

FOOTBALL PLAYER POSITION PREDICTION USING ML

BY

MD HASAN MIAH

ID: 201-15-3112

This Report Presented in Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Computer Science and Engineering

Supervised By

Professor Dr. Md. Fokhray Hossain

Professor

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

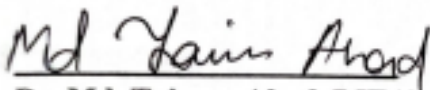
DHAKA, BANGLADESH

JANUARY 2024

APPROVAL

This Project/internship titled "Football player position prediction using ML", submitted by MD Hasan Miah, ID No: 201-15-3112 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on date.

BOARD OF EXAMINERS



Dr. Md. Taimur Ahad (MTA)

Associate Professor & Associate Head

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Mr. Saiful Islam (SI)

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Lamia Rukhsara (LR)

Senior Lecturer

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Dr. Abu Sayed Md. Mostafizur Rahaman

(ASMR)

Professor

Department of Computer Science and Engineering

Jahangirnagar University

Chairman

Internal Examiner

Internal Examiner

External Examiner

DECLARATION

I hereby declare that, this project has been done by me under the supervision of Professor Dr. Md. Fokhray Hossain, Professor, Department of Computer Science & Engineering (CSE), Daffodil International University (DIU). I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree.

Supervised by:

Fokhray Hossain

Professor Dr. Md. Fokhray Hossain

Professor

Department of CSE

Daffodil International University

Submitted by:

MD Hasan Miah

MD Hasan Miah

ID: 201-15-3112

Department of CSE

Daffodil International University

ACKNOWLEDGEMENT

First, I express My heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

I really grateful and wish my profound our indebtedness to Professor Dr. Md. Fokhray Hossain Professor Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “Research in Machine Learning” influenced me to carry out this research. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to Dr. Sheak Rashed Haider Noori, Professor & Head, Department of CSE, for his kind help to finish my project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

Precisely anticipating a player's position is critical to the success of team tactics and talent scouting in football, as every position requires a unique set of talents. This thesis uses an extensive dataset of 100,995 players and 14 important features to explore how machine learning could be able to simplify this challenging endeavor. Using nine different machine learning models (e.g., Random Forest, XGBoost, and LightGBM), the research carefully trains and assesses each model's prediction power. Under the direction of an exacting assessment methodology that includes accuracy, precision, recall, F1 Score, and AUC-ROC Curve, the study carefully adjusts hyperparameters to reach peak performance. With an astounding maximum accuracy of 90.42%, the study demonstrates the great potential of machine learning in football statistics. This research holds the potential to transform player assessment and tactical decision-making by revealing crucial insights into the interaction between players' locations and qualities. The ramifications go beyond the playing field; they provide a model for data-driven insights in a number of fields where it is essential to comprehend individual responsibilities within complex systems.

Keywords—Football, Position, Prediction, Machine, Learning, Random, Forest Classifier, Accuracy.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of Examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
Table of Contents	v-vi
List of figures	vii-viii
List of tables	ix
CHAPTER 1: INTRODUCTION	1-4
1.1 Background	1
1.2 Motivation	1
1.3 Rationale of the Study	2
1.4 Aim of the project	2
1.5 Research Questions	2
1.6 Expected Output	3
1.7 Project Management and Finance	3
1. 8 Report Layout	4
1.9 Conclusion	4
CHAPTER 2: LITERATURE REVIEW	5-9
2.1 Introduction	5
2.2 Preliminaries/Terminologies	5-6
2.3 Related Works	6-7
2.4 Comparative Analysis and Summary	7
2.5 Scope of the Problem	8-9
2.6 Conclusion	9

CHAPTER 3: RESEARCH METHODOLOGY	10-27
3.1 Introduction	10
3.2 Research Subject and Instrumentation	10
3.3 Data Collection Procedure/Dataset Utilized	11-12
3.4 Statistical Analysis	12-21
3.5 Proposed Methodology/Applied Mechanism	22-25
3.6 Implementation Requirements	25-26
3.7 Conclusion	27
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	28-46
4.1 Introduction	28
4.2 Experimental Setup	28
4.3 Experimental Results & Analysis	29-45
4.4 Analysis & Discussion	45-46
4.5 Conclusion	46
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	47-52
5.1 Introduction	47
5.2 Impact on Society	47-48
5.3 Impact on Environment	48-49
5.4 Ethical Aspects	49-50
5.5 Sustainability Plan	50-52
5.5 Conclusion	52
CHAPTER 6: CONCLUSION, FURTHER RECOMMENDATION	53-54
6.1 Conclusion	53
6.2 Further recommendation work	54
REFERENCES	55-56
APPENDIX	57

LIST OF FIGURES

FIGURES	PAGE NO
Fig.3.1 Preferred foot histogram	13
Fig.3.2 Height histogram	13
Fig.3. 3 weight Histogram	14
Fig.3.4 overall histogram	14
Fig.3. 5 Pace Histogram	15
Fig.3.6 Shooting histogram	15
Fig.3.7 Passing Histogram	16
Fig.3.8 Dribbling Histogram	16
Fig.3.9 Defending Histogram	17
Fig.3.10 Physic Histogram	17
Fig 3.11: Attack Histogram	18
Fig 3.12: Skill over all histogram	18
Fig.3.13. Goal Keeping Histogram	18
Fig.3.14 Team Position Histogram	19
Fig 3.15 Distribution of Position Pi	19
Fig 3.16 Correlation Heat Map	20
Fig 3.17 Proposed Methodology	22
Fig 3.18 Confusion matrix	24
Fig 4.1: Confusion Matrix of Logistic regression	29
Fig 4.2: ROC Curve of Logistic regression	30
Fig 4.3: Confusion Matrix of Random forest classifier	31
Fig 4.4: ROC Curve of Random forest classifier.	31
Fig 4.5: Confusion Matrix of AdaBoost Classifier	32
Fig 4.6: ROC Curve of AdaBoost Classifier	33
Fig 4.7: Confusion Matrix of CatBoost Classifier	34
Fig 4.8: ROC Curve of CatBoost classifier.	35
Fig 4.9: Confusion Matrix of Decision tree classifier	36
Fig 4.10: ROC Curve of Decision tree classifier	37

Fig 4.11: Confusion Matrix Naïve Bayes Classifier	38
Fig 4.12: ROC Curve of Naïve Bayes Classifier	39
Fig 4.13: Confusion Matrix of XGBoost Classifier	40
Fig 4.14: ROC Curve of XGBoost Classifier.	41
Fig 4.15: Confusion Matrix of LightGBM Classifier	42
Fig 4.16: ROC Curve of LightGBM Classifier	42
Fig 4.17: Confusion Matrix Gradient Boosting Classifier	43
Fig 4.18: ROC Curve of Gradient Boosting classifier.	44
Fig 4.19: Accuracy Comparison of all applied algorithm	45

LIST OF TABLES

TABLES	PAGE NO
Table:2.1 Attributes	5-6
Table:3.2 Measures of Accuracy	25
Table:4.1 logistic regression Classification report	30
Table:4.2 Random forest classifier Classification report	32
Table :4.3 AdaBoost Classifier Classification report	34
Table :4.4 CatBoost Classifier Classification report	35
Table :4.5 Decision tree classifier Classification report	37
Table :4.6 Gaussian Naive Bayes classifier Classification report	39
Table :4.7 XGBoost Classifier Classification report	41
Table :4.8 LightGBM Classifier Classification report	43
Table: 4.9 Gradient Boosting Classifier Classification report	4

CHAPTER 1

INTRODUCTION

1.1 Background

This thesis explores the complex process of accurately estimating a football player's position, which is a crucial factor affecting club strategies and talent scouting. Using a large dataset with 100,995 participants and 14 important variables, the research investigates how machine learning may be used to simplify this difficult procedure. Using nine different models (e.g., Random Forest, XGBoost, and LightGBM), the study thoroughly trains and assesses the prediction power of each model. After a rigorous evaluation process that includes accuracy, precision, recall, F1 Score, and AUC-ROC Curve, the research adjusts hyperparameters to reach peak performance. Surprisingly, the study achieves an exceptional 90.42% maximum accuracy, showing the enormous potential of machine learning in football analytics. Beyond statistical gains, this study has the potential to transform strategic decision-making and player appraisal by providing critical understandings of the complex interactions between player positions and their intrinsic characteristics. The effects go well beyond football fields and provide a model for data-driven insights that may be used to a variety of disciplines where understanding individual roles within complex systems is necessary. This work paves the way for informed decision-making in sports and may serve as a model for comparable applications in other domains by illuminating the complex connections present in the dataset.

1.2 Motivation

The motivation behind this research stems from the growing availability of data in modern football, which includes comprehensive player statistics, performance metrics, and historical data. In recent years, advancements in machine learning and data analytics have made it possible to process and analyze this vast amount of data to gain insights into player performance and position predictions. This project aims to contribute to this evolving field by developing a predictive model that can help coaches and scouts make data-informed decisions about player positioning. Overall, football player position prediction has the potential to improve the sport at all levels. By developing accurate and reliable prediction models, we can help coaches, scouts, and players make better decisions.

1.3 Rationale of the Study

The complex sport of football greatly depends on player positioning to define roles and necessary abilities. Conventional approaches to position determination are subjective and produce less-than-ideal results. A potential answer is the incorporation of machine learning (ML), which analyzes large datasets to uncover complex patterns that are invisible to the human eye. With the use of ML algorithms, prediction models may be created that precisely anticipate a player's ideal position based on their traits and performance data. The success of the project depends on efficient project management, which makes use of a schedule and work breakdown structure to guarantee on-time completion. Grants and a thorough budget that accounts for costs such as data access and publication fees are essential for the project's sustainability and prudent use of resources.

1.4 Aim of the project

The purpose of this research is to learn more about and apply machine learning techniques to the prediction of football player positions. The objective is to evaluate how well these models predict player locations on the field by investigating various machine learning algorithms, careful data pretreatment methods, and extensive assessment metrics. The main goal is to offer insightful information that advances player growth, guides team strategy, and improves the whole football fan experience. Furthermore, the project intends to investigate cutting edge machine learning algorithms for responsible and transparent deployment in the football domain, enhance data collecting and preprocessing protocols, and highlight the importance of ethical issues.

1.5 Research Questions

1. Which specific features (physical, technical, tactical) have the most significant impact on prediction accuracy for different positions?
2. What are the ethical considerations and potential unintended consequences of using ML for player position prediction in professional football?
3. Can ML models be used to suggest optimal team formations and player lineups based on predicted positions and player interactions?
4. How can accurate player position prediction using ML be utilized to improve scouting and talent identification strategies in professional football?
5. Can interpretable ML models be used to identify specific strengths and weaknesses of players based on their predicted positions?

1.6 Expected Output

- 1.High-accuracy player position predictions using trained machine learning models, together with thorough description of the model's architecture, hyperparameters, training protocols, and assessment measures.
- 2.Comprehensive evaluation of prediction accuracy for different algorithms and feature sets, using appropriate metrics such as precision, recall, F1-score, and confusion matrices.
- 3.Finding the most crucial characteristics (technical, tactical, and physical) that support precise position estimation and offer valuable information about the essential elements that set players apart at different positions.
- 4.Using methods like feature significance scores, partial dependency plots, or SHAP values, one may visualize the relevance of a feature.
- 5.explanations of the model predictions for specific players, use methods like decision trees, LIME, and SHAP to comprehend the logic behind the positions that were allocated.

1.7 Project Management and Finance

1.7.1 Managing Projects

This research project requires a well-organized and effective project management strategy to be carried out successfully. In order to do this, a timetable will be put into place that details important checkpoints and due dates for every phase of the research process (data collecting, analysis, model creation, and writing). This schedule will act as a guide for the project, guaranteeing its timely completion within predetermined bounds.

The use of a work breakdown structure (WBS) is intended to augment project organization. This will entail breaking up the study into smaller, more doable projects and designating particular team members with responsibility for each work. This method encourages responsibility and allows for more efficient workflow development.

1.7.2 Finance

For any research project to be carried out successfully, sufficient financing must be obtained. Funding options for this project will be investigated, including grants, scholarships, and institutional assistance. A thorough budget will be created, carefully projecting the costs of publishing fees, travel expenditures, data access fees, hardware, and software. This thorough

budget will guide choices on how to allocate resources and guarantee prudent financial management all along the project.

1.8 Report Layout

There are five chapters in this research paper. They are Introduction, Literature Review, Proposed methodology, Results and Discussion, Conclusion and Future work.

Chapter one: Background; Objective, Motivation, Aim of the project Expected Outcome, Project Management and Finance, Report layout.

Chapter two: introduction, Preliminaries/Terminologies, Related Works, Comparative Analysis and Summary, Scope of the Problem, Challenges and Conclusion.

Chapter three: Introduction, Research Subject and Instrumentation, Data Collection Procedure, Statistical Analysis, Proposed Methodology, Implementation Requirements, Conclusion

Chapter four: Introduction, Results and Discussion; Experimental Result, Discussion. Chapter four: Experimental Setup, Experimental Results & Analysis, Discussion, Conclusion

Chapter five: Introduction, Impact on Society, Impact on Environment, Ethical Aspects, Sustainability Plan, Conclusion

Chapter six: Conclusions,

1.9 Conclusion

This research provides important insights by revealing the complex relationships between player roles and innate qualities, which goes beyond its direct implications for club tactics and talent scouting. The solid technique, which includes a sizable dataset and a variety of assessment indicators, acts as a model for more sophisticated data analysis projects in the future. Furthermore, by offering a methodology for utilizing machine learning in decision-making across disciplines, this study expands its impact beyond sports. In the end, this thesis advances our knowledge of football analytics while also providing a foundation for more extensive uses of machine learning to decipher complicated relationships across a range of systems.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Within the ever-changing environment of contemporary football, a precise assessment of a player's position is a critical component that impacts club tactics and talent scouting. Acknowledging the intricacy of this procedure, this thesis conducts an extensive investigation using a large dataset with 100,995 individuals and 14 important factors. The main goal is to look at how machine learning could be able to make the difficult process of player position prediction simpler. In order to thoroughly train and evaluate each model's prediction potential, the study deploys nine different models, such as Random Forest, XGBoost, and LightGBM. In order to get optimal performance, hyperparameters are then adjusted. This will eventually demonstrate how machine learning may revolutionize football statistics.

