A Comparative Study of Predictive Analysis Using Machine Learning Techniques: Performance Evaluation of Manual and AutoML Algorithms

Karim Mohammed Rezaul[1], Md. Jewel[2], Anjali Sudhan[3], Mifta Uddin Khan[4], Maharage Roshika Sathsarani Fernando[5], Kazy Noor e Alam Siddiquee[6], Tajnuva Jannat[7], Muhammad Azizur Rahman[8], Md Shabiul Islam[9]

Wrexham University, Faculty of Arts, Science and Technology, Wrexham LL11 2AW, UK[1]

Centre for Applied Research in Software & IT (CARSIT), 80a Ashfield Street, London, England, E1 2BJ, UK[2, 3, 4, 5, 7]

Multimedia University, Faculty of Engineering (FOE), Cyberjaya 63100, Malaysia[6, 9]

Cardiff Metropolitan University, Department of Computer Science, Llandaff Campus, Western Avenue, Cardiff, CF5 2YB, UK[8]

***Abstract*—In this study, we have compared manual machine learning with automated machine learning (AutoML) to see which performs better in predictive analysis. Using data from past football matches, we tested a range of algorithms to forecast game outcomes. By exploring the data, we discovered patterns and team correlations, then cleaned and prepped the data to ensure the models had the best possible inputs. Our findings show that AutoML, especially when using logistic regression can outperform manual methods in prediction accuracy. The big advantage of AutoML is that it automates the tricky parts, like data cleaning, feature selection, and tuning model parameters, saving time and effort compared to manual approaches, which require more expertise to achieve similar results. This research highlights how AutoML can make predictive analysis easier and more accurate, providing useful insights for many fields. Future work could explore using different data types and applying these techniques to other areas to show how adaptable and powerful machine learning can be.**

Keywords*—*Machine learning; predictive analytics; sports forecasting; automated machine learning (AutoML); feature engineering; model evaluation; data pre-processing; algorithm comparison; football analytics; sports betting; team performance metrics; exploratory data analysis (EDA); cross-validation techniques

I. Introduction

Millions of football fans from all around the world attend the UEFA European Championship, also referred to as the UEFA Euro. This esteemed competition, which is hosted by UEFA, features the top teams from throughout Europe, showcasing their talent, tenacity, and competitive spirit [1]. Analysts, enthusiasts, and commentators eagerly engage in predicting the outcomes of this highly anticipated and often unpredictable event. Recent advancements in machine learning (ML) algorithms, combined with the availability of extensive historical football data, have opened new avenues for predicting match results and identifying potential tournament winners. These sophisticated algorithms can detect patterns in complex datasets, providing valuable insights for predictive analysis and strategic decision-making.

Using ML algorithms to forecast the UEFA Euro winner involves a detailed analysis of historical match data, team statistics, player performance metrics, and other factors that influence team success. By understanding the intricate interactions of these factors, ML models can predict future outcomes based on data from past tournaments. In this study, a variety of manual ML algorithms known for their efficiency in predictive modelling were used. These include Ada Boost, Random Forest, XGBoost, Decision Tree, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbors (KNN), and Naive Bayes. Each algorithm has unique strengths, making them suitable for different predictive tasks in forecasting the winner.

Additionally, the study explores the realm of Automated Machine Learning (AutoML), utilizing advanced techniques to streamline and optimize the model development process. The AutoML framework incorporates a broad set of algorithms, such as Ridge, Quadratic Discriminant Analysis, Linear Discriminant Analysis, Extra Trees Classifier, Extreme Gradient Boosting, Light Gradient Boosting Machine, and Dummy Classifier, among others. This comprehensive approach allows for a thorough evaluation of predictive efficacy.

This study aims to demonstrate the efficacy of both manual and automated machine learning (AutoML) approaches in sports analytics, especially in forecasting the results of the UEFA Euro 2024. This study aims to demonstrate the potential for widespread acceptance and innovation in sports analytics by analysing the performance of several machine learning algorithms in projecting tournament results. The findings of this study are expected to enlighten stakeholders such as analysts, coaches, and investors, allowing them to make more informed judgements and strategic assumptions during the tournament.

The paper is organized as follows: Section I presents the introduction. In Section II, we define the problem and outline the objectives. Section III covers the literature review, addressing manual and automated machine learning separately. Section IV explains the research methodology. Section V details the preprocessing, cleaning, and data preparation steps. Section VI provides a comparative analysis of manual and automated machine learning across various classifiers. Section VII presents the results and evaluation, and finally, Section VIII concludes the paper.

