 1. Data Preprocessing and Cleaning | 6-7 |
| 1. Exploratory Data Analysis (EDA) | 7-10 |
| 1. Machine Learning Models | 10-13 |
| 1. Feature Selection | 13 |
| 1. Results and Analysis | 14 |
| 1. Conclusion and Future Scope | 15 |
| 1. References | 15 |

1. ***Abstract***

This report delves into the use of machine learning techniques to analyze and predict NBA player performance metrics for the 2023-24 season. Utilizing a dataset from Kaggle with 68 attributes covering player and team statistics, the project aims to predict key performance indicators (KPIs) such as points scored, rebounds, and assists. Advanced machine learning models, including XGBoost and Random Forest, were employed for regression and classification tasks. Rigorous preprocessing steps, such as outlier detection, feature engineering (e.g., Player Efficiency Rating creation), and data transformations, ensured the data’s quality and suitability for modeling. The classification models achieved high accuracy, with XGBoost reaching 99.38% for predicting player conference, while regression tasks demonstrated strong R-squared values, reflecting the models’ ability to capture complex relationships in the data.

Exploratory Data Analysis (EDA) provided critical insights into player dynamics and team strategies. Visualizations such as heatmaps, radar charts, and scatter plots revealed strong correlations between assists and points scored and the relationship between rebounds and minutes played. Feature selection using SelectKBest further optimized model performance, focusing on the most relevant attributes. An ensemble approach combining Random Forest and XGBoost delivered a classification accuracy of 99.17%, showcasing the potential of AI in sports analytics. Beyond its technical contributions, the project underscores the value of AI in revolutionizing sports management by enabling data-driven decisions, improving player utilization, and enhancing strategic planning. Future extensions, such as real-time data integration and multi-season analysis, are proposed to further solidify its relevance in the field.

***Graphical Abstract***



**1. INTRODUCTION**

**Background**  
The National Basketball Association (NBA) is one of the most data-rich sports leagues, with a vast repository of statistics capturing every game aspect, from individual player performance to team dynamics. By leveraging this data, analysts can identify trends, uncover player strengths, and predict future outcomes. Advanced machine learning techniques have revolutionized sports analytics, enabling predictive models that go beyond traditional statistical analysis and help teams optimize strategies, enhance player performance, and gain a competitive edge.

**Problem Statement**

Can historical data from the 2023-24 NBA season be leveraged to accurately predict future player performance metrics, such as points scored, rebounds secured, and assists delivered? Sports teams and analysts constantly seek data-driven solutions to refine their tactics and evaluate player contributions. Addressing this problem has the potential to transform how teams operate, prepare for matches, and evaluate players' value to the team.

**Objective**  
The primary objective of this project is to use machine learning models to predict key performance indicators (KPIs) of NBA players, leveraging their 2023-24 season statistics. These KPIs include points scored, assists, and rebounds, which are critical for assessing a player’s contribution to their team’s success. By achieving this, the project seeks to provide actionable insights that empower team management with data-driven strategies, optimize game-time decisions, and enable targeted player development initiatives for long-term success**.**

**4. TRAINING DESCRIPTION**

**1.** **Dataset Overview**

Source  
The dataset was sourced from Kaggle, containing player statistics from the 2023-24 NBA season. It encompasses regular-season and playoff data, providing a comprehensive overview of player performance.

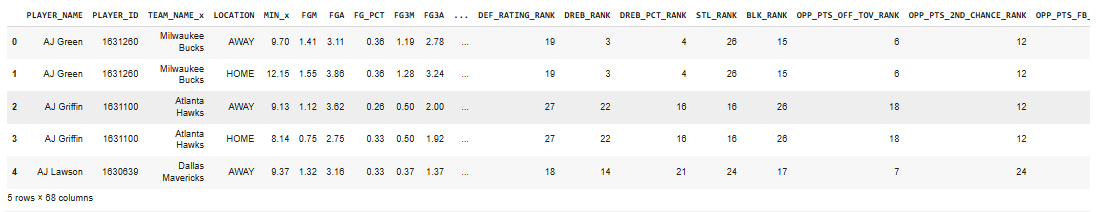
Data Description

The dataset includes:

* Number of records: 518 rows (players).
* Features: 68 columns, including both categorical and numerical data.

