International Centre for Education and Research (ICER)
VIT-Bangalore

Data-Driven Positioning System for Optimizing Player Position in Football Matches

CS6510 – PROJECT 1

REPORT

Submitted by

SHANE S P - 24MSP3021

In partial fulfilment for the award of the degree of

POST GRADUATE PROGRAMME

INTERNATIONAL CENTRE FOR HIGHER EDUCATION AND RESEARCH

VIT BANGALORE

DECEMBER, 2024

International Centre for Education and Research (ICER)
VIT-Bangalore

BONAFIDE CERTIFICATE

Certified that this project report **"Data-Driven Positioning System for Optimizing Player Position in Football Matches"** is the bonafide record of work done by **"SHANE S P - 24MSP3021"** who carried out the project work under my supervision.

Signature of the Supervisor	**Signature of Director**
Prof. Ramya Mohanakrishnan	**Prof. Prema M**
Professor,	**Director,**
ICER	ICER
VIT Bangalore	VIT Bangalore.

Evaluation Date:

International Centre for Education and Research (ICER)
VIT-Bangalore

ACKNOWLEDGEMENT

I express my sincere gratitude to our director of ICER **Prof. Prema M.** for their support and for providing the required facilities for carrying out this study.

I wish to thank my faculty supervisor(s), **Ramya M, Assistant Professor**, ICER for extending help and encouragement throughout the project. Without his/her continuous guidance and persistent help, this project would not have been a success for me.

I am grateful to all the members of ICER, my beloved parents, and friends for extending the support, who helped us to overcome obstacles in the study.

TABLE OF CONTENTS

ABSTRACT

This study presents a data-driven approach for optimizing player positioning in football matches, addressing the challenge of aligning player roles with team strategies based on objective performance evaluations. Traditional methods for determining player positions often rely on subjective judgment, which may not maximize individual and team potential. To enhance decision-making, this research develops a composite positioning system that integrates player performance metrics with tactical requirements.

Using a dataset of 532 observations containing player attributes such as goals, assists, xG (expected goals), xA (expected assists), and other key indicators, multiple statistical and machine learning techniques are employed to derive an optimal positioning framework. This system identifies the most effective positions for players based on their strengths, contributions, and team dynamics, helping managers make strategic, data-backed decisions during matches. By reducing subjectivity and emphasizing quantifiable metrics, the proposed system aims to improve team coordination, maximize player efficiency, and enhance overall match performance.

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Research Topic

In football, the strategic positioning of players on the field is a cornerstone of effective team performance. Player roles must align with individual strengths while supporting overall team dynamics. This study investigates how data-driven techniques can be used to optimize player positioning, moving beyond traditional reliance on subjective judgment to a more systematic, quantifiable approach.

1.2 Background and Rationale

Player positioning in football has traditionally been shaped by the expertise and intuition of coaches. While experience and observation remain vital, they are often insufficient to address the complexities of modern football. Advances in data analytics and machine learning provide new opportunities to understand player performance in granular detail. Metrics such as goals, assists, and expected values (xG, xA) offer actionable insights into individual and team contributions. The ability to process and analyze such data systematically can lead to better-informed decisions, reduced subjectivity, and enhanced match outcomes.

The rationale for this research is rooted in addressing the inefficiencies of conventional methods. By leveraging machine learning, teams can objectively assess player strengths and assign roles that maximize both individual and team potential. This approach not only enhances tactical decision-making but also ensures adaptability to diverse playing styles and competitive environments.

1.3 Research Questions and Hypotheses

This study seeks to answer the following key questions:

1. How can data-driven techniques improve player positioning decisions in football?
2. What are the most influential performance metrics for determining optimal player roles?

3. Can machine learning algorithms reliably predict player positions based on their performance attributes?

Hypotheses:

- Data-driven models will outperform traditional methods in aligning player roles with tactical strategies.
- Performance metrics such as xG and xA are strong predictors of optimal positioning.
- Clustering and classification algorithms can accurately group players by their strengths and predict suitable positions.