II. Problem Definition and Objectives

The purpose of this study is to show that both manual and automatic machine learning (AutoML) methodologies can accurately forecast UEFA Euro 2024 outcomes. The study creates prediction models for team performance by analysing historical data on international football matches and team characteristics. The findings, which highlight the potential of machine learning, particularly AutoML, in improving prediction accuracy, seek to enlighten analysts, coaches, and bettors, supporting greater use and innovation in sports analytics.

The main objective of this study is to demonstrate that both manual and automatic machine learning (AutoML) methods can effectively predict the outcomes of the UEFA Euro 2024 competition. The goal is to create prediction models that accurately assess each team's likelihood of success by thoroughly analysing historical data from international football matches and considering various team characteristics. This research aims to highlight the potential of machine learning (ML), particularly AutoML, in expediting predictive analytics processes and enhancing model accuracy.

The key objectives include the careful collection and pre-processing of historical international soccer match data and team attributes, followed by a detailed exploratory data analysis to identify relevant features. The study will then involve selecting informative features and applying rigorous feature engineering techniques to capture essential team and match characteristics. Various machine learning algorithms will be evaluated and compared to identify the most effective models based on performance metrics. The selected models will undergo thorough training using pre-processed data, with hyperparameter optimization to ensure optimal performance. Rigorous evaluation of predictive performance using appropriate metrics, along with the implementation of cross-validation techniques to ensure model generalizability and reduce overfitting, are crucial parts of this research.

Additionally, this project will test the performance of the models using benchmark datasets. The study aims to demonstrate the potential of these models in sports analytics by providing significant insights and encouraging wider adoption and innovation. The focus will be on evaluating the effectiveness of AutoML approaches.

III. Literature Review

A. Use of Manual Machine Learning

Forecasting tournament winners is an enduring challenge in sports analysis, extensively explored across research. Machine learning emerges as a prominent tool in this domain, offering predictive capabilities based on historical data. Leveraging past records, machine learning models discern patterns and variables correlated with auspicious outcomes, thereby enabling forecasts for future events. This methodology capitalizes on the wealth of information contained within historical datasets, facilitating the identification of key determinants of success. Through iterative learning processes, these algorithms refine their predictive accuracy, contributing to the advancement of sports analytics. The utilization of machine learning in predicting tournament winners emphasises the significance of data-driven approaches in enhancing the understanding of sports dynamics and informing strategic decision-making processes within the realm of athletics. The challenge of predicting football match outcomes is addressed in the research by Hucaljuk & Rakipović, acknowledging the complexity stemming from numerous unquantifiable factors. A software solution is developed to tackle this challenge, undergoing testing to optimize feature and classifier combinations. Results demonstrate satisfactory predictive capabilities surpassing reference methods, with an accuracy exceeding the initial goal of 60%. However, the study suggests areas for enhancement, particularly in feature selection, proposing the inclusion of player form data for improved accuracy [2]. Additionally, increasing the size of the dataset for training could further enhance predictive performance. This project exemplifies successful advancement in football match prediction methodologies while highlighting avenues for future research and refinement in feature engineering.

Another research investigates the efficacy of utilizing machine learning techniques to predict football match outcomes by incorporating pre-game features instead of relying solely on post-game goal statistics. Custom-generated features are developed and compared against in-game data features using the XGBoost algorithm. Results indicate superior prediction accuracy with custom features, demonstrating higher precision, recall, f1 score, and accuracy compared to in-game features. The research suggests that leveraging comprehensive player and team statistics, such as dribbling and expected goals, could further enhance predictive performance. Additionally, considering factors like team formation and fan sentiment from social media could provide valuable insights into match outcomes. The study underscores the potential for enhanced predictive modelling in football matches through the incorporation of diverse pre-game features and data sources [3]. The studies by Hucaljuk & Rakipović and Rose et al. 2022 both address the challenge of predicting football match outcomes using machine learning techniques [2] [3]. Hucaljuk & Rakipović develop a software solution to optimize feature and classifier combinations, surpassing an initial accuracy goal of 60% [2]. They highlight the importance of feature selection and suggest incorporating player form data to enhance predictive accuracy. Conversely, focus on incorporating pre-game features, such as comprehensive player and team statistics, using custom-generated features [3]. This study demonstrates superior prediction accuracy compared to relying solely on post-game goal statistics, emphasizing the potential for enhanced predictive modelling through diverse pre-game features and data sources. Both studies contribute to advancing football match prediction methodologies and highlight avenues for future research in feature engineering and data analysis. The research by Groll et al. [4] introduces a hybrid modelling approach for predicting soccer match scores, combining random forests with two ranking methods: Poisson ranking and bookmakers' odds. By incorporating team covariate information and ability parameters derived from both ranking methods, the model accurately estimates team strengths. The approach is applied to FIFA Women's World Cups 2011, 2015, and 2019, with simulations favouring the USA as the top contender for the 2019 title, followed by France, England, and Germany. The study highlights the effectiveness of integrating

diverse methodologies for robust predictions in soccer tournaments, offering insights into team performance and tournament outcomes [4].