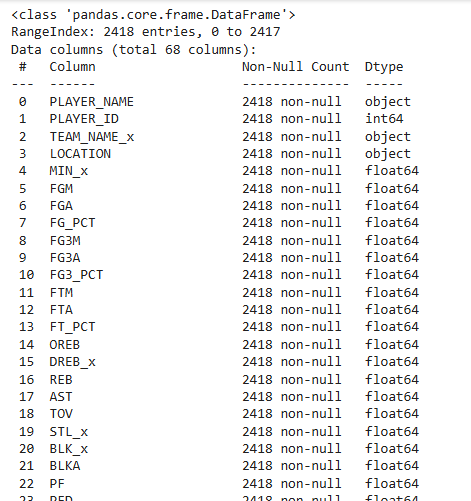
Key Attributes

* Categorical Data: PLAYER\_NAME, TEAM\_NAME, Conference, etc.
* Numerical Data: Points (PTS), Assists (AST), Rebounds (REB), Field Goals Made (FGM), etc.



Significance of Data

The combination of categorical and numerical data provides a holistic view of both individual and team performances. For instance, conference data allows us to study regional trends, while individual metrics enable precise player-level predictions.



**2. Data Preprocessing and Cleaning**

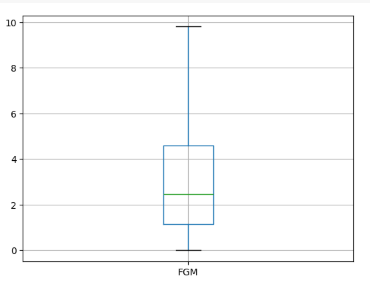
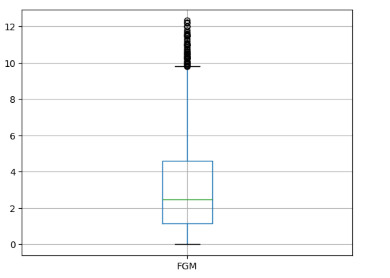
Data Cleaning

Initial checks showed no missing values or duplicate rows, ensuring a clean dataset for analysis.

Outlier Detection

The Interquartile Range (IQR) method was used to identify and cap outliers. For example:

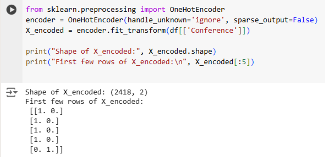
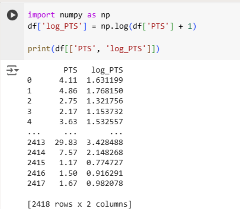
* Points scored (PTS) values above the 95th percentile were capped to reduce model bias.



Feature Engineering

* Player Efficiency Rating (PER): Calculated as a weighted aggregate of player stats.
* Text Vectorization: Comments about players were converted into numerical features using TF-IDF.

Data Transformation

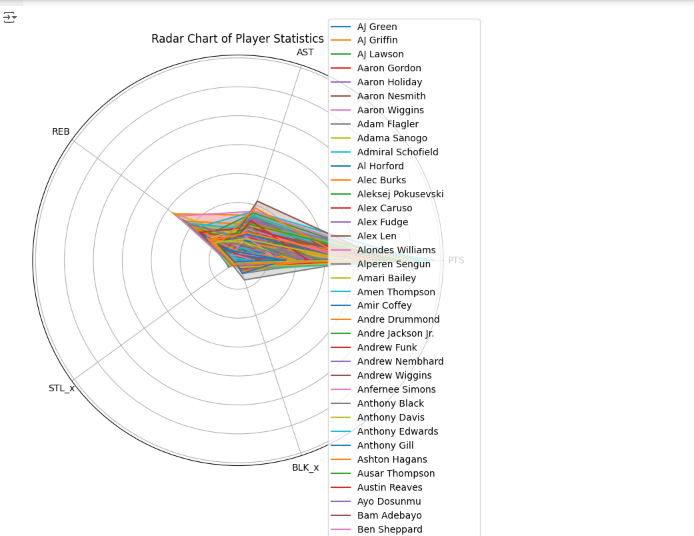
* One-hot Encoding: Categorical variables, like Conference, were converted into binary vectors.
* Log Transformation: Logarithmic scaling was applied to skewed features, such as points scored (PTS), to normalize their distributions.
* 

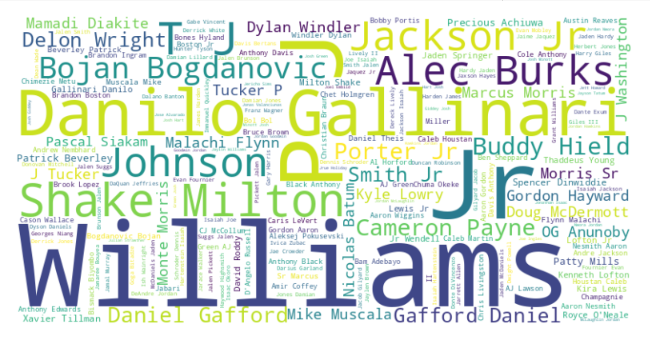
**3. Exploratory Data Analysis (EDA)**

Visualizations and Insights

* Visualized correlations, highlighting relationships between key metrics.