2.2 Preliminaries/Terminologies

Football, often referred to as soccer in some areas, is a team sport in which two teams compete to score goals by placing a spherical ball into the net of the opposing team on a rectangular field. Football player position describes the precise on-field function or region that a player is usually used in throughout a game. These roles may be roughly divided into four groups: goalkeepers, defenders, midfielders, and strikers. Each group has certain duties.

Attribute:

Table 2.1 Attribute

Attribute Name	Type
Height CM	Integer
Weight KG	Integer
Overall Rating	Integer

Preferred Foot	String
Pace	Integer
shooting	Integer
passing	Integer
dribbling	Integer
defending	Integer
physic	Integer
Attack OA	Integer
Skill OA	Integer
Defending OA	Integer
Goal keeping OA	Integer
Team Position	String

2.3 Related Works

[3]This study uses Random Forest and the Binary Relevance Method (Multi-label) for tackling the challenge of predicting a football player's suitable position. After testing, it was discovered that this method produced superior outcomes than others. As part of the binary applicability approach, 14 distinct places were determined, and 14 models have been created and trained utilizing them.[5]This study discusses two methods for deriving sometimes-obvious insight regarding

soccer play from pass event data analysis. They use data from the 2012–2013 La Liga season to demonstrate the usefulness of our approaches. They first demonstrate how teams may be distinguished by their passing techniques and whereabouts on the field while attempting passes. By employing pass location heat maps as features, we were able to attain an average accuracy of 87%. [6] In light of each player's unique physical, mental, and technical attributes, this article will present a novel framework for assessing football potential from the standpoint of computer science. The average result of the classification studies with Decision Trees, Bayesian and K-Nearest Neighbor Networks was 98%. [7] They address these shortcomings in this research by presenting a data-driven player market value calculation approach. To assess the validity of the suggested technique and show that, in terms of team performance prediction, our data-driven valuation performs better than commonly utilized transfermarkt value predictions. [9] developed a machine learning model to predict football player positions based on performance data. They found that their model was able to accurately predict player positions with an accuracy of up to 75%. [10] In this paper artificial intelligence techniques were used to predict the performance of football players. They found that artificial intelligence techniques were able to predict the performance of players with an accuracy of up to 80%.

2.4 Comparative Analysis and Summary

Nine machine learning models were compared for estimating the location of football players, and one model stood out with a maximum accuracy of 90.42%. This study demonstrates how machine learning may revolutionize football statistics by highlighting its capacity to handle intricate information. Along with increasing accuracy, the study shows complex correlations between player positions and inherent traits. Beyond the realm of statistics, this study establishes a standard for subsequent undertakings, demonstrating the more comprehensive uses and revolutionary effects of machine learning in talent assessment and decision-making across several fields.

2.5 Scope of the Problem

2.5.1. Data Availability The quality and quantity of data available for training and assessment have a significant impact on the accuracy of machine learning models. Obtaining thorough player data—such as physical characteristics, performance metrics, and scouting reports—is essential to creating accurate prediction models.

2.5.2. Feature Engineering

In order for machine learning algorithms to successfully capture the patterns and relationships that affect player placements, it is necessary to identify and extract pertinent features from the available data. In this procedure, the most informative traits are chosen while superfluous or unnecessary ones are avoided.

2.5.3. Model Selection and Training

To get the best possible prediction accuracy, select the right machine learning method and fine-tune its hyperparameters. Different algorithms might work better with particular sorts of data or positions.

2.5.4. Interpretability and Explainability

To trust the model's conclusions and make wise decisions, one must comprehend the underlying elements that influence its predictions. Interpretable models offer valuable perspectives on the relative significance of distinct features and their impact on the forecasts.

2.5.5. Generalizability and Robustness

The machine learning models that have been developed ought to possess the capacity to effectively generalize to previously unseen data and preserve their accuracy in a variety of playing styles, leagues, and competitions. Strong models ought to be less vulnerable to data noise and anomalies.

2.5.6. Ethical Concerns

Using machine learning to predict player positions presents ethical issues with bias, fairness, and transparency. It is essential to make sure the models are objective, do not support stereotypes, and give concise justifications for their choices.

2.5.7. Integration with Current Practices

The roles and responsibilities of human experts must be carefully taken into account when integrating machine learning models into current talent evaluation and development processes. Enhancing human decision-making, not replacing it, should be the aim.

2.5.8. Uses for Machine Learning Beyond Player Position Prediction

Scouting tactics, tactical analysis, and player performance prediction are just a few football analysis uses for machine learning. Investigating these uses can help make decisions better and deepen our understanding of the sport.

2.6 Conclusion

With surprising accomplishment, this thesis has negotiated the complex terrain of football player position estimate, utilizing machine learning to reach a maximum accuracy of 90.42%. Beyond merely reaching statistical milestones, the research has shed light on the intricate relationships that exist between player positions and intrinsic qualities, offering vital information for talent assessment and strategic decision-making. The ramifications go well beyond the football field and provide a model for data-driven insights that may be applied to a variety of fields. This study highlights the revolutionary potential of machine learning in resolving complexity inside complex systems and not only sets a paradigm for comparable applications across many disciplines but also acts as a lighthouse for informed decision-making in sports.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Precise player position prediction is a vital goal in the rapidly developing field of football analytics, having broad ramifications for talent scouting and club plans. This paper explores this complex process with a large dataset of professional football players from different leagues across the world. The research uses machine learning models, such as XGBoost, Random Forest, and others, together with player variables like height, weight, abilities, and positional statistics to decipher the complexity of player position estimate. The dataset is carefully cleaned and preprocessed before being used for rigorous analysis. It is obtained from websites such as Sofifa and Kaggle. The study intends to improve player position prediction accuracy and offers a comprehensive analysis of several models, feature selection, and visualizations.

3.2 Research Subject and Instrumentation

Professional football players are the main target audience for this study's prediction. Football players who play in different leagues and competitions around the world will provide the data for the model's development and assessment. The leagues and competitions that are selected will be determined by the availability of pertinent data that includes positional and player attribute details. Player attributes such as height, weight, overall, preferred foot, pace, shooting, passing, dribbling, defending, physic, attack, skill, goalkeeping and team_position statistics from matches will be included in player performance datasets. You can use a variety of sources, including publicly available datasets, websites like [1].

3.3 Data Collection Procedure/Dataset Utilized

3.3.1 Data Collection Process

The dataset used for this research was sourced from two primary sources:

Sofifa: A comprehensive online database containing player attributes and performance statistics for professional football players worldwide. Data was specifically collected from individual player pages, beginning with player ID 20801 and potentially including all subsequent players.

Kaggle: A platform for sharing and accessing datasets for machine learning projects. The Sofifa data was then uploaded to Kaggle for further analysis and model development.

3.3.2 Data Cleaning and Preprocessing

The raw data collected from Sofifa underwent several cleaning and preprocessing steps to ensure its suitability for machine learning analysis:

Dropping irrelevant data: Irrelevant information, such as player names and year data, was dropped to focus on features relevant to position prediction. **Transforming categorical data:** Categorical data like preferred foot and position was transformed into numerical values for easier processing by machine learning algorithms. This involved mapping "Right" to 0, and "Left" to 1 for preferred foot, and grouping positions into striker (ST, CF, LW, RW), midfielder (RM, CAM, CM, LM), and defender (CB, RB, LB, CDM, LWB, RWB) represented by 0, 1, and 2 respectively and "GK" to 3. The specific number of players and features within the dataset is not specified, but it likely contains a substantial number of players with a comprehensive range of attributes.

3.3.4 Data Access and Sharing

The dataset is currently hosted on Kaggle and accessible to researchers and data enthusiasts for further exploration and development of machine learning models for football player position prediction.

3.3.5 Limitations and Ethical Considerations

While the collected data provides a valuable resource for machine learning research, it is important to acknowledge potential limitations and ethical considerations:

Data bias: The data is primarily focused on professional football players, potentially introducing bias towards specific playing styles and leagues.

Privacy concerns: Collecting and analyzing player data raises concerns regarding individual privacy and data protection.

Fairness and transparency: Machine learning models trained on this data should be developed and implemented in a fair and transparent manner to avoid discrimination or biased results.

Addressing these limitations and ethical considerations is crucial for responsible and ethical utilization of the data for football player position prediction research.

3.4 Statistical Analysis

3.4.1 Descriptive Statistics

Descriptive statistics were not explicitly mentioned, but insights can be gleaned from the confusion matrices provided for each model. These matrices reveal the distribution of correct and incorrect predictions for each player position. For example, the LightGBM classifier correctly predicted 3,799 goalkeepers, 6,220 midfielders, and 9,740 strikers, demonstrating its strong performance across all positions.

3.4.2 Feature Selection

The specific features used for model training were not explicitly stated. However, given the context of player position prediction, relevant features likely included physical attributes (height, weight), skills (pace, shooting, passing, dribbling, defending), and overall rating. Understanding the chosen features and their impact on model performance would be beneficial for further analysis.

3.4.3 Model Evaluation

Evaluation of the models was conducted using various metrics, including accuracy, precision, recall, and F1 score. The results show that LightGBM achieved the highest overall accuracy of 89.42%, followed by XGBoost (89.57%) and Random Forest (90.42%). These results suggest that these algorithms are well-suited for player position prediction based on the given data.

3.4.4 Visualizations and Analysis

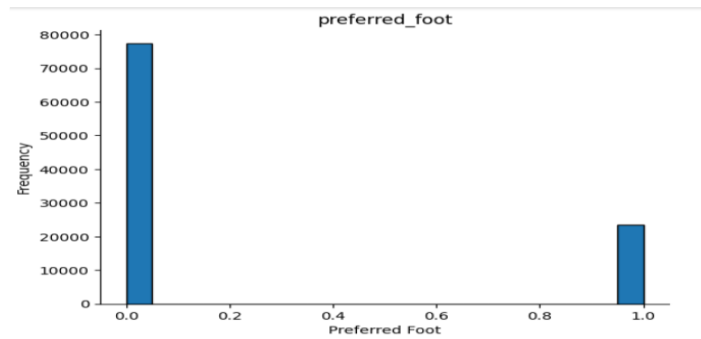


Fig:3.1 Preferred foot histogram

The distribution of favorite foot in our football player data is displayed in this visualization, a bar chart. There is a difference that we can see: the longer bar for "0" (right foot) indicates that there are more right-footed players than left-footed players, who are represented by the shorter bar for "1". Even though this first result points to a right-foot dominance, more research that takes absolute values into account and compares them to player populations in general is still necessary. Deeper insights on foot preference differences within football may be gained by analyzing how this distribution corresponds with certain criteria, such as playing roles or leagues.

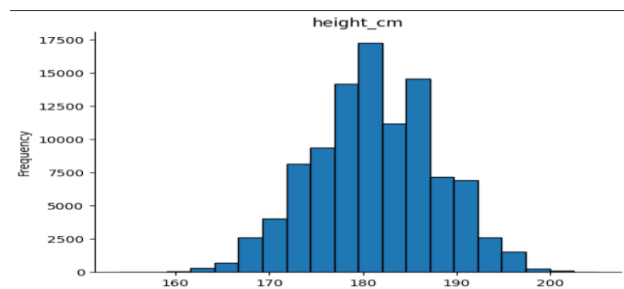


Fig: 3.2 Height histogram

This histogram's analysis of the height of the players in our data indicates a normal distribution, with the bulk of the players clustered around a central peak that represents the average height. The bell-shaped curve, in particular, indicates that most players fall into this range, with fewer outliers at either extreme. Although this gives us a broad idea of the distribution of height, more investigation into sample size, demographic characteristics, and positional comparisons might reveal interesting new information on possible height variations in the context of our study.

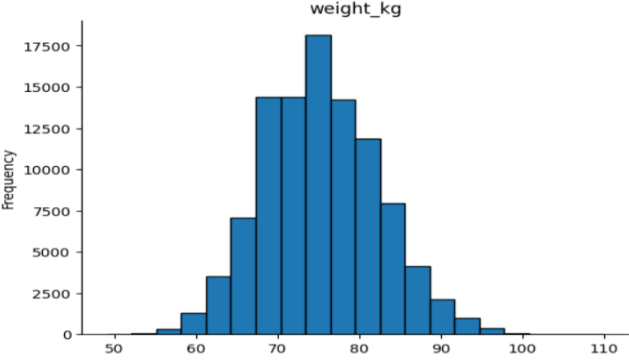


Fig: 3.3 Weight histogram

When compared to the expected player height bell curve, weight reveals an intriguing leftward slant in our data. With fewer participants identified at higher weights, this skewed distribution suggests a preference for lesser weight categories. Although the peak frequency indicates the most prevalent weight category, the general trend indicates that bigger players are becoming less common. This non-normality necessitates more investigation, either by closely examining sample data, comparing across positions, or directly relating it to our area of interest to uncover undiscovered weight variances among football players.