Another study focuses on employing machine learning techniques to predict the winner of the ICC Men's T20 World Cup 2020. Four algorithms, including Random Forest, Extra Trees, ID3, and C4.5, were compared, with Random Forest exhibiting the highest proficiency at 80.86% custom accuracy. Australia emerged as the predicted champion. Future directions include optimizing the predictive models and incorporating additional parameters like match venue and weather forecast to enhance accuracy. The study underscores the utility of machine learning in sports prediction and offers insights into potential improvements for future analyses, emphasizing the importance of considering various factors for more accurate forecasts in cricket tournaments [5]. The both studies focus on utilizing machine learning techniques for sports prediction, albeit in different contexts. Groll and the team introduce a hybrid modelling approach for predicting soccer match scores, incorporating random forests with Poisson ranking and bookmakers' odds. This study demonstrates the effectiveness of integrating diverse methodologies for robust predictions in soccer tournaments, providing insights into team performance and outcomes. In contrast, Basit and the team concentrate on predicting the winner of the ICC Men's T20 World Cup 2020 using machine learning algorithms such as Random Forest, Extra Trees, ID3, and C4.5 [4] [5]. This research underscores the utility of machine learning in sports prediction, emphasizing the importance of considering various factors for more accurate forecasts, particularly in cricket tournaments. Both studies contribute to advancing predictive modelling in sports and offer valuable insights into improving accuracy in tournament predictions. In examining cricket match prediction models, emphasize the development of a machine learning model specifically tailored for Indian Premier League (IPL) matches, achieving nearly 90% accuracy [6]. Conversely, Kumar, et al. [7] focus on Decision Trees and Multilayer Perceptron Network models, highlighting the superiority over traditional statistical methods. This CricAI system offers a user-friendly prediction tool, emphasising the flexibility of machine learning approaches. Vistro et al [8] similarly explore IPL match prediction using various machine learning algorithms, achieving high accuracies up to 94.87%. While all studies underscore the significance of data science in sports analytics, two studies emphasize the potential applicability of the methodologies beyond cricket, which could inform predictive analytics in UEFA Euro and other sports contexts [7][8].

The research by Elmiligi & Saad presents a novel hybrid approach combining machine learning and statistical methods to predict soccer match outcomes [9]. Analysing a dataset comprising over 200,000 match results from 2000/2001 to 2016/2017, the research explores various features including team and player statistics, home/away advantage, and data recency. Two hybrid models are developed, with the best achieving 46.6% prediction accuracy on a test set. The study also evaluates hypotheses regarding feature engineering, finding no significant improvement with recent match data or separate models for each league [9]. Additionally, the research plans to extend its analysis to other sports and conduct comparative feature significance studies. This work contributes to advancing predictive modelling in sports and lays the groundwork for future research directions. Another study delves into predicting football match outcomes, focusing on the 2022 FIFA World Cup, leveraging Exploratory Data Analysis (EDA) and various machine learning algorithms. Notably, Random Forests, Decision Trees, K-Nearest Neighbours, XGBoost, and Gradient Boosting are tested, with XGBoost and Gradient Booster achieving the highest average accuracy of 98.34%. The study introduces a novel approach combining EDA and machine learning to address the challenges of sports match prediction, proposing Multi-output Regressor as a solution. It suggests that this method could accurately forecast sporting event outcomes and encourages further research into incorporating additional factors like current world ranking and new age metrics. The findings contribute to advancing predictive modelling in football and offer potential avenues for enhancing prediction accuracy in future studies [10]. The research by Athish et al. [11] explores the application of the Bayesian approach in predicting soccer match outcomes, leveraging authentic squad information and match results sourced from platforms like Kaggle and Sofifa.com. The Gaussian Naive Bayes model demonstrates 85.43% accuracy in match result prediction, surpassing the 79.81% accuracy achieved by the Decision Tree Classifier. The study offers a tool for users to assess team probabilities in tournaments, although it emphasizes individual discretion in betting due to uncertainties inherent in sports outcomes. The findings contribute to the understanding of machine learning techniques in soccer prediction and provide a basis for further research in the field. The studies discussed present diverse methodologies and approaches for predicting soccer match outcomes using machine learning and statistical techniques [9] [10][11]. Research by Elmiligi & Saad introduces a hybrid model that analyses team and player statistics, achieving a prediction accuracy of 46.6%. It emphasizes the importance of feature engineering and plans to extend its analysis to other sports [9]. In contrast, Majumdar and team focus on the 2022 FIFA World Cup, employing exploratory data analysis and various machine learning algorithms. Their approach yields high accuracy, with XGBoost and Gradient Boosting achieving 98.34% [10]. This study proposes a novel method combining EDA and machine learning, suggesting avenues for further research. Athish et al. explores the Bayesian approach for predicting soccer match outcomes, achieving an accuracy of 85.43% with the Gaussian Naive Bayes model. This study provides insights into machine learning techniques for soccer prediction, emphasizing individual discretion in betting [11]. Overall, these studies contribute to advancing predictive modelling in sports and offer valuable insights for future research directions.