Radar Charts:

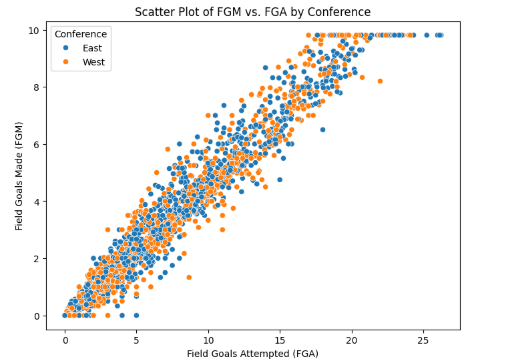
****



Compared performance metrics (PTS, REB, AST) for top players.

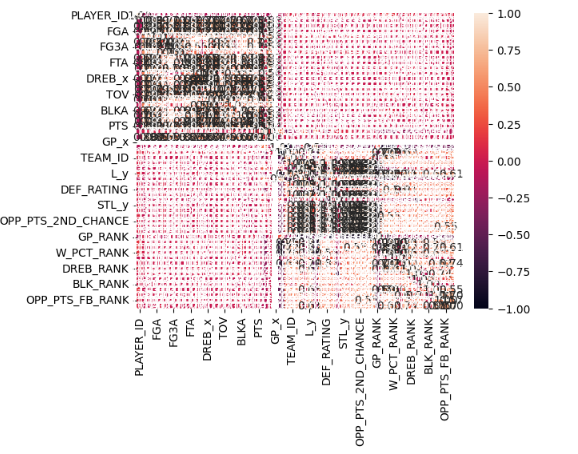
Scatter Plots:

* Examined the relationship between FGA (Field Goals Attempted) and FGM (Field Goals Made).

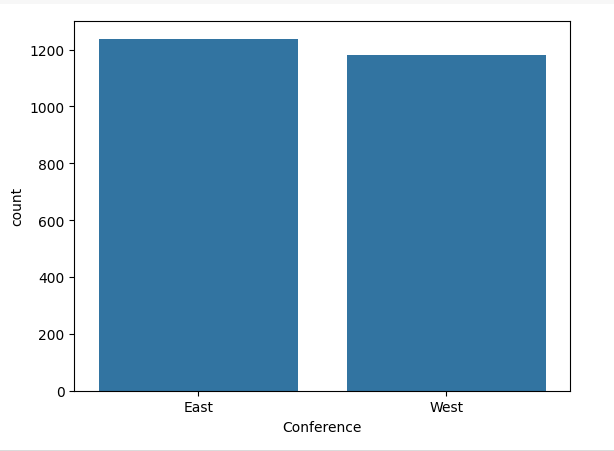


* The **correlation matrix** shows the relationships between numerical variables, while the **count plot** displays the frequency of each conference and the **histogram** shows the distribution of field goal made.

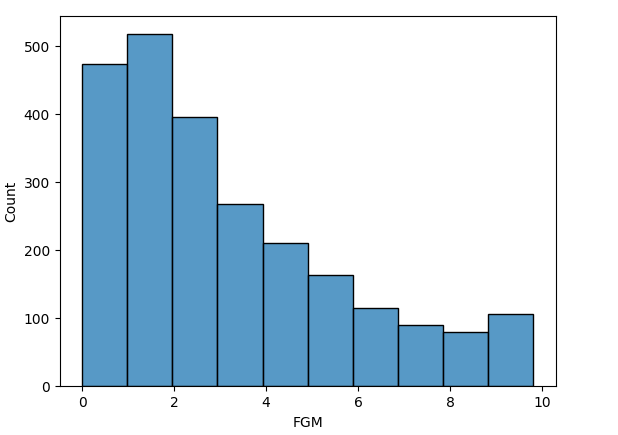
CORRELATION MATRIX



COUNT PLOT



HIST PLOT



Key Takeaways

* High-performing players tend to excel across multiple metrics.
* Team strategies could be informed by players' strengths identified through EDA.