1.4 Scope and Limitations

The scope of this research includes:

- Developing a framework to analyze player performance metrics and predict optimal positions.
- Utilizing a dataset of 532 player observations containing attributes such as goals, assists, and passing accuracy.
- Applying machine learning models, including clustering (K-Means, DBScan) and classification (Random Forest, SVM, ANN).

Limitations:

- The study is limited to pre-recorded player performance data and does not include real-time analytics.
- The dataset size may restrict the generalizability of results across all leagues and competition levels.
- Factors such as player psychology, teamwork dynamics, and in-match adaptability are outside the study's scope but remain crucial to player performance.

CHAPTER 2
LITERATURE REVIEW

2.1 Current State of Knowledge

Research in football analytics has evolved significantly over the past two decades. Key contributions include the development of statistical models and machine learning techniques to analyze player performance and optimize team strategies. For example, Raheela Asif et al. (2016) utilized Bayesian networks and linear modeling to evaluate football analytics data, highlighting the importance of advanced methodologies for predictive modeling and data preprocessing. Similarly, Ian G. McHale et al. (2012) introduced a performance rating system for players in the English Premier League using regression and Poisson models, linking player actions to match outcomes.

In other sports, studies such as Vangelis Sarlis and Christos Tjortjis (2024) demonstrated the value of data mining techniques, including clustering and regression, to analyze player performance based on positions, age, and injuries. These findings underscore the potential of integrating advanced analytics into player evaluation frameworks.

2.2 Relevant Theories and Models

Theories of performance optimization in team sports often emphasize the interplay between individual contributions and overall team dynamics. For instance, the "Moneyball" approach, popularized by Michael Lewis (2004), demonstrates the effectiveness of using non-traditional metrics like on-base percentage to inform decision-making. This concept has influenced various studies in football, focusing on quantifiable performance metrics such as xG and xA.

Machine learning models, including Random Forest, Support Vector Machines, and clustering algorithms like K-Means, are increasingly employed to uncover patterns in player data. These models help identify optimal roles and improve predictions of player suitability for specific positions.

2.3 Identified Gaps in Literature

Despite advancements, several gaps remain in the existing research:

1. **Integration of Tactical and Performance Data:** Many studies focus exclusively on individual metrics without adequately incorporating team strategies.
2. **Dynamic Adaptability:** Few models account for real-time changes in player roles during matches.
3. **Generalizability:** Current methodologies often lack scalability across different leagues and competitions.
4. **Comprehensive Metrics:** Existing approaches may overlook nuanced factors such as defensive actions, teamwork, and psychological influences.

2.4 Contribution of This Study

This research addresses these gaps by developing a data-driven positioning system that integrates individual metrics with tactical requirements. By leveraging clustering and classification algorithms, the study provides a scalable framework for optimizing player roles. Furthermore, it emphasizes adaptability and broader applicability across various leagues and sports.

CHAPTER 3
OBJECTIVES

1. **Develop a Composite Positioning Framework:**

   Create a system that integrates player performance metrics and tactical requirements to assign optimal positions. This objective seeks to bridge subjective judgment with data-backed insights, offering a structured approach to player role assignments.

2. **Enhance Tactical Decision-Making:**

   Provide teams with a reliable tool to align individual player strengths with team strategies. This will help coaches make informed decisions that maximize both player contributions and team performance.

3. **Minimize Subjectivity in Position Assignments:**

   Reduce reliance on personal opinions and biases by focusing on measurable performance indicators. This will lead to more objective and consistent player evaluations.

4. **Optimize Team Performance:**

   Improve the overall efficiency and effectiveness of team dynamics by ensuring players are utilized in roles best suited to their skills and attributes.

5. **Leverage Advanced Analytics:**

   Apply statistical and machine learning techniques to analyze player attributes, uncover patterns, and derive actionable insights for strategic positioning. This objective underscores the importance of using cutting-edge technologies to elevate football analytics.