The studies by Chin et al. and Daundkar & Kandhway both employ machine learning techniques to enhance predictive capabilities in sports, focusing on ice hockey and NBA match outcomes, respectively. Chin and the team analysed various machine learning techniques using NHL data from 2015-2021, with Logistic Regression achieving the highest accuracy at 77.82% [12][13]. This study highlights the significance of incorporating match-specific data for improved predictive accuracy [12]. Conversely, Daundkar & Kandhway predict NBA match outcomes based on past team performances,

achieving a prediction accuracy of approximately 66%. This research underscores the relevance of machine learning in sports betting and offers insights into the predictive capabilities of historical data in forecasting NBA match outcomes, aligning closely with human expert accuracy [13]. Both studies contribute to advancing predictive analytics in sports, offering valuable implications for future research and applications [12][13]. The research by Kumar, et al. [7] introduces an advanced approach to football analysis by utilizing predictive data like expected goals instead of descriptive data like shots taken and goals scored. By applying fixed parameters on machine learning algorithms, the method aims to evaluate teams and players based on performance rather than results, enhancing scouting and strategy formation. Results indicate that the light XGBoost machine learning model provides a better match of shot quality, as measured by McFadden's pseudo-R-squared score. Incorporating a "big chance" component further improves assessment criteria, although this capability is not available in the dataset used. Feature importance measurements highlight variables crucial to the model's outputs, offering valuable insights for performance analysis and talent identification. Another study focuses on predicting halftime results and league winners in the English Premier League using classification models. Leveraging ensemble techniques, the study achieves 80% accuracy in halftime result prediction and up to 95% accuracy in league winner prediction [14]. Building upon previous work, the research incorporates additional features such as team form and form points, enhancing prediction accuracy. By training models at a match week level, the study offers insights into predicting league winners throughout the season. The findings highlight the potential of utilizing dynamic datasets and simple features for accurate football match predictions and league analysis [14].

The studies by Tiwari et al. and Jaeyalakshmi et al. explore machine learning techniques for predicting football match outcomes, albeit with different emphases [15][16]. Tiwari and the research team focus on utilizing Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) to leverage the abundance of statistical football data, aiming to enhance prediction accuracy for various match-related information. The conclusion of this study highlights the superiority of LSTM-based RNNs over traditional methods, suggesting further improvements by incorporating player statistics and larger datasets [15]. Conversely, Jaeyalakshmi and the team introduce a machine learning approach for forecasting football match results, emphasizing feature selection, data imbalance treatment, and model generalization. Achieving over 81% accuracy with the Random Forest Algorithm, this study emphasises the unpredictable nature of football and advocates responsible betting, suggesting future enhancements through additional statistics incorporation [16]. Another study introduces a machine-learning model employing a random forest algorithm for predicting football player performance and optimizing fantasy football team line-ups. Back-testing on historical data yielded a Mean Square Error (MSE) of 4.921 and a Root Mean Square Error (RMSE) of 2.2275, with the model presenting the potential for profitability in fantasy football betting [17]. The analysis involves web scraping, data segregation, and hyperparameter tuning. Results showcase the model's capability in team formation for different formations, with future scope including subscription services and incorporation of additional data sources to enhance predictive accuracy while acknowledging inherent limitations in predicting future events and relying exclusively on past performance data [17].

Proposing a data-driven approach, the study Al-Asadi & Tasdemır [18] aims to estimate football players' market values using machine learning algorithms applied