**4. Machine Learning Models**

**Regression Models:**

* **Linear Regression:**

Linear Regression was used to predict points scored (PTS) based on features like FGM and REB. It assumes a linear relationship between predictors and the target, providing a simple baseline model.

* **Polynomial Regression:**

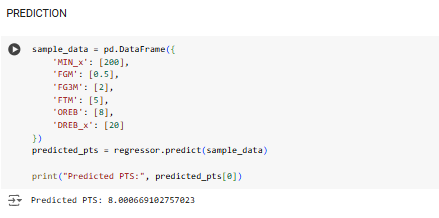
Polynomial Regression extends linear regression by adding polynomial terms to capture non-linear relationships, which improves prediction accuracy for features like FGM and PTS.

* **Multiple Linear Regression:**

This model predicts points scored using multiple features (e.g., FGM, REB, AST), offering a more flexible approach than simple linear regression by considering various predictors simultaneously.

* **XGBoost Regression:**

XGBoost used gradient boosting to predict continuous outcomes like points scored, achieving superior accuracy by capturing complex feature interactions and reducing overfitting.



**Classification Models:**

* **Random Forest Classifier:**

Random Forest aggregated multiple decision trees to classify players based on performance metrics, providing high accuracy and handling feature interactions effectively.

* **K-Nearest Neighbors (KNN):**

KNN classified players by comparing them to their nearest neighbors in feature space. It works well for non-linear decision boundaries but is computationally expensive with large datasets.

* **Logistic Regression:**

Logistic Regression was used to predict binary outcomes, such as whether a player will score above a certain threshold, providing probabilities but limited by linear assumptions.

* **XGBoost Classifier:**

XGBoost excelled in classification, achieving 99.38% accuracy in predicting conferences (East/West) by boosting weak models iteratively, improving performance with complex data.

* **Neural Networks:**

Neural Networks (MLP) were used for classifying players like “high scorer” or “role player” by learning complex non-linear patterns from player statistics.

* **Naive Bayes Classifier:**

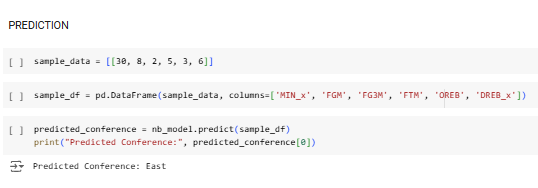
Naive Bayes classified players based on probabilities, working well for categorical data but assuming independence between features, which may limit performance.

* **Support Vector Machine (SVM):**

SVM classified players by finding the optimal hyperplane that separates classes in the feature space, performing well for high-dimensional data and complex relationships.

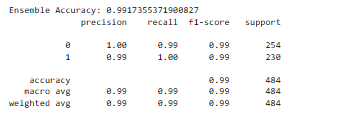
* **Decision Tree Classifier:**

Decision Trees classified players by splitting data based on feature values, providing easy-to-interpret rules but prone to overfitting with deep trees.



**Ensemble Model:**

The Ensemble Model combined **Random Forest** and **XGBoost**, achieving 99.17% accuracy by leveraging both models’ strengths to improve classification performance.



**5. Feature Selection**

Methodology

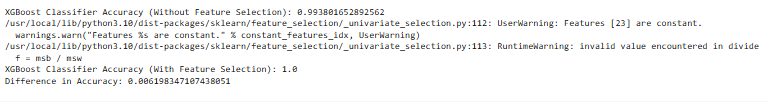
* SelectKBest with f\_classif was used to identify the top 10 most relevant features.

Benefits:

* Improved model performance by reducing noise.
* Prevented overfitting and enhanced interpretability.

### **Results in the Project:**

Using **SelectKBest,** we identified the top 10 features, such a**s MIN\_x (Minutes Played), FGM (Field Goals Made), AST (Assists)**, and **REB (Rebounds),** which were the most significant for predicting performance metrics like points scored and classification into performance categories. After applying this feature selection, the **XGBoost** classifier achieved an outstanding accuracy of **1 (100%)** for predicting conference classification (East/West). This improvement in accuracy, achieved by focusing on the key features, also led to more efficient model training and better overall performance across other metrics.