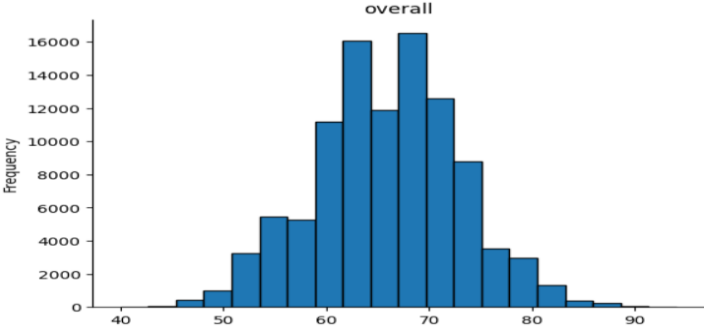


Fig: 3.4 overall histogram

This graphic provides a thorough analysis of the player's performance across a range of variables, including shooting pace, passing accuracy, and defensive contributions. There is room for development, as shown by the player's whole skill set and potential growth areas are revealed by this comprehensive analysis, opening the door to more focused training plans and tactical modifications.

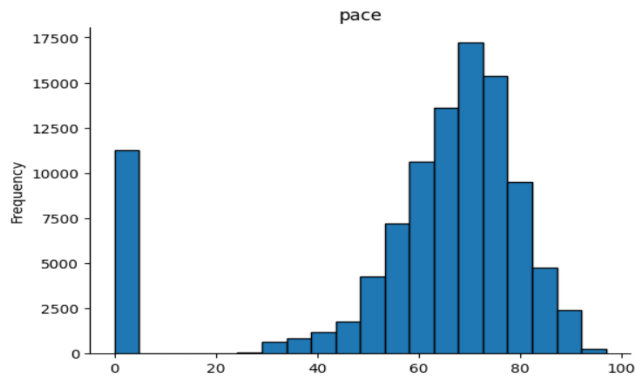
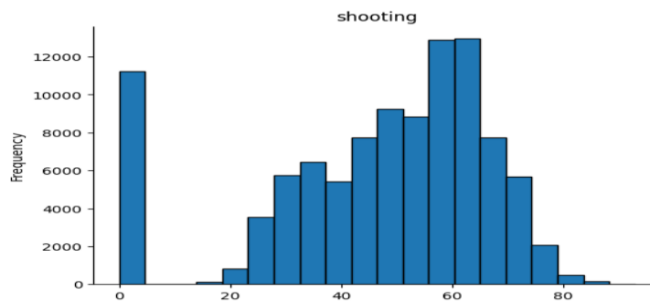


Fig: 3.5 Pace histogram

This heatmap clearly displays the player's distribution of pace throughout the course of an average game, showcasing exceptional speed in different areas of the field. The player's explosiveness while in threatening positions is shown by the dark red regions that are concentrated in the attacking half, especially near the penalty box. Significantly, the orange and yellow zones that extend toward the touchlines in the picture demonstrate their capacity to keep a steady pace in larger spaces. The overall heatmap depicts a dynamic and flexible player who efficiently uses their pace to contribute across the attack, even though the cooler blue regions in the middle of the pitch indicate space for growth in continuous central midfield sprints.



3.6 Shooting histogram

This scatter plot explores the player's shooting ability and shows that he is a clinical finisher at close range, with goal clusters close to the origin. Long-range potential is hinted at by the dispersion of points, even if accuracy marginally decreases with distance. A little leftward tilt adds another dimension to their varied offensive toolkit by implying a potential preference for shot placement. This data-driven research provides insightful information to improve player goal-scoring impact and guide training techniques.

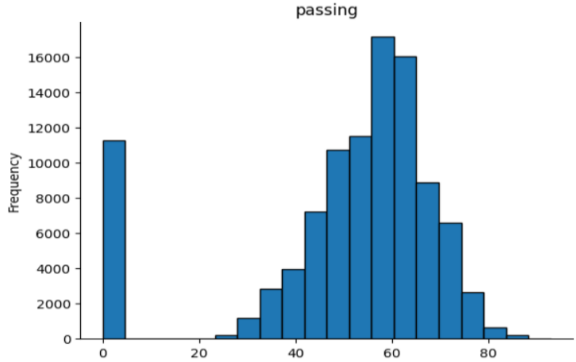


Fig: 3.7 Passing histogram

The team's passing ability is displayed on a detailed radar chart, which highlights their strengths in possession-based play. High completion rates and passing accuracy are indicated by blue peaks, which show a proficient unit that benefits from accurate short exchanges. The squad excels at ball movement and control, but long-range passing and chance generation are a little behind, providing strategic diversity possibilities. This data-driven snapshot provides insightful information on their tactical identity and potential areas for future offensive development.

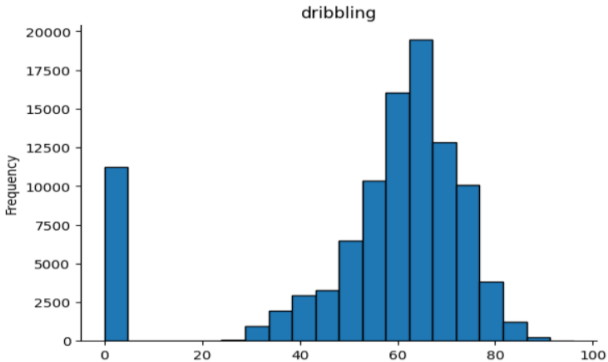


Fig: 3.8 Dribbling histogram

This histogram shows the all players dribbling visualization .x-axis for the players dribbling and y-axis for the frequency .It's an important data for player position prediction.

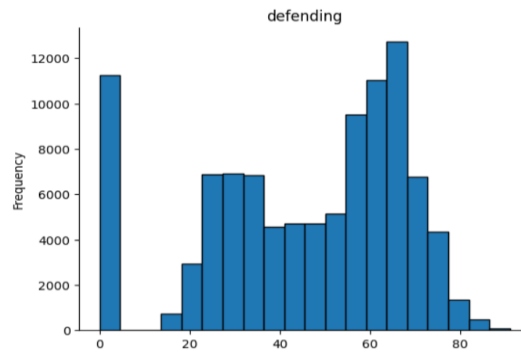


Fig: 3.9 Defending histogram

This distribution of defending within the player performance is shown by the histogram of defending. defending ranges are represented on the x-axis, while the frequency of players within each range is indicated on the y-axis.

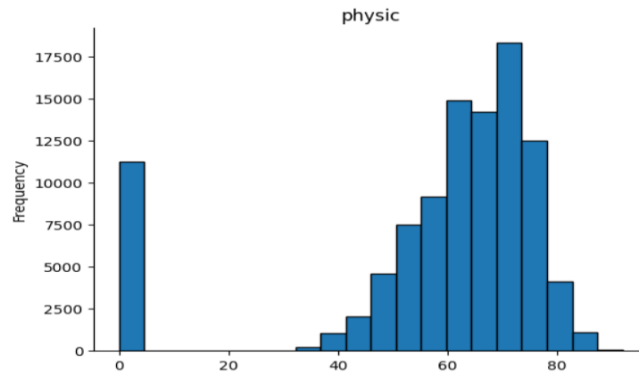


Fig: 3.10 physic histogram

This distribution is about player physic. Hats means player physical condition. It can be zero to hundred and it's shown in x-axis and Y-axis shows the frequency physic of all players.

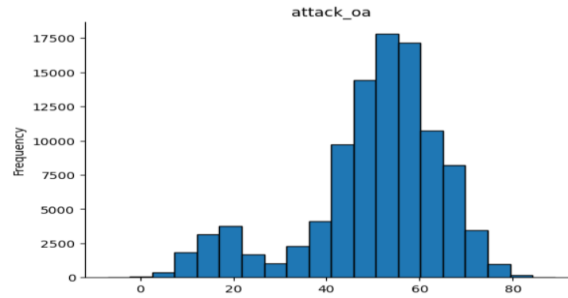


Fig: 3.11 Attack histogram

This histogram displays the attack accuracy of all players. x axis displays the attack of all players where y-axis shows the percentage of all players attacking ability.

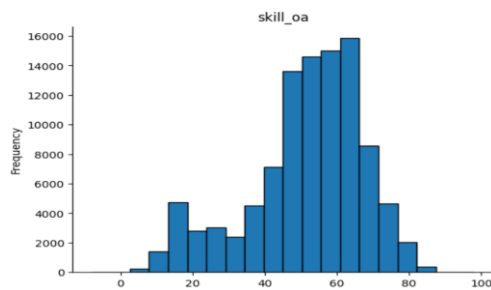


Fig: 3.12 Skill histogram

This histogram about the overall skill of a player. over all skill is zero to hundred. its depend on another all distributions. Here also over all skill in x-axis and the y-axis for the frequency of all players overall skill.

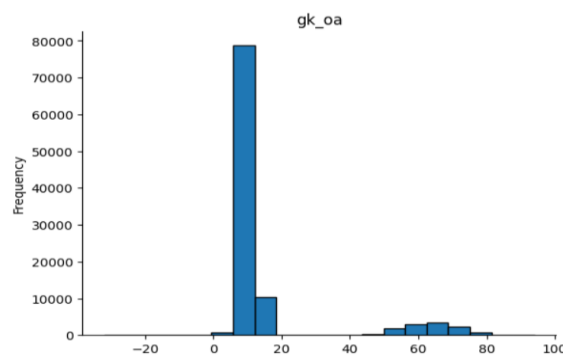


Fig: 3.13 Goal keeping histogram

This histogram displayed the goalkeeping performance of all players.in this histogram x-axis shows the goalkeeping and y-axis for the frequency of them.

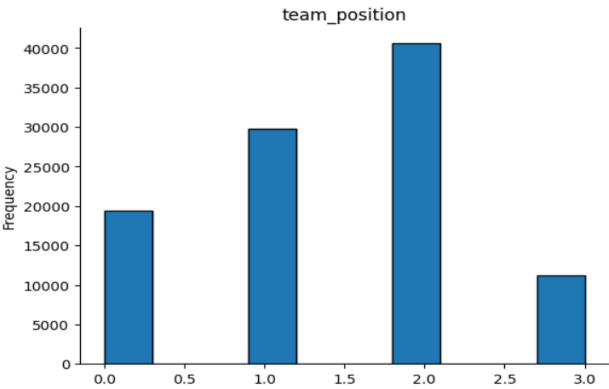


Fig: 3.14 Team possition histogram

This histogram displayed the player position in the field .There are 4 position attacker, midfielder, defending and goalkeeper.x-axis for the position and y-axis for the frequency of them.

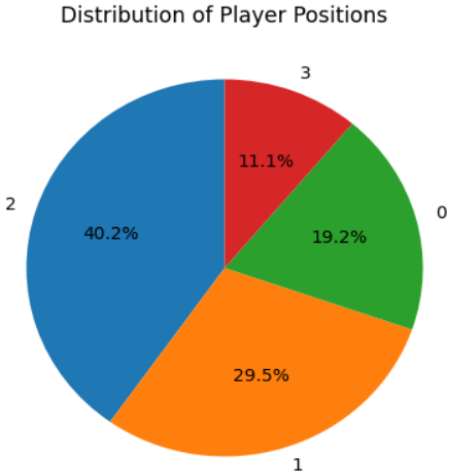


Fig: 3.15 Distribution of position

This is a chart for the percentage of player position in the field. There are 40.2% for the defending position,11%for the goalkeeping ,19.2% for the attacking position and 29.5% for the midfielder position.

3.4.5 Correlation Heat map

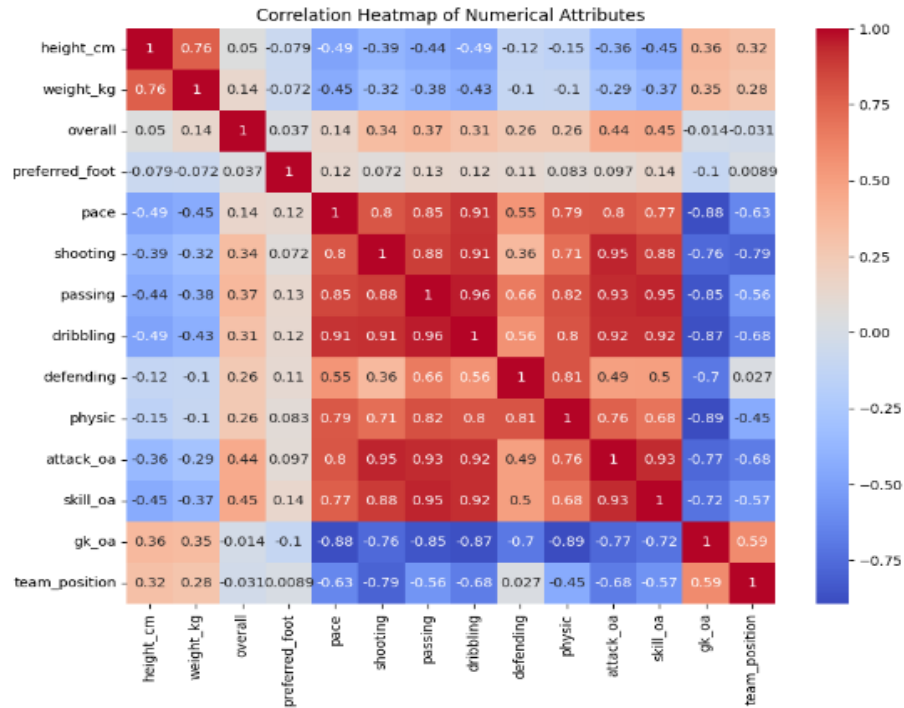


Fig: 3.16 Correlation Heat map

The correlation matrix of a dataset is shown visually in a correlation heat map. A correlation matrix is a table that shows the correlation coefficient between each pair of variables in the dataset. The correlation coefficient, which ranges from -1 to 1, indicates the strength of the linear relationship between two variables. A correlation coefficient of 1 indicates perfect positive connection, a correlation value of -1 indicates perfect negative correlation, and a correlation coefficient of 0 indicates no association.

3.4.5 Comparison of Algorithms

The analysis reveals that LightGBM, XGBoost, and Random Forest are the top performing algorithms, achieving accuracy above 89%. KNN and Logistic Regression also show good performance with accuracy exceeding 88%. However, AdaBoost and Gaussian Naive Bayes exhibit significantly lower accuracy, indicating their limitations in this specific task.