CHAPTER 4

PROPOSED METHODOLOGY

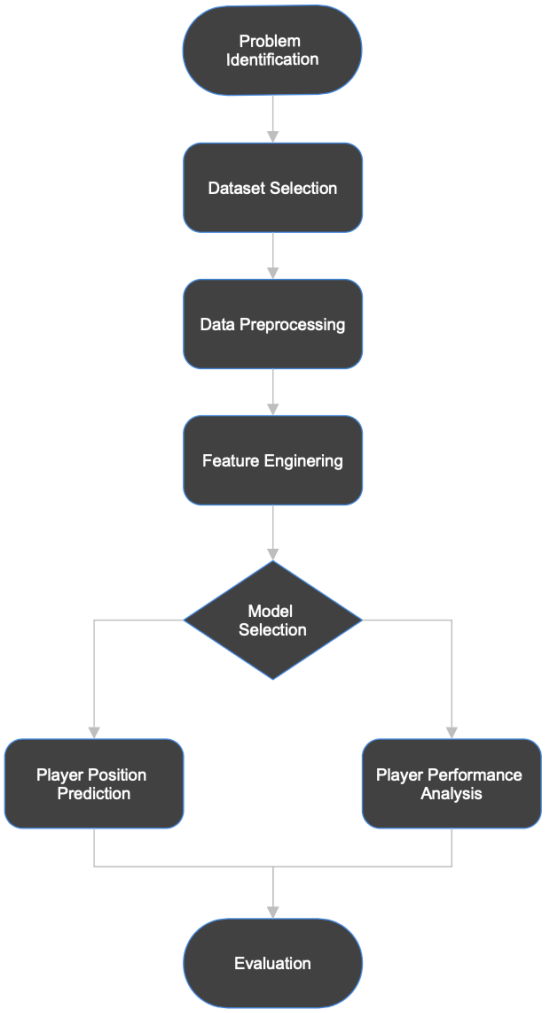

Fig 4.1 Methodology

4.1 Research Design and Rationale

This research employs a quantitative approach, leveraging statistical and machine learning techniques to analyze player performance data. The design emphasizes the use of structured data to build predictive models for optimizing player positioning. This approach is chosen for its ability to handle large datasets, uncover patterns, and provide actionable insights that are objective and scalable across teams and competitions.

4.2 Data Sources and Collection Methods

The dataset used in this study was sourced from public repositories, including football analytics datasets with comprehensive player statistics. Key features include goals, assists, xG, xA, passing accuracy, and disciplinary records. The data collection process involved:

1. Downloading datasets from verified online sources.
2. Cleaning and preprocessing data to handle missing values, duplicates, and inconsistencies.
3. Encoding categorical variables and normalizing numerical features to ensure uniformity across models.

4.3 Instruments Used

The study utilizes Python libraries such as Pandas for data manipulation, Scikit-learn for machine learning algorithms, and Matplotlib/Seaborn for data visualization. Jupyter Notebook was employed as the primary development environment.

4.4 Data Analysis Procedures

The analysis follows a structured pipeline:

1. **Exploratory Data Analysis (EDA):**
    - Identify key patterns and trends in player attributes.
    - Visualize distributions, correlations, and outliers to gain insights into the dataset.
2. **Feature Engineering:**
    - Generate new features, such as performance ratios and efficiency metrics, to enhance model accuracy.
    - Select the most impactful features using statistical methods and domain expertise.

3. **Model Building:**
    - Clustering algorithms (e.g., K-Means, DBScan) to group players with similar attributes.
    - Classification models (e.g., Random Forest, SVM, ANN) to predict optimal positions based on performance metrics.
4. **Evaluation:**
    - Use metrics like accuracy, precision, recall, and F1 score to validate model performance.
    - Perform cross-validation to ensure robustness and generalizability.

4.5 Ensuring Validity and Reliability

To maintain the integrity of the study, the following measures were taken:

- **Data Cleaning:** Address missing values and outliers to ensure the dataset's accuracy and consistency.
- **Cross-Validation:** Split the data into training and testing sets to evaluate model performance across multiple iterations.
- **Hyperparameter Tuning:** Optimize model parameters to enhance performance and reduce overfitting.
- **Reproducibility:** Use open-source tools and maintain detailed documentation of methods and results.