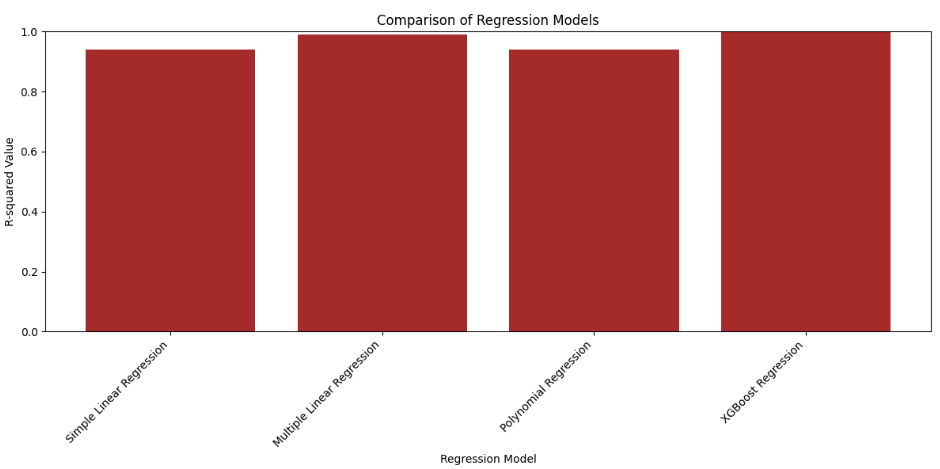


**EXPERIMENTAL RESULTS**

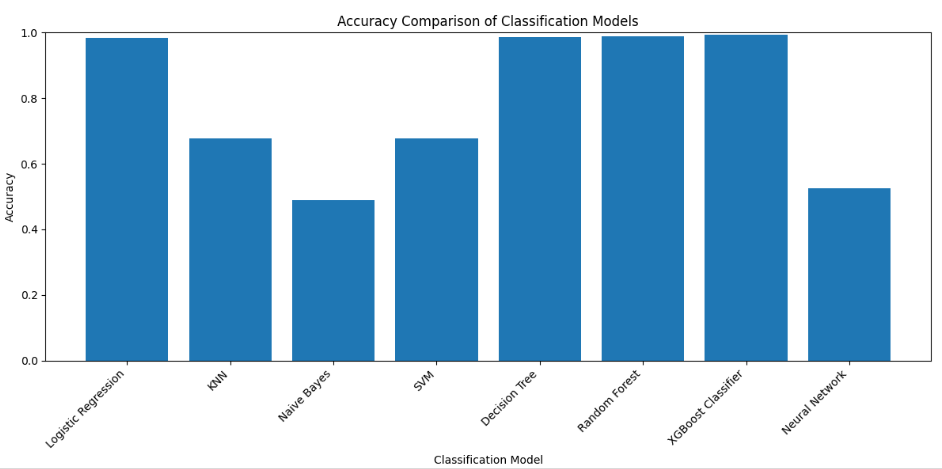
**6. Results and Analysis**

**Performance Metrics**

* Regression Models: XGBoost Regression explained variance in PTS with the highest R-squared.



* Classification Models: XGBoost outperformed with a classification accuracy of 99.38%.



**7. Conclusion and Future Scope**

Conclusion  
The project successfully demonstrated the application of machine learning to predict NBA player performance. Insights derived from the models can help teams optimize strategies and improve decision-making.

**Future Scope**

The project demonstrates significant potential for expanding the role of machine learning in sports analytics. Future enhancements could include:

* **Real-Time Data Integration:** Incorporate real-time game data, such as player movements and live statistics, to enable in-game predictions and assist teams in making tactical adjustments during matches.
* **Advanced Neural Networks:** Explore neural network architectures like RNNs for time-series analysis and CNNs for video data to improve predictions. Multi-label models could predict multiple outcomes simultaneously, such as player roles and scoring probabilities.
* **Broader Dataset Coverage:** Extend the dataset to include historical seasons, playoff data, or international leagues, enabling more robust and generalized models applicable across various basketball contexts.
* **Player Injury Prediction:** Develop predictive models to identify injury risks based on workload, game intensity, and historical health data, helping teams manage player health proactively.
* **Simulating Strategies:** Use predictive models to simulate the outcomes of different game strategies, offering coaches a virtual environment for strategy testing and optimization.
* **IoT Device Integration:** Leverage data from wearable devices monitoring player vitals and fitness to enhance predictions related to performance, fatigue, and recovery.
* **Team Metrics Predictions:** Extend the framework to predict team-level metrics, such as total points or win probabilities, using aggregated player statistics.
* **Fan Engagement:** Develop AI-powered tools for fans, such as personalized performance insights or fantasy league predictions, to enhance engagement.

These directions pave the way for creating more dynamic, accurate, and impactful applications of AI in basketball and beyond.

References

1. Kaggle Dataset: NBA 2023-24 Player Stats.
2. Python Libraries: Scikit-learn, Pandas, Matplotlib, XGBoost.