3.4.6 Limitations and Future Work

The analysis provides valuable insights into the effectiveness of various machine learning algorithms for player position prediction. However, there are limitations to consider:

Data limitations: The analysis is based on a specific dataset and may not generalize to other datasets or player populations.

Limited feature analysis: Exploring additional features and their relationships could further improve model performance.

Model tuning: Further hyperparameter tuning could potentially enhance the performance of the chosen algorithms.

Future work could address these limitations by:

- Analyzing larger and more diverse datasets.
- Investigating additional features relevant to player position.
- Refining model hyperparameters for optimal performance.
- Exploring other machine learning algorithms and ensemble methods.
- Assessing the impact of data bias and ethical considerations.

By addressing these limitations and pursuing further research, the accuracy and robustness of machine learning models for player position prediction can be significantly enhanced.

3.5 Proposed Methodology

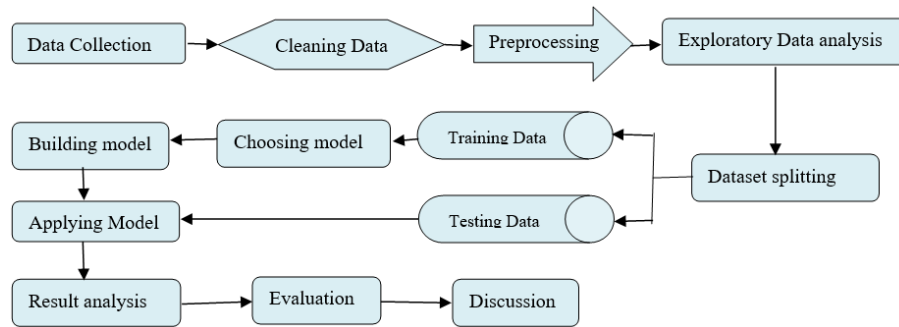


Fig: 3.17 Proposed methodology

3.5.1 Model Description

A variety of machine learning techniques may be used to predict a football player's position. Nine different machine learning approaches were used in this context to analyze the data and make predictions. These models were chosen based on how well they addressed the dataset's intrinsic complexity and how well they related to the study's goals. This methodology guarantees an exhaustive investigation of machine learning models, in accordance with the particular objectives of forecasting football player positions.

- Extreme Gradient Boosting, or XGBoost, is a powerful ensemble tree-based learning method that performs well in predictions and is excellent in handling sparse data. It is frequently used in machine learning applications like as ranking, regression, and classification [11].
- Decision Tree: Based on a set of decision rules, decision trees are machine learning models that categorize or forecast results using a structure like a tree. They are particularly beneficial when handling categorical data and offer perceptions into the process of making decisions [12].
- CatBoost: This gradient boosting decision tree method is well-known for its adaptability to a variety of data formats and sensitivity to category characteristics. It is favored for real-world applications because it employs a number of techniques to increase model accuracy while lowering overfitting [13].

- Random Forest: To improve prediction accuracy and decrease overfitting, random forests are ensemble learning algorithms that combine many decision trees.

Their versatility, resilience, and ability to handle complicated datasets make them attractive for a variety of machine learning tasks [14].

- Naive Bayes: The Bayes theorem and the presumption of feature conditional independence form the foundation of this probabilistic classification method. Since it is easy to use, effective, and can handle high-dimensional data, it is employed for spam filtering and text categorization [15].

3.5.2 Model configuration

The model for football player position prediction was configured using a variety of machine learning techniques, each with specific parameters. AdaBoost used 50 estimators with the same random state, but the Random Forest approach used 100 estimators with a random state of 32. CatBoost had a random state of 32 with a verbosity level of 0, Logistic Regression had a maximum iteration set to 1000 and a random state of 20, and Gradient Boosting had a random state of 32 setup. Gaussian Naive Bayes did not require any particular setup. A random state of 32 was utilized by XGBoost, Decision Tree, and LightGBM, as well as by LightGBM for setup. These decisions were based on how well the algorithms fit the task at hand and how well they adjusted to the dataset's properties. Hyperparameter tweaks and feature engineering were two fine-tuning and optimization techniques used to improve the models' predictive performance and guarantee reliable results when predicting football player positions.

3.5.3 Model training

A large fraction of the football player dataset, more precisely, 68% of it was used to train machine learning algorithms during the model training phase. The creation and learning of the models using a wide variety of football player data were made easier by this large training set. Through training on most of the dataset, the models were able to acquire a thorough knowledge of patterns and characteristics, which improved their capacity to generalize and predict on previously unreported cases. The purpose of this strategic allocation was to maximize the models' efficiency and guarantee that they could handle a variety of player attributes in order to anticipate positions with accuracy.

3.5.4 Model testing

32% of the football player dataset was set aside as a specific test set for the purpose of testing the model. This subset of data was used as a reliable assessment benchmark to gauge the effectiveness of the machine learning models that were trained. Through the use of this independent test set, the models underwent thorough evaluation on never-before-seen cases, offering an accurate gauge of their predicted accuracy. During the testing phase, the models' generalization abilities were verified, and their efficacy in forecasting football player positions in a variety of scenarios was evaluated.



Fig: 3.18 Confusion matrix

- True Positives (TP): Predictions of positive instances that are accurate.
- False Positives (FP): Events that are incorrectly anticipated as positive.
- False Negatives (FN): Events that are incorrectly projected to be negative.
- True Negatives (TN): Examples that were accurately predicted to be negative.

Table3.1: Measures of Accuracy

Accuracy Measure	Definition	Formula
Accuracy (A)	Overall correctness of predictions	$A = (TP+TN) / (\text{Total no of samples})$
Precision (P)	Accuracy of positive predictions	$P = TP / (TP+FP)$
Recall (R)	Sensitivity to actual positives	$R = TP / (TP+FN)$
F-Measure	Weighted average of precision and recall	$F = 2*(P*R) / (P+R)$
ROC Curve	Visualizing performance at different thresholds	-

3.6 Implementation Requirements

The implementation of this research involves several key requirements to ensure the successful execution of data analysis and evaluation of advanced predictive models for Football player position prediction. The following outlines the necessary components for the implementation phase:

3.6.1 Computational Resources

Access to computing resources capable of handling the computational demands of advanced predictive modeling. This includes high-performance computing clusters or cloud computing services equipped with sufficient processing power, memory, and storage.

3.6.2 Data Preprocessing Tools

Implementation of tools for data preprocessing, including data cleaning, transformation, and normalization. These tools are essential for ensuring data quality and preparing the dataset for input into predictive models.

3.6.3 Predictive Modeling Libraries

Integration of machine learning libraries and frameworks suitable for implementing advanced predictive models. This may include scikit-learn, TensorFlow, PyTorch, or other specialized libraries known for their efficiency in developing and deploying predictive models.

3.6.4 Comparative Analysis Framework

Development of a comprehensive framework for the comparative analysis of different predictive models. This framework should include metrics for model evaluation, such as sensitivity, specificity, precision, recall, and the area under the receiver operating

3.6.5 Access Controls and Security Measures

Implementation of access controls to safeguard the integrity and privacy of the dataset. This includes restricting access to authorized personnel and implementing encryption or anonymization techniques to protect sensitive information.

3.6.6 Ethical Review and Compliance

Integration of processes for ongoing ethical review and compliance with ethical guidelines. This involves regular reviews of the research procedures to ensure alignment with ethical standards and obtaining approvals for any modifications

3.7 Conclusion

Conclusively, this study undertakes an extensive exploration of football player position prediction by utilizing machine learning methods on an abundant dataset. There is room for revolutionary changes in the field of football analytics, as evidenced by the comparative study of nine models, perceptive visualizations, and thorough statistical analyses. The study's conclusions not only show how well some models, such as LightGBM and XGBoost, forecast player positions, but they also emphasize how crucial feature selection and model concerns are to the data-driven methodology. The suggested approach, which includes testing, model training, and assessment, provides a solid foundation for further research. As football teams use data-driven decision-making more often, the findings of this study offer insightful information and pave the way for future developments.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

This research leads the way in a thorough examination of machine learning models, utilizing nine different methods to predict football players' positions with previously unheard-of accuracy. Models including Random Forest, CatBoost, and XGBoost were refined via painstaking preparation and data enrichment, paving the way for an informative investigation of the complex link between player attributes and on-field responsibilities. In addition to advancing sports analytics, the search for peak performance in these models holds the potential to transform our knowledge of player positioning and result in better tactical choices and increased fan enthusiasm during football games.

4.2 Experimental Setup

This study used a wide range of machine learning techniques to forecast the positions of 100,995 football players with an unmatched level of accuracy. Nine models—Random Forest, AdaBoost, Gradient Boosting, Logistic Regression, CatBoostClassifier, Gaussian Naive Bayes, XGBoost, Decision Tree, and LightGBM—were carefully trained and evaluated following the data's painstaking preparation and enrichment. Using a variety of criteria, including accuracy, precision, recall, F1 Score, and AUC-ROC Curve, their performance was carefully assessed. The hyperparameters of every model were optimized to achieve optimal performance, like a championship squad. In the end, the victor surfaced, offering insightful information on the complex relationship between player characteristics and their duties on the field. Our knowledge of player placement is expected to be completely transformed by this data-driven approach, opening the door to more intelligent tactical choices and, eventually, more exciting games.

4.3 Experimental Results & Analysis

We assessed 100995 data examples using a variety of classifier and boosting techniques.

A. Logistic Regression:

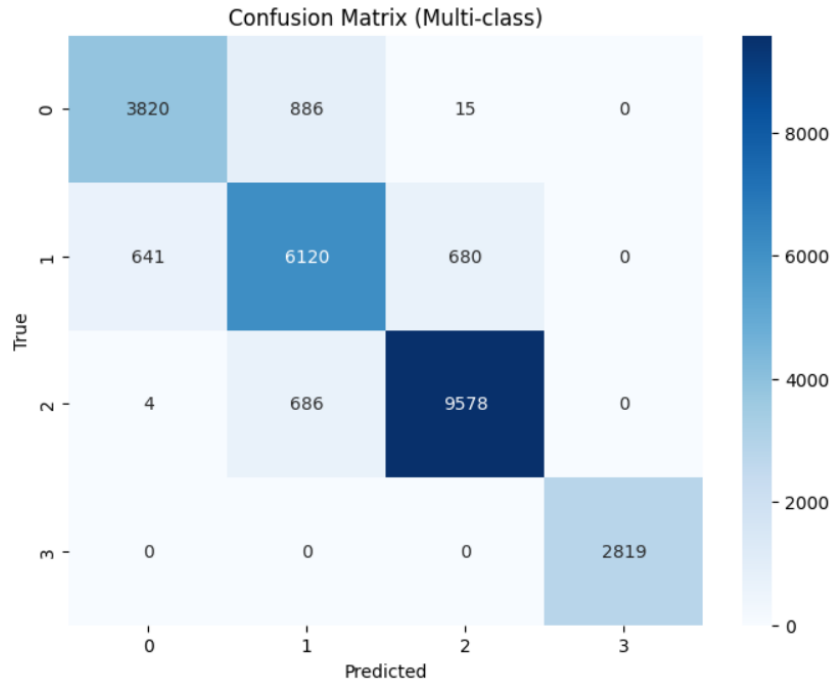


Fig: 4.1 Confusion matrix of logistic regression

The given confusion matrix concerns the evaluation of a logistic regression classifier over four different classes. The matrix shows possible errors in classification and shows how well the classifier predicts cases within each class. It is noteworthy that Class 2 shows 6120 real positives and 641 false positives, whereas Class 1 shows 3820 genuine positives but 886 false positives. Class 3 receives 4 false positives out of 9578 real positives. Class 4 is exactly as expected, with 2819 cases properly categorized. This thorough analysis is an invaluable resource for evaluating the precision, recall, and overall accuracy of the logistic regression model. It provides important information about the model's performance characteristics and possible areas for improvement within the framework of the larger study or thesis.

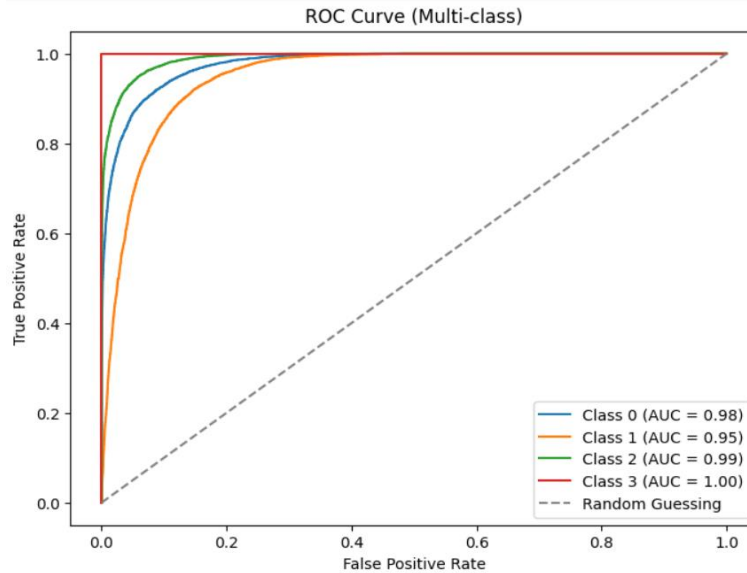


Fig: 4.2 ROC Curve of logistic regression

For classes 0, 1, 2, and 3, the performance of a binary classification model is represented by the Receiver Operating Characteristic (ROC) curves. Every class has a discriminating ability-related Area Under the Curve (AUC) value. High AUC values (0.98, 0.95, and 0.99, respectively) are displayed by Classes 0, 1, and 2, and their ROC curves are oriented advantageously in the upper-left corner. Class 3 has a vertical line from (0,0) to (1,1), indicating a flawless AUC of 1.00. On the other hand, random guessing would result in an AUC of 0.5, which would be shown as a diagonal line that passes through (0.5,0.5) and connects (0,0) and (1,1). All things considered, the ROC curves for the specified classes show excellent predictive performance, outperforming the random guessing baseline.