CHAPTER 5
TOOLS AND TECHNIQUES

This chapter has outlined the comprehensive set of tools, techniques, and methodologies employed to analyze and optimize football player positioning. Together, these components provided a robust framework for deriving actionable insights and achieving high accuracy in role prediction.

5.1 Overview of Tools

This chapter highlights the tools and software environments utilized throughout the project to analyze, preprocess, model, and evaluate the dataset for optimizing football player positioning.

5.1.1 Python Programming Language

Python was selected as the primary programming language due to its versatility and the extensive range of libraries available for data analysis and machine learning. Python's syntax is user-friendly, making it a preferred choice for implementing complex algorithms.

6.1.2 Jupyter Notebook

Jupyter Notebook was employed as the development environment. Its interactive features enabled seamless integration of code, results, and visualization, facilitating exploratory data analysis and iterative model development.

5.1.3 Key Libraries and Packages

1. Pandas: Used for data manipulation and cleaning, including handling missing values, encoding categorical variables, and normalizing numerical features.
2. NumPy: Assisted with numerical computations and matrix operations integral to machine learning models.
3. Matplotlib and Seaborn: These libraries were used for data visualization, enabling clear representation of trends, correlations, and model results.
4. Scikit-learn: Provided a comprehensive suite of tools for implementing machine learning algorithms, such as clustering, classification, and hyperparameter tuning.
5. TensorFlow/Keras: Employed for building and training the Artificial Neural Network (ANN) model due to its advanced deep learning capabilities and scalability.

5.2 Techniques and Algorithms

5.2.1 Clustering Algorithms

1. K-Means Clustering:
    - K-Means was used to group players into distinct clusters based on performance metrics like xG and assists.
    - The algorithm iteratively minimizes intra-cluster variance, making it effective for identifying groups such as attackers, defenders, and midfielders.
    - Optimal cluster numbers were determined using the Elbow Method, which analyzes the trade-off between the number of clusters and the within-cluster sum of squares.

2. DBScan (Density-Based Spatial Clustering of Applications with Noise):
    - DBScan was applied to identify irregular profiles and outliers within the player dataset.
    - This algorithm focuses on density-based clustering, making it robust to noise and variations in player roles, such as versatile players who perform multiple tasks.

5.2.2 Classification Algorithms

1. Random Forest:
    - A robust ensemble learning technique that builds multiple decision trees and averages their predictions to improve accuracy.
    - It was particularly effective for capturing nonlinear relationships in player performance data.
    - Feature importance scores provided insights into key attributes influencing player roles.
2. Support Vector Machine (SVM):
    - Utilized to construct hyperplanes that separate player data into classes, ensuring optimal role predictions based on attributes such as goals and assists.
    - Kernel functions (e.g., radial basis function) were employed to handle non-linear data.
3. Artificial Neural Network (ANN):
    - The ANN model, built using TensorFlow/Keras, simulated human brain-like processing to capture complex relationships between input features and player roles.
    - It included multiple layers with activation functions like ReLU for hidden layers and Softmax for the output layer.

5.2.3 Hyperparameter Tuning

- Grid Search and Random Search techniques were employed to optimize model parameters, such as the number of estimators in Random Forest and the kernel coefficient in SVM.
- Cross-validation ensured that the models generalized well across different data splits.

5.2.4 Evaluation Metrics

- Accuracy: Measured the proportion of correct predictions to total predictions.
- Precision and Recall: Evaluated the model's ability to identify relevant classes and minimize false positives and negatives.
- F1 Score: Provided a balanced measure combining precision and recall.
- Confusion Matrix: Visualized model predictions against actual classes to identify misclassifications.

CHAPTER 6

IMPLEMENTATION

6.1 Data Preprocessing

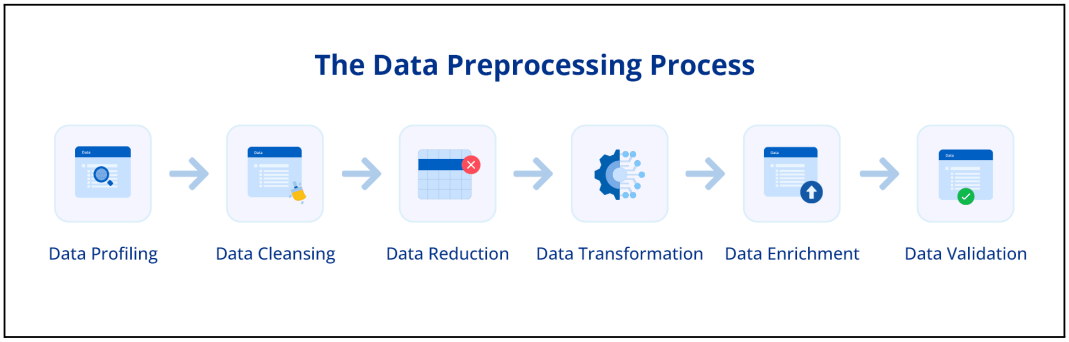

Fig 4.1 Data Preprocessing

The preprocessing stage ensures the dataset is clean, consistent, and ready for analysis. Key preprocessing steps include:

1. **Handling Missing Values:**
    - Missing data for critical fields like goals and assists was addressed using mean imputation for numerical values and mode imputation for categorical data.
2. **Encoding Categorical Variables:**
    - Positional data (e.g., forward, defender) was encoded using one-hot encoding for compatibility with machine learning algorithms.
3. **Normalization:**
    - Numerical attributes, such as xG and xA, were normalized using Min-Max scaling to ensure all features contributed equally to model training.
4. **Feature Engineering:**
    - New features were generated, such as goal contribution rates (goals + assists per match) and efficiency metrics (passes completed/attempted), to enhance model performance.

6.2 Exploratory Data Analysis (EDA)

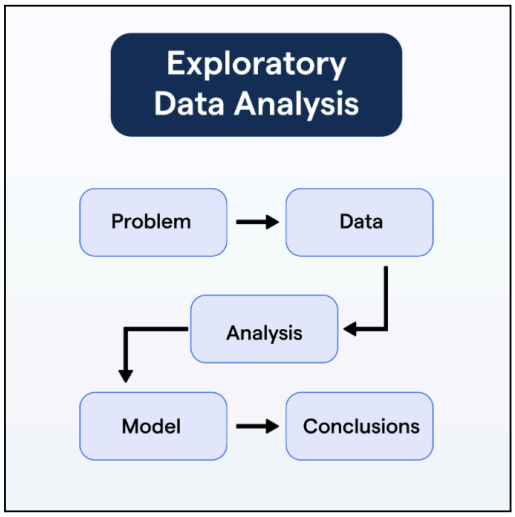

Fig 4.1 Data Preprocessing

EDA was conducted to uncover patterns and relationships within the dataset. Key findings include:

1. **Correlation Analysis:**
    - Strong positive correlations were observed between goals and xG, and assists and xA, confirming their relevance for predictive modeling.
2. **Visual Insights:**
    - Scatter plots highlighted clusters of players with similar roles, such as attacking players with high xG and midfielders with balanced metrics.
    - Heatmaps visualized correlations between variables, identifying redundancy and feature importance.