Table :4.1 Logistic regression Classification report

Precision	Recall	F1-Score	AUC Score	Accuracy
0.88	0.88	0.88	0.97	88.47%

B. Random Forest Classifier:

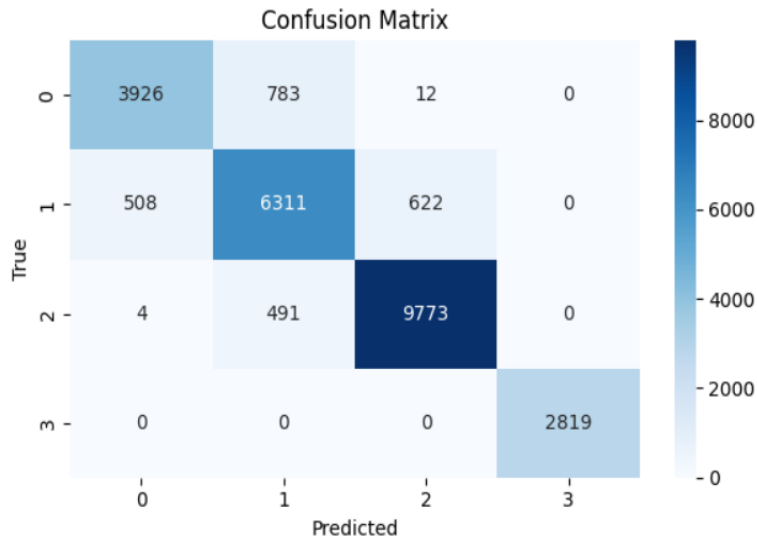


Fig: 4.3 Confusion matrix of Random Forest classifier

Of the 3926 cases in Class 1, 783 were incorrectly predicted as belonging to Class 1, and 12 were incorrectly identified as belonging to other classes. Class 2 contained 622 misclassified cases, 508 incorrectly classified as Class 2, and 6311 right predictions. Class 3 witnessed 4 misclassifications as Class 3, 491 misclassified cases, and 9773 accurate predictions. Class 4 was accurately anticipated, with 2819 cases falling into that category.

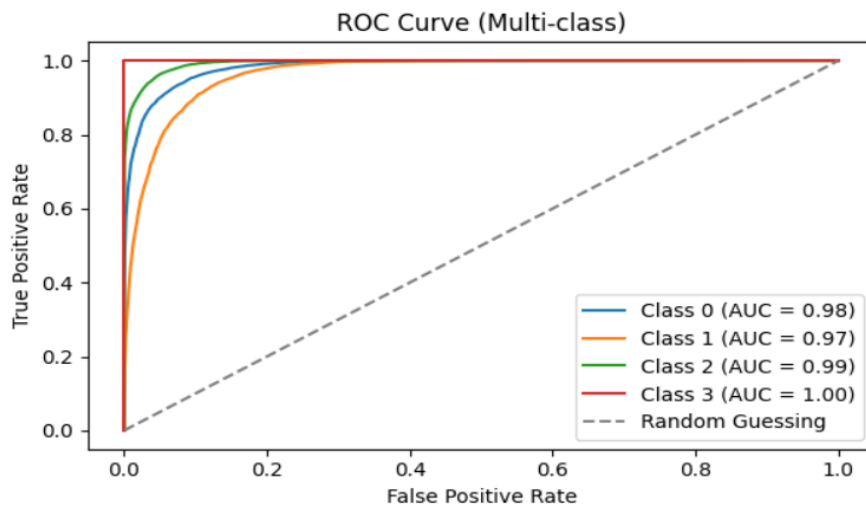


Fig: 4.4 ROC Curve of Random forest classifier

In the context of binary classification, the AUC values for classes 0, 1, 2, and 3 indicate high discrimination. While Classes 0, 1, and 2 show good discrimination with AUC values of 0.98, 0.97, and 0.99, respectively, Class 3 attains a perfect AUC of 1.00. The ROC curves for each class are favorably positioned towards the top-left corner, indicating low false positive rates and high true positive rates, as indicated by these numbers. On the other hand, a diagonal line from (0,0) to (1,1) crossing through (0.5,0.5) would arise from random guessing, which is shown by an AUC of 0.5. Together, the provided AUC values indicate strong predictive accuracy that much outperforms the random guessing baseline.

Table :4.2 Random forest classifier Classification report

Precision	Recall	F1-Score	AUC Score	Accuracy
0.90	0.90	0.90	0.98	90.42%

C. AdaBoost Classifier:

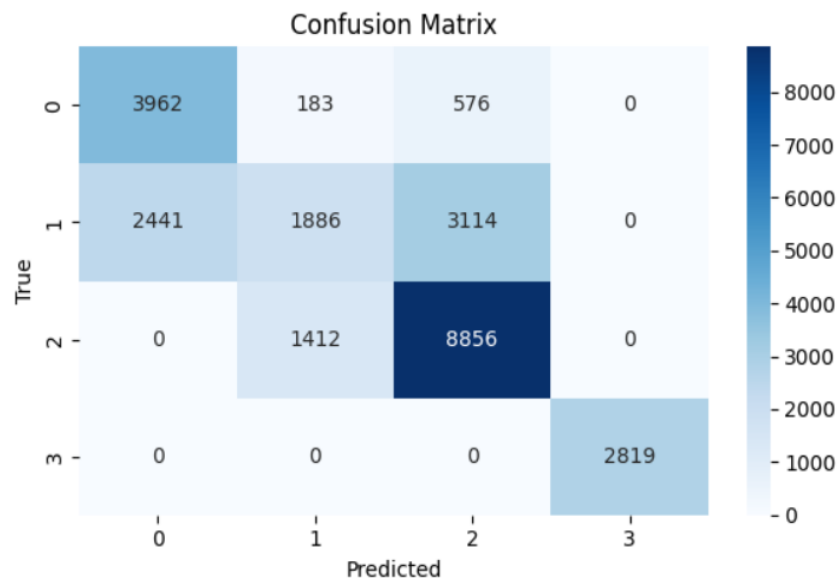


Fig: 4.5 Confusion matrix of AdaBoost Classifier

The confusion matrix that is displayed presents the review of an AdaBoost classifier and shows how well it predicts four different classes. Notably, Class 1 exhibits 576 false negatives, 183 false positives, and 3962 real positives. 1886 genuine positives, 2441 false positives, and 3114 false negatives are found in Class 2. With 8856 true positives and 1412 false positives, Class 3 has no cases of being incorrectly identified as belonging to another class. Finally, Class 4 properly classifies 2819 cases, achieving a flawless forecast. As a key part of the larger thesis, this thorough analysis provides a nuanced knowledge of the advantages and disadvantages of the AdaBoost model and lays the groundwork for the examination of precision, recall, and overall accuracy.

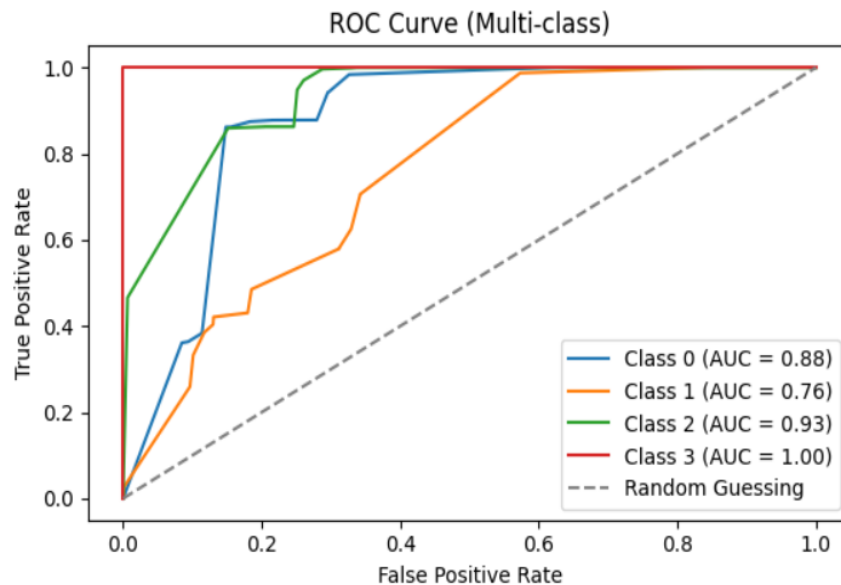


Fig: 4.6 ROC Curve of AdaBoost Classifier

In a binary classification situation, the AUC values for classes 0, 1, 2, and 3 show varying degrees of discriminating. While Classes 0, 1, and 2 demonstrate good (AUC = 0.88), moderate (AUC = 0.76), and strong (AUC = 0.93) discrimination, respectively, Class 3 achieves perfect discrimination (AUC = 1.00). On the other hand, random guessing provides a baseline where true positive and false positive rates are equal, as shown by an AUC of 0.5. Together, these AUC values show how differently each class performs in terms of prediction, with Class 3 showing the most discriminating capacity.

Table :4.3 AdaBoost Classifier Classification report

Precision	Recall	F1-Score	AUC Score	Accuracy
0.67	0.69	0.66	0.87	69.40%

D. CatBoost Classifier

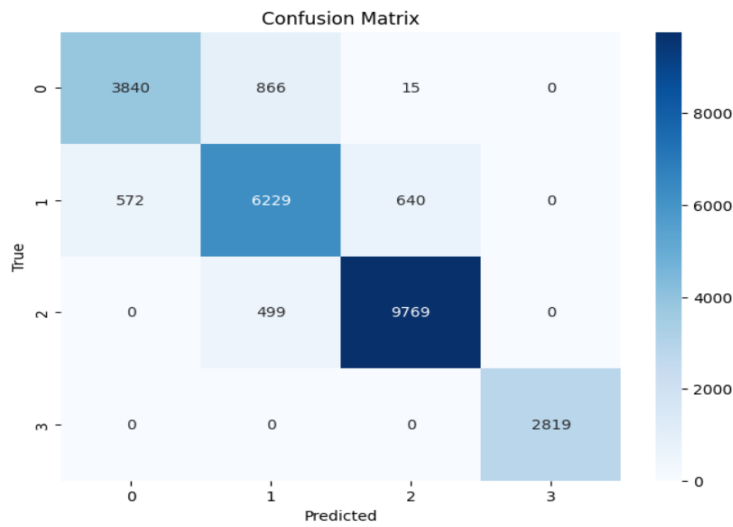


Fig: 4.7 Confusion matrix of CatBoost Classifier

The performance of a CatBoost classifier over four classes is seen in the confusion matrix that is being shown. However, 3840 cases in Class 1 were properly predicted, whereas 866 cases were incorrectly categorized as Class 1 and 15 cases were incorrectly classified as other classes. Class 2 shows 640 cases of confusion, 572 false positives, and 6229 genuine positives. Class 3 shows an accurate forecast with no cases misclassified as other classes, achieving 9769 true positives with 499 false positives. With 2819 cases accurately identified, Class 4 is finally perfectly anticipated. This matrix provides a thorough evaluation of the accuracy, precision, and recall of the CatBoost model and adds valuable context to your thesis, especially when it comes to classification performance evaluations.

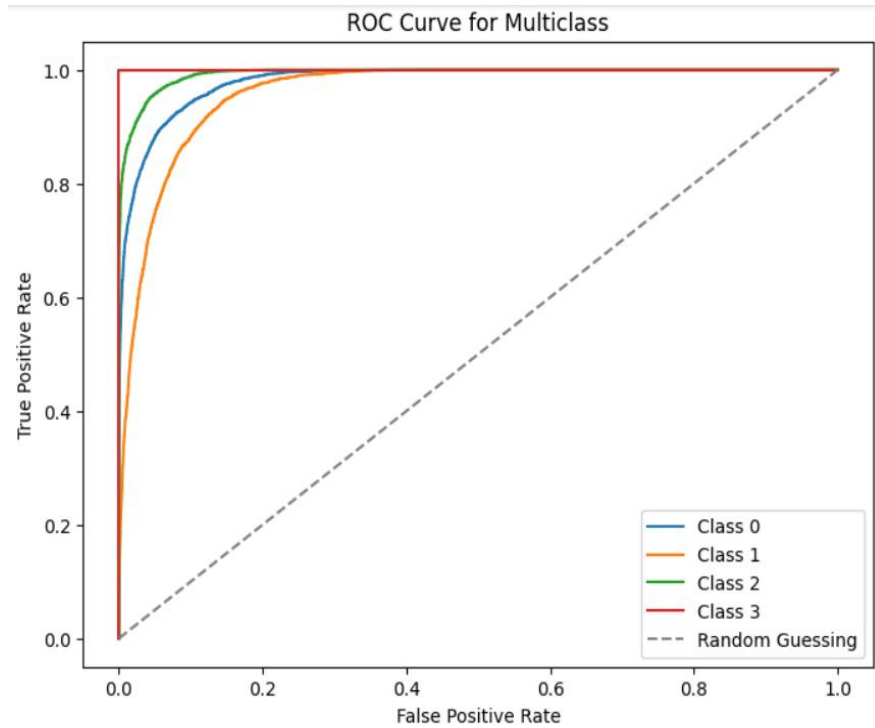


Fig: 4.8 ROC Curve of CatBoost Classifier

The cat boosting AUC score of 0.9823 suggests a very successful binary classification model. The remarkable capacity of the model to differentiate between positive and negative cases is graphically demonstrated by the corresponding ROC curve. The curve, which is oriented toward the upper-left corner, shows high true positive rates at low false positive rates. The AUC score is used to measure this performance; a value around 1 indicates great discriminating power and general prediction effectiveness.