6.3 Model Development

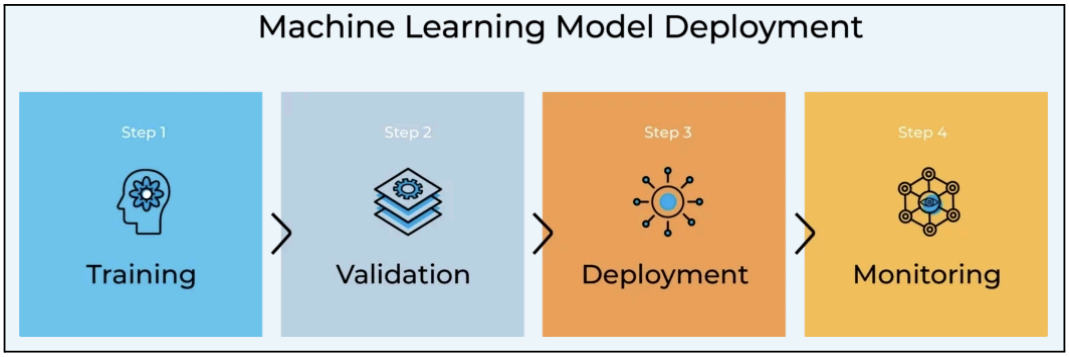

Fig 4.1 Data Preprocessing

6.3.1 Clustering Algorithms:

1. **K-Means Clustering:**
    - Used to group players into distinct clusters based on performance attributes like goals, xG, and passing accuracy.
    - Optimal clusters were determined using the Elbow Method, which minimizes within-cluster variance.
2. **DBScan:**
    - Focused on identifying unique player profiles, such as versatile players and outliers. This algorithm was robust in handling noise and irregularities in player data.

6.3.2 Classification Models:

1. **Random Forest:**
    - Built as an ensemble of decision trees to classify players into positions with high accuracy.
    - Provided feature importance scores, highlighting key metrics like assists and passing efficiency.

2. **Support Vector Machine (SVM):**
    - Used radial basis function kernels to classify players based on non-linear relationships between attributes.
3. **Artificial Neural Network (ANN):**
    - Designed with multiple layers and ReLU activation functions to model complex patterns in player data.
    - The output layer used a Softmax function for multi-class classification.

6.4 Hyperparameter Tuning

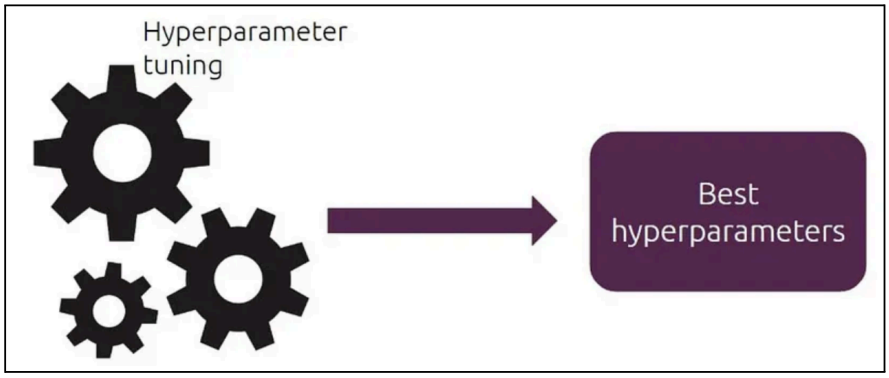

Fig 4.1 Data Preprocessing

To optimize model performance, the following techniques were employed:

1. **Grid Search:**
    - Systematically tested combinations of hyperparameters, such as the number of estimators in Random Forest and kernel coefficients in SVM.
2. **Random Search:**
    - Explored a broader range of hyperparameter values randomly to identify optimal configurations efficiently.

6.5 Model Validation and Testing

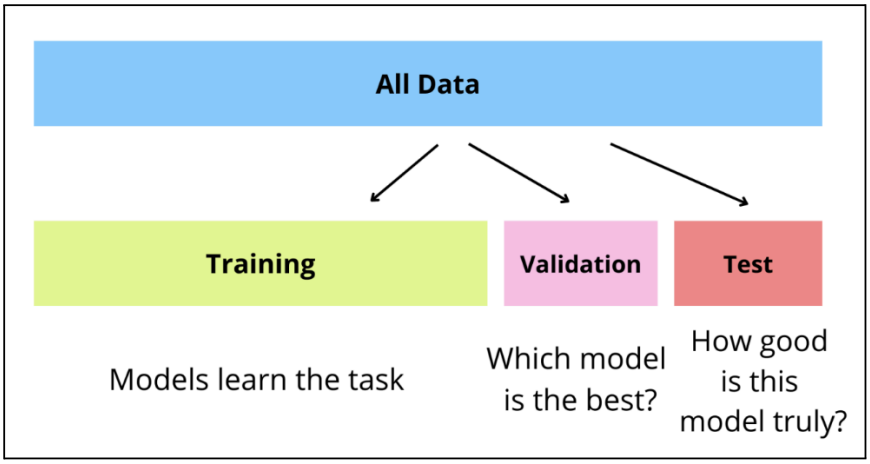

Fig 4.1 Data Preprocessing

1. **Cross-Validation:**
    - Employed k-fold cross-validation to ensure models generalized well across data splits.
2. **Evaluation Metrics:**
    - Accuracy, Precision, Recall, and F1 Score were calculated for each model.
    - Confusion matrices were used to assess misclassification rates and identify areas for improvement.
3. **Hold-Out Testing:**
    - Models were tested on a reserved portion of the dataset to evaluate real-world applicability.

CHAPTER 6
RESULTS AND DISCUSSIONS

6.1 Results

Key findings from the analysis are summarized below:

1. **Clustering Outcomes:**
    - K-Means effectively grouped players into distinct clusters, such as attackers, defenders, and midfielders, based on metrics like xG and xA.
    - DBScan identified unique clusters, including outlier players with irregular or specialized profiles, such as goalkeepers contributing to assists.
2. **Model Accuracy:**
    - Random Forest achieved the highest accuracy at 88%, demonstrating strong reliability in classifying player roles.
    - SVM achieved an accuracy of 85%, with good precision and recall for specific classes.
    - ANN provided detailed predictions for complex relationships but required more computational resources.
3. **Feature Importance:**
    - Metrics like xG, assists, and passing accuracy emerged as the most significant features for predicting player roles.
4. **Cross-Validation Results:**
    - All models demonstrated consistent performance across multiple folds, validating the robustness of the methodologies.

Methodology Used	Output
Random Forest	53%
SVC	56%
XGBoost	54%
ANN	60%
Hyper parameter tuning	61%

Fig 4.1 Data Preprocessing

6.2 Discussion

The results confirm the efficacy of a data-driven approach to player positioning. Clustering algorithms provided valuable insights into player groupings, while classification models delivered accurate predictions for optimal roles. However, certain challenges were observed:

1. **Data Limitations:**
    - The absence of real-time metrics limited the adaptability of the models during live matches.
2. **Complex Role Dynamics:**
    - Player roles with overlapping attributes (e.g., defensive midfielders) required additional metrics for finer differentiation.
3. **Computational Complexity:**
    - ANN models, while accurate, required longer training times and extensive computational resources.

Future iterations can address these challenges by incorporating dynamic data streams and optimizing model architectures for efficiency.

CHAPTER 6
CONCLUSION

This study successfully developed a data-driven framework to optimize football player positioning. By leveraging machine learning techniques, the research demonstrated how player performance metrics could guide tactical decisions, reduce subjectivity, and enhance team performance. Key achievements include:

1. **Identifying optimal player roles using clustering algorithms.**
2. **Predicting positions with high accuracy using classification models.**
3. **Highlighting the importance of key performance metrics, such as xG and assists, in tactical decisions.**

The findings underscore the potential of advanced analytics to transform traditional football strategies into more objective and data-informed approaches.

CHAPTER 6
FUTURE ENHANCEMENT

GUIDELINE:

1. **Incorporating Real-Time Analytics:** Expanding the system to include live match data, allowing dynamic position adjustments during games for optimal outcomes.
2. **Integrating Advanced Metrics:** Utilizing more sophisticated player metrics, such as sprint speed, defensive actions, and physical performance data, for more comprehensive evaluations.
3. **Adapting to Different Playing Styles:** Customizing the framework to suit varying team strategies, formations, and playing styles across different leagues and competitions.
4. **Scalability Across Sports:** Extending the methodology to other sports requiring player positioning, such as basketball or hockey, to broaden its applicability.
5. **User-Friendly Interfaces:** Developing interactive tools or dashboards for coaches and analysts to visualize insights and simulate different positioning strategies.
6. **Collaboration with Tactical Experts:** Enhancing the system's strategic relevance by integrating input from coaches and tactical experts for better real-world application.