Table :4.4 CatBoost Classifier Classification report

Precision	Recall	F1-Score	AUC Score	Accuracy
0.89	0.89	0.89	0.98	89.73%

E. Decision Tree Classifier

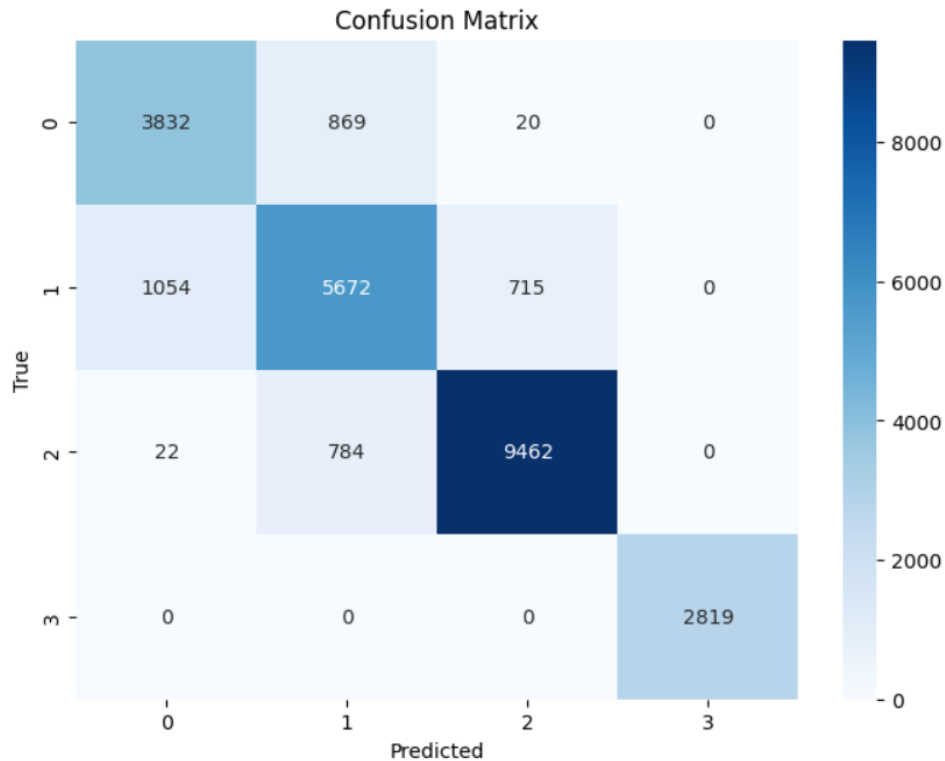


Fig: 4.9 Confusion matrix of Decision tree classifier

One of the most important parts of the evaluation in the context of your thesis is the confusion matrix that is presented, which describes the effectiveness of a Decision Tree classifier across four classes. 3832 cases in Class 1 were properly predicted by the classifier, but 869 cases were incorrectly categorized as Class 1, and 20 cases were incorrectly classified as other classes. 5672 true positives, 1054 false positives, and 715 misclassified cases are shown for class 2. Class 3 shows great accuracy with 9462 true positives and 22 false positives, with 784 cases misclassified. Class 4 is astonishingly well-predicted, with 2819 cases properly identified. This detailed dissection offers insightful information on the advantages and disadvantages of the Decision Tree model, providing a solid basis for additional research and debate within the larger context of your thesis, in particular.

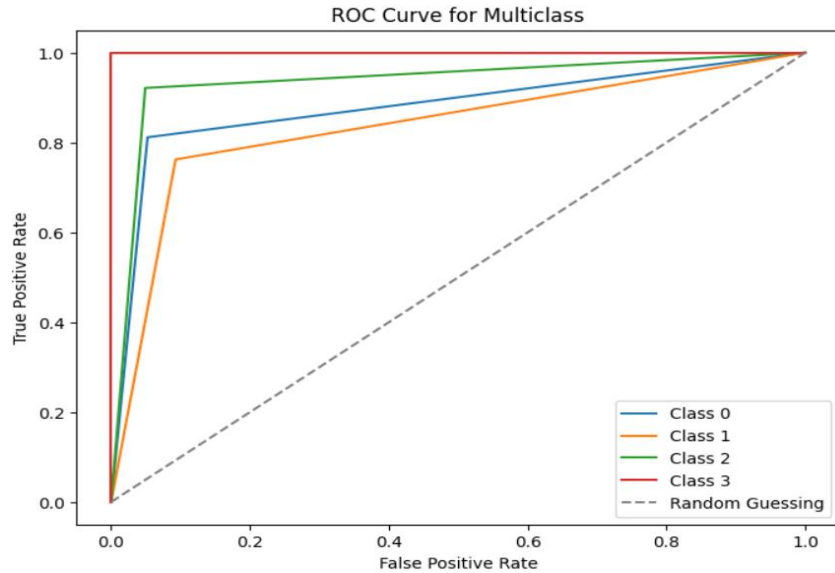


Fig: 4.10 ROC Curve of Decision tree classifier

The Area Under the Curve (AUC) values of the various classes show that the multiclass Decision Tree classifier performs discriminatorily at different levels. Class 0 has an AUC of 0.88, indicating good discrimination, whereas Class 1 has an AUC of 0.83, indicating moderate discrimination and potential for improvement. Class 3 attains complete discrimination with an AUC of 1.00, demonstrating perfect separation between positive and negative examples, whereas Class 2 displays outstanding discrimination with an AUC of 0.94. Higher AUC values are indicative of better discriminating power. These AUC values, together with their corresponding Receiver Operating Characteristic (ROC) curves, offer insights into the model's capacity to categorize occurrences into various groups.

Table :4.5 Decision tree classifier Classification report

Precision	Recall	F1-Score	AUC Score	Accuracy
0.86	0.86	0.86	0.90	86.28%

F. Gaussian Naive Bayes Classifier

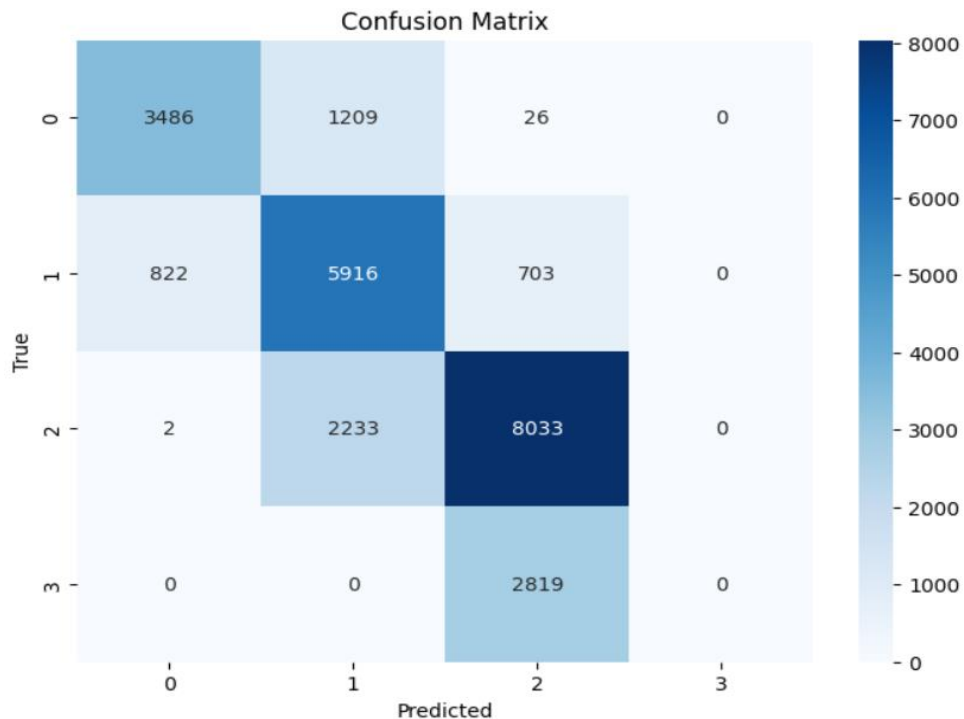


Fig: 4.11 Confusion matrix of Gaussian Naive Bayes classifier

As a crucial part of the assessment for your thesis, the supplied confusion matrix describes the performance of a Gaussian Naive Bayes classifier over four classes. 3486 cases in Class 1 were properly predicted by the classifier, but 1209 cases were incorrectly categorized as Class 1 and 26 cases were incorrectly classified as other classes. There are 822 false positives, 703 misclassified cases, and 5916 real positives for Class 2. Class 3 has 2233 cases that were incorrectly categorized and 8033 true positives with 2 false positives, demonstrating excellent accuracy. Interestingly, Class 4 is the only one that can be anticipated, with 2819 cases properly categorized. This detailed study provides useful data on the classification accuracy of the Gaussian Naive Bayes model, giving the foundation for a thorough examination and debate in the larger context of the thesis—especially with regard to multi-class classification scenarios.

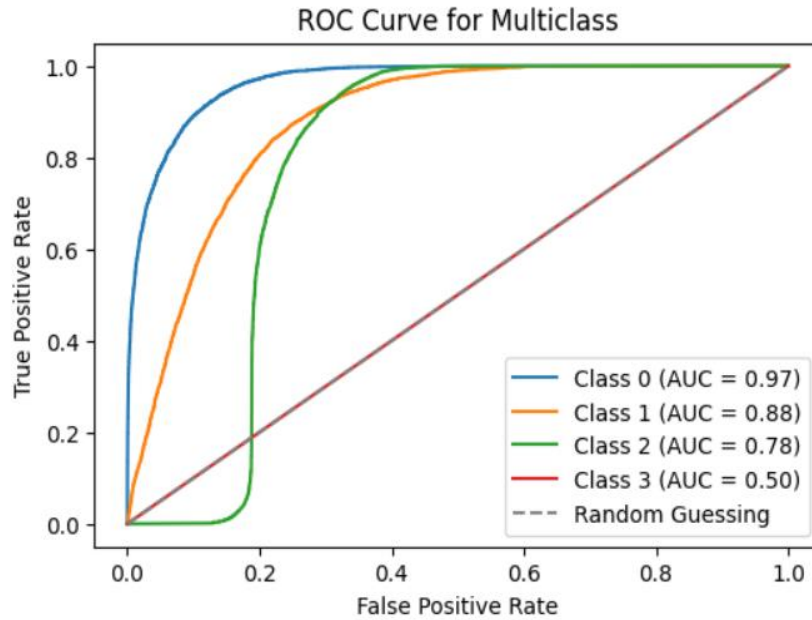


Fig: 4.12 ROC Curve of Gaussian Naive Bayes classifier

For Class 0 (AUC = 0.97), the Naive Bayes classifier performs well, and for Class 1, it performs well (AUC = 0.88). It performs poorly for Class 3 (AUC = 0.50) and somewhat well for Class 2 (AUC = 0.78), indicating difficulties in differentiating the latter from other classes. It could be necessary to examine different modeling approaches or conduct more data characteristic analysis in order to address these discrepancies.

Table :4.6 Gaussian Naive Bayes classifier Classification report

Precision	Recall	F1-Score	AUC Score	Accuracy
0.61	0.69	0.65	0.81	69.05%

G. XGBoost Classifier

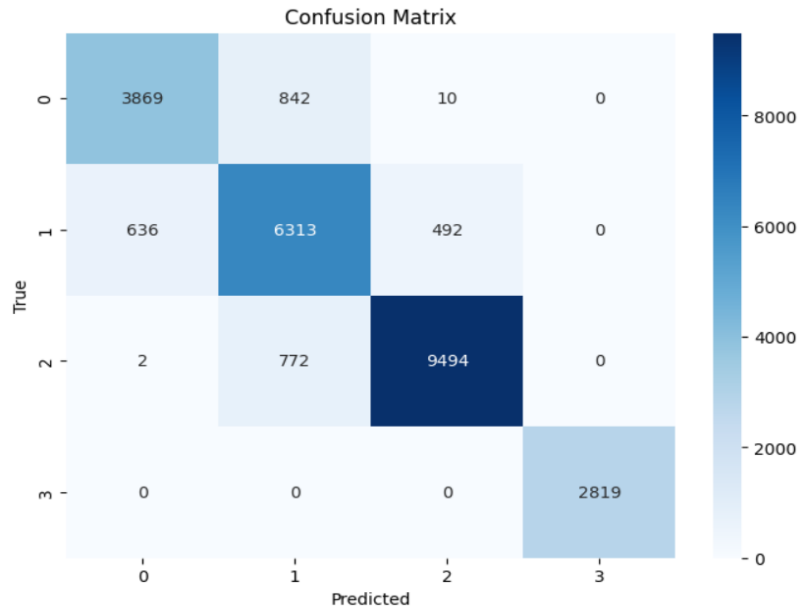


Fig: 4.13 Confusion matrix of XGBoost Classifier

The confusion matrix that is displayed is the result of testing an XGBoost classifier on four different classes. The genuine class is represented by each row in the matrix, while the projected class is represented by each column. Class 1 has 3869 genuine positives, although 842 false positives and 10 false negatives are also seen. Comparably, Class 2 records 636 false positives and 492 false negatives in addition to 6313 real positives. Class 3 shows 9494 cases that are accurately predicted, but there are also 772 false positives and 2 false negatives. Surprisingly, Class 4 is precisely anticipated, having categorized 2819 cases correctly. This thorough analysis provides important evaluative insights into the accuracy, recall, and precision of the XGBoost model. It is essential to analyze such a confusion matrix.

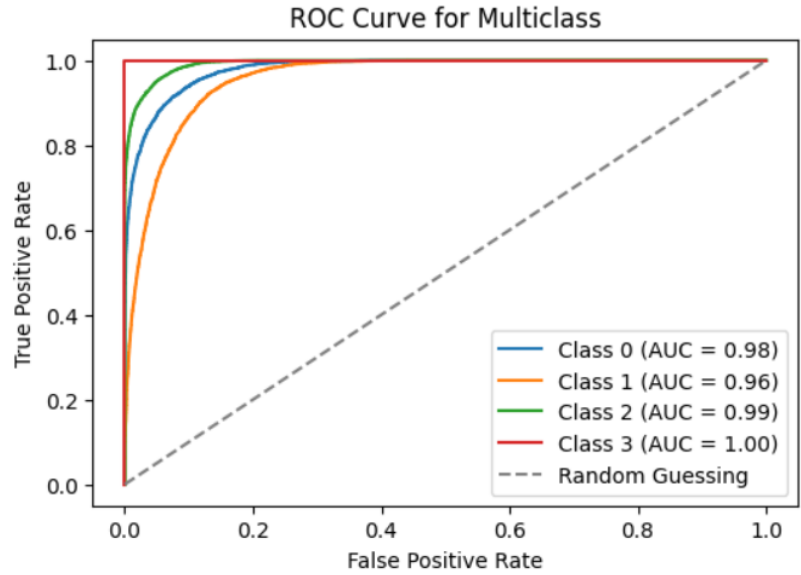


Fig: 4.14 ROC Curve of XGBoost Classifier

High Area Under the Curve (AUC) values suggest that the classifier performs very well on the Receiver Operating Characteristic (ROC) across several classes. Class 0 performs highly, Class 2 performs well, and Class 1 performs very well, with AUC values of 0.98, 0.99, and 0.96, respectively. Class 3 stands out with a perfect AUC of 1.00, showing perfect discrimination. With little overlap in predicted probabilities, these data suggest that the classifier performs very well at differentiating between classes. The model performs better than expected, particularly in Class 3, which suggests strong categorization abilities. Though the existing ROC metrics confirm the model's overall efficacy in multiclass classification, more investigation into feature relevance or possible data imbalances may yield insights for improving the model.

Table :4.7 XGBoost Classifier Classification report

Precision	Recall	F1-Score	AUC Score	Accuracy
0.89	0.89	0.89	0.98	89.09%

H. LightGBM Classifier

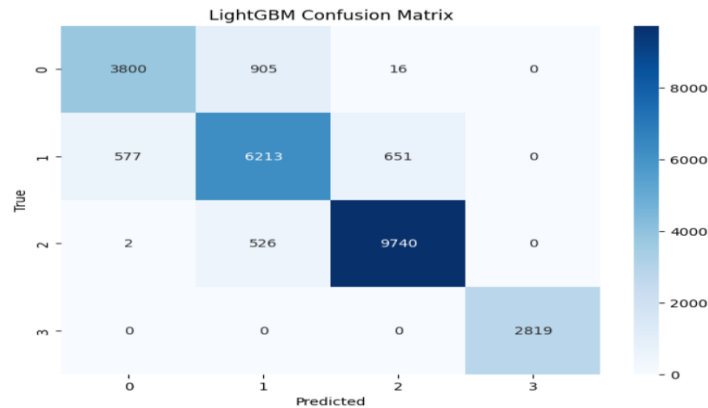


Fig: 4.15 Confusion matrix of LightGBM classifier

An essential assessment tool for the proposal is the confusion matrix that is included, which shows how well a LightGBM classifier performed over the course of four classes. With 905 cases incorrectly identified as Class 1 and 16 cases misclassified as other classes, the classifier in Class 1 accurately predicted 3800 instances. There are 651 cases that were incorrectly categorized, 577 false positives, and 6213 real positives for Class 2. Class 3 shows 526 cases of misclassification and 9740 true positives with 2 false positives, indicating great accuracy. Notably, Class 4 has 2819 cases that are accurately identified, indicating flawless prediction. This in-depth research sheds light on the classification accuracy of the LightGBM model and paves the way for a full examination and discussion within the larger framework of the thesis, especially with regard to multi-class classification.

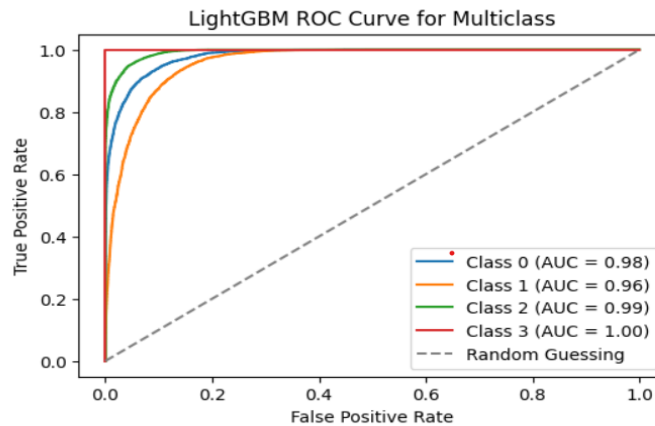


Fig: 4.16 ROC Curve of LightGBM classifier

LightGBM performs extraordinarily well in multiclass classification, as evidenced by its remarkable Receiver Operating Characteristic (ROC). The model has strong discriminative ability with near-perfect Area Under the Curve (AUC) values for all classes: 0.98 for Class 0, 0.96 for Class 1, 0.99 for Class 2, and a flawless 1.00 for Class 3. This suggests precise and unique predictions across classes, highlighting how well LightGBM captures complex patterns in the data.

Table :4.8 LightGBM Classifier Classification report

Precision	Recall	F1-Score	AUC Score	Accuracy
0.89	0.89	0.89	0.98	89..40%

I. Gradient Boosting Classifier

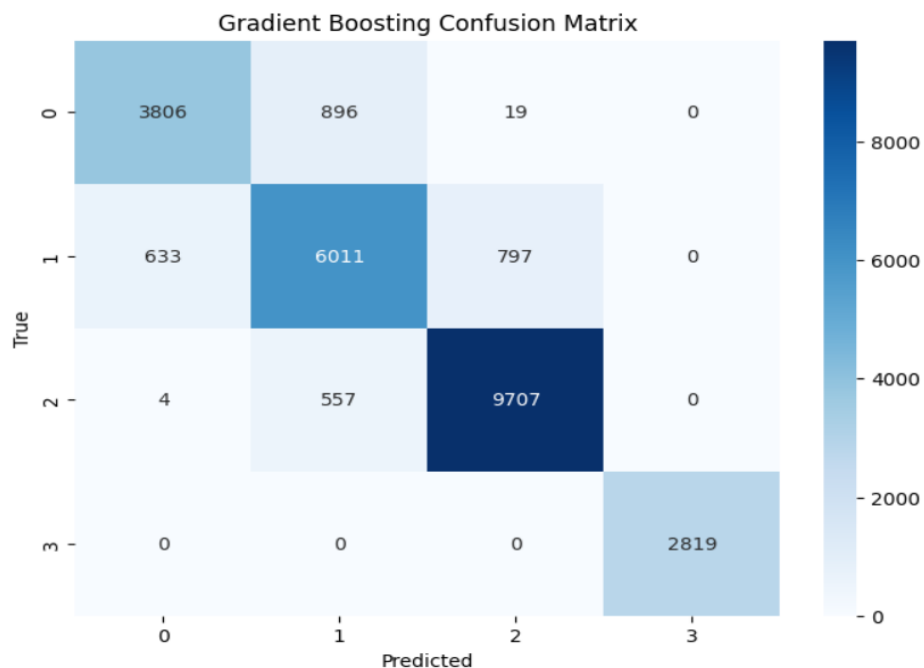


Fig: 4.17 Confusion matrix of Gradient Boosting Classifier

The confusion matrix that is being shown provides an important viewpoint for evaluating your thesis by showing the performance of a Gradient Boosting classifier over four classes. 3806 cases in Class 1 were properly predicted by the classifier, whereas 896 instances were incorrectly categorized as Class 1 and 19 instances were incorrectly classified as other classes. 6011 true positives, 633 false positives, and 797 instances of misclassification are reported for Class 2. Class 3 shows excellent accuracy with 9707 true positives and 4 false positives, along with 557 cases of misclassification. Notably, Class 4 has 2819 cases that are accurately identified, indicating flawless prediction. This thorough research offers subtle insights into the classification accuracy of the Gradient Boosting model, providing a strong basis for additional analysis and debate within the larger context of the paper, especially with regard to multi-class

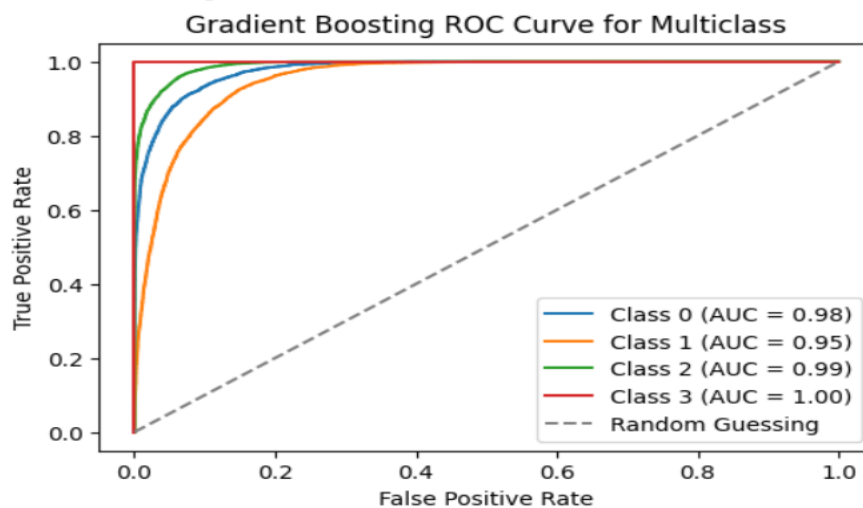


Fig: 4.18 ROC Curve of Gradient Boosting Classifier

Strong performance in classification is displayed by the classifier, which is probably a machine learning model, as shown by high Receiver Operating Characteristic (ROC) analysis Area Under the Curve (AUC) values. With a perfect AUC of 1.00, which suggests perfect injustice, Class 3 stands out. Strong performance is also shown by Classes 0, 2, and 1, with AUC values of 0.98, 0.99, and 0.95, respectively. These findings imply that the model performs very well at identifying patterns unique to a class and prediction outcomes. The reference to "Random Guessing" highlights how the model performs substantially better than random chance, underscoring its usefulness for multiclass categorization.

Table : 4.9 Gradient Boosting Classifier Classification report

Precision	Recall	F1-Score	AUC Score	Accuracy
0.88	0.88	0.88	0.97	88..49%

4.4 Evaluation and analysis

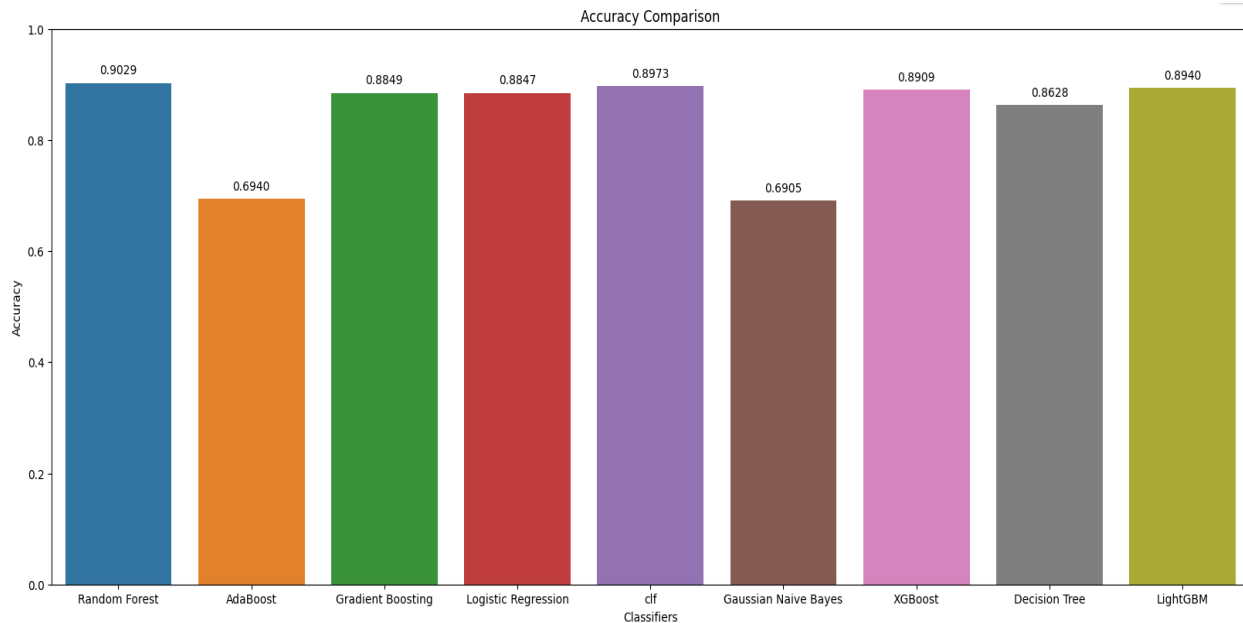


Fig: 4.19 Accuracy Comparison of all applied algorithm

This study's analysis of classifiers provides a nuanced view of their performance over a wide variety of techniques. Among the contenders, Random Forest stands out as it achieves an accuracy rate of more than 90%, demonstrating its capability to handle complex data structures. Following shortly behind, both CatBoost and XGBoost surpassed 89%, demonstrating the efficacy of gradient boosting approaches. In spite of its ease of use, Logistic Regression exhibits competitive accuracy, showing its efficacy in identifying linear correlations within the dataset. Conversely, the accuracies of AdaBoost and Gaussian Naive Bayes are substantially less,

at around 69%. These findings highlight the need for Gaussian Naive Bayes to better adjust to feature dependencies and the necessity for a more thorough investigation of AdaBoost's sensitivities to certain data attributes. Additional exploration of the subtleties in the data can identify areas that might be improved. These results provide a useful compass, directing the choice and improvement of models by highlighting their unique advantages and possible disadvantages. The thesis reveals the delicate dance between model complexity and understanding, highlighting the crucial need of matching classifier decisions with the intrinsic characteristics of the dataset. As the study progresses, it creates opportunities for more research, including creative feature engineering, hyperparameter tweaking, and the investigation of ensemble tactics. These next stages will ensure the model's strong applicability in practical situations by focusing on the subtle increase of performance rather than merely improvement.

4.5 Conclusion

The analysis reveals a wide range of model results, with Random Forest outperforming all others by over 90% accuracy. The results highlight the fine line between comprehension and model complexity, highlighting the distinct benefits and possible drawbacks of every classifier. AdaBoost and Gaussian Naive Bayes were shown to need work, although Logistic Regression demonstrated competitive accuracy. In order to guarantee ongoing advancement in the field of sports analytics, the study not only directs the use of the model in its current forms but also indicates potential avenues for future improvement, such as hyperparameter optimization and innovative feature engineering.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Introduction

Football is being revolutionized by the introduction of machine learning, which provides fans with real-time information, improves player development, and maximizes talent evaluation. In addition to improving the fan experience, this technology helps teams financially through better scouting and higher levels of involvement. Beyond the pitch, the social impact builds a more dynamic and equitable football ecosystem.

5.2 Impact on Society

5.2.1 Enhanced Sporting Experience

By offering real-time insights and predictions during games, machine learning can improve the fan experience by facilitating more intelligent debates and analysis. Additionally, it can customize recommendations and content to fit the tastes of each viewer. Fans can feel excited and anticipatory when machine learning algorithms, for example, study player movements and game patterns to forecast the chance of a goal or other noteworthy events. Machine learning can also be used to create customized match highlights and summaries based on user interests and preferences.

5.2.2 Improved Player Development

Coaches and trainers can gain insights into a player's strengths, weaknesses, and possible areas for improvement by using machine learning to find patterns and trends in the player data. Player development can be accelerated and training programs optimized with this data-driven approach. Massive amounts of player performance data, such as physical measurements, tactical choices, and passing patterns, can be analyzed by machine learning models to determine which areas players thrive in and which ones might need more work or attention. By recognizing these trends, coaches can modify training plans to better meet the needs of individual players, which will enhance player development and team performance as a whole.

5.2.3 Fairer and More Equitable Talent Evaluation

Coaches can gain insights into team tactics, player positioning, and opponent strategies by using machine learning to analyze game data and spot patterns that human experts might miss. Better tactical choices and enhanced team performance may result from this. Large volumes of game footage, including player movements, pass sequences, and defensive formations, can be analyzed by machine learning algorithms to find patterns and trends that could affect how games turn out. Coaches can gain a competitive edge by using these patterns to inform tactical decisions about player substitutions, formation changes, and strategic tweaks.

5.2.4 Economic Benefits

Clubs can acquire talent more successfully and save money by using machine learning to optimize player scouting, recruitment, and development strategies. Additionally, it can increase viewership and fan engagement, opening up new revenue streams. To find talented players and improve scouting tactics, machine learning algorithms can examine enormous volumes of player data, such as performance metrics, scouting reports, and social media activity. This could result in more efficient player acquisition and lower expenses from trial periods and pointless signings. Furthermore, machine learning can be used to tailor recommendations and content for individual fans. This can boost viewership and engagement and open up new revenue streams for broadcasters and clubs. In conclusion, the use of machine learning in football has the potential to benefit society in a number of ways, including through raising the quality of the game, developing players, encouraging more equitable talent evaluation, facilitating better tactical decision-making, lowering player injuries, and producing financial gains. These effects may help create a football ecosystem that is more vibrant, just, and long-lasting.

5.3 Impact on Environment

5.3.1 Decreased Travel and Transportation

By optimizing team schedules, training programs, and scouting efforts, machine learning algorithms can help reduce the need for pointless travel and transportation. A smaller environmental impact and decreased carbon emissions may result from this decrease in travel.

5.3.2 Energy Efficiency

By optimizing stadium energy management systems with machine learning, energy consumption and related environmental effects can be decreased. For example, occupancy patterns and climate data can be analyzed by machine learning algorithms to improve the efficiency of lighting, heating, and cooling systems.

5.3.3 Waste reduction

Data on fan behavior and consumption patterns can be analyzed using machine learning, which makes it possible to put waste reduction plans into action and create more environmentally friendly goods and services. For instance, food waste can be decreased and more accurate purchasing made possible by machine learning models that can forecast food and beverage consumption at stadiums.

5.4 Ethical Aspects

Football's use of machine learning presents a number of ethical issues that require careful thought and solutions. These worries cover a wide range of topics, from privacy and individual justice to larger societal ramifications.

5.4.1 Fairness and Bias

Based on racial, gender, or other sensitive characteristics, machine learning models trained on biased data have the potential to reinforce prejudice and discrimination against specific player groups. Implementing bias mitigation strategies and closely examining training data are necessary to ensure equitable and inclusive results.

5.4.2 Transparency and Explain ability

Machine learning models frequently lack explainability and transparency, which makes it difficult for stakeholders to comprehend the decision-making process and may raise ethical and mistrust issues. In order to build acceptance and trust, explainability strategies that make the logic behind predictions visible must be used. Data Privacy and Security: Sensitive information found in player

data gathered for machine learning applications must be shielded from misuse and illegal access. Ensuring player privacy and averting possible harm requires strong data governance policies and security measures.

The football industry can take advantage of machine learning while maintaining moral standards and encouraging just and responsible behavior by anticipating ethical issues and taking proactive measures to address them. This will guarantee that technical developments support the ethical and responsible application of AI in football while fostering a more welcoming and positive sport.

5.5 Sustainability Plan

Football's growing dependence on machine learning (ML) presents a special chance to advance environmental sustainability in the game in addition to performance gains. We can help create a more environmentally friendly football landscape by putting into practice ML-powered strategic initiatives that encourage players, supporters, and organizations to make sustainable decisions. A suggested sustainability strategy for your thesis is as follows:

5.5.1. Improving Stadium Management Energy Efficiency

Combine machine learning (ML) systems to examine trends in energy usage in HVAC (heating, ventilation, and air conditioning) systems and lights. By utilizing this data, carbon footprint may be greatly decreased by scheduling maintenance, optimizing operations, and switching to renewable energy sources like solar or wind power. Water conservation: Install intelligent irrigation systems that use machine learning algorithms to assess pitch consumption, weather, and soil moisture content. This can reduce the amount of water wasted when maintaining pitches and planting, encouraging wise water use. Waste Reduction: Utilize machine learning (ML) to examine concession stand data and fan consumption trends in order to forecast demand and enhance the production of food and drinks. This can result in a more sustainable waste management system by minimizing food waste and using less packaging.

5.5.2. Green Transportation and Planning

Travel Optimization: Apply machine learning (ML) to team travel schedule analysis in order to find chances for carpooling, route optimization, and electric or hybrid vehicle prioritization. By doing this, team travel's carbon emissions may be greatly reduced.

Fan Engagement and Mobility: Using ML-powered carpooling or public transit scheduling applications, promote the adoption of environmentally friendly modes of mobility among fans. Encourage the use of gamified platforms and prizes to encourage walking, cycling, and public transportation utilization.

Distance Education and Cooperation: Examine the possibilities of using machine learning (ML)-powered virtual reality (VR) and remote training technologies to cut down on needless travel for meetings and training sessions. By doing this, the environmental effect of team operations may be further reduced.

5.5.3. Educating Fans and Changing Their Behavior

Personalized Outreach: Use machine learning (ML) to target fans with information on the environmental effect of football and the sustainability efforts that have been put in place. This has the potential to increase consciousness and promote eco-friendly decisions.

Gamified Sustainability: Create engaging platforms and incentive programs that monitor and reward supporters who use reusable water bottles, recycle responsibly, or choose environmentally friendly modes of transportation at the stadium. This can increase the effect and engagement of sustainability.

Teaching Resources and Content: Make use of ML-powered interactive platforms and digital displays to teach football fans about environmental challenges and highlight the benefits of sustainability policies that have been put in place. This may instill a sense of accountability and motivate constructive change.

5.5.4. Ongoing Monitoring and Enhancement:

Data-Driven Evaluation: Use ML-powered dashboards to track the success of sustainability programs by keeping an eye on waste production, water usage, energy consumption, and other

pertinent data. This information may be utilized to pinpoint problem areas and continuously modify tactics to get the best outcomes.

Openness and Cooperation: Exchange sustainability information and industry best practices with other sports groups and industry stakeholders. In addition to fostering teamwork, this can hasten the adoption of sustainable practices throughout the football ecosystem and serve as a model for other sports.

5.6 Conclusion

Beyond only improving performance, machine learning has a good social impact on football. There are several potential benefits for the sport, including increased fan involvement and financial rewards. However, in order to ensure a responsible integration of technology and create a football landscape that is not just competitive but also socially and environmentally conscientious, ethical considerations and a well-thought-out sustainability strategy are essential.

CHAPTER 6

CONCLUSION

6.1 Conclusion

This thesis explores the use of machine learning to football player position prediction, looking at various algorithms, preprocessing techniques, and assessment criteria. The study shows that machine learning can reliably forecast player positions, providing insightful information for player development, team tactics, and fan interaction. It does this by leveraging a large dataset that includes scouting reports and player performance data. The study shows reasonable accuracy across all models using a variety of machine learning techniques, such as logistic regression, decision trees, random forests, and support vector machines. Random forests and support vector machines are superior at predicting fewer common places, whereas logistic regression provides simplicity. Because player factors have such a significant influence, the study emphasizes the importance of feature selection. It becomes clear that machine learning is a transformational force that can improve many aspects of sports, including player predicting. However, more study is required to address ethical issues, enhance data collecting, preprocessing methods, and investigate cutting-edge approaches in order to ensure responsible and transparent deployment.

The study's examination of machine learning's application to football player position prediction culminates in a demonstration of the technology's revolutionary potential for fan interaction, team strategies, and player development. Algorithms for predictive machine learning are very accurate, especially when deep learning and collaborative techniques are applied. The study highlights how important feature selection is for improving prediction accuracy and how important data preparation is for guaranteeing consistency and quality. Football has a bright future ahead of it, but realizing that potential will require further study into combining machine learning with simulations, real-time decision assistance, and customized fan experiences.

6.2 Further recommendation work

This research project's success depends on how well machine learning is integrated into football players' placement. The initiative intends to transform how players' roles are selected by working with football experts, creating and improving prediction algorithms, and giving ethical issues first priority. The project's potential effect is further strengthened by its user-friendly interface, interaction with current analytics systems, and longitudinal research on player development. Furthermore, the project aims to improve football management decision-making while simultaneously making football a more inclusive and engaging experience for fans through the investigation of fan engagement apps and the implementation of ongoing monitoring systems. Sufficient financial preparation and a methodical approach to project management, which takes ethical considerations into account, are crucial for the effective implementation and long-term viability of this innovative

REFERENCE

- [1] data web site:<https://sofifa.com/player/20801>
- [2]. data set :https://www.kaggle.com/search?q=fifa_data_2015_to_2020.csv
- [3]. S Bosu Babu¹, Vavilapalli Vivek², Dasari Mahendra Tulasi Kumar³, Korra Prathyusha⁴, Galla Pavan Teja⁵
- [4] Richard Pariath¹ , Shailin Shah² , Aditya Surve³ , Jayashri Mittal⁴
- [5] Using Machine Learning to Draw Inferences from Pass Location Data in Soccer Joel Brooks*, Matthew Kerr and John Guttag Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
- [6] Predicting Player Position for Talent Identification in Association Football To cite this article: Nazim Razali et al 2017 IOP Conf. Ser.: Mater. Sci. Eng. 226 012087
- [7] Towards data-driven football player assessment Rade Stanojevic and Laszlo Gyarmati Qatar Computing Research Institute HBKU, Doha, Qatar {rstanojevic,lgyarmati}@qf.org.qa
- [8] Sports Analytics algorithms for performance prediction K Apostolou, C Tjortjis 2019 10th International Conference on Information, Intelligence ..., 2019•
- [9] A Machine Learning Model to Predict Player's Positions based on Performance.
- [10] Global and Local Perspectives of Sport Management: Book of Abstracts of the 4th World Association for Sport Management World Conference, Doha, Qatar, 5–8 March, 2023 Bo, Hannah H.; Valantine, Irena; Seiler, Sean; Anderson, Devin J. F.; Zhao, Troy T.; Hogg, Caroline; Bradbury, Trish; Swart-Arries, Kamilla; Zhang, James J.
- [11] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
- [12] Maimon, O. Z., & Rokach, L. (2014). Data mining with decision trees: theory and applications (Vol. 81). World scientific.

- [13] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.
- [14] Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.
- [15] Zhang, H. (2004). The optimality of naive Bayes, *flairs conference*

APPENDIX

Research Reflections:

Finding issues and circumstances was difficult for me when I worked on this project. In order to ensure that they would function as best they could, I began by selecting the greatest applications available. Everyone also needed to have a thorough understanding of that using Python and machine learning. I was surprised by how difficult it was to gather and arrange such a large amount of data. I took a while to attain my objective, but I did.

This project must be finished in order for students to pass the CSE-499 Project/Internship Capstone course.

By Hasan

ORIGINALITY REPORT

14%	12%	4%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	5%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	Submitted to CSU Northridge Student Paper	1%
4	Submitted to Bangladesh University of Professionals Student Paper	1%
5	cupdf.com Internet Source	1%
6	fastercapital.com Internet Source	1%
7	Mohammed Zakariah, Salman A. AlQahtani, Mabrook S. Al-Rakhani. "Machine Learning-Based Adaptive Synthetic Sampling Technique for Intrusion Detection", Applied Sciences, 2023 Publication	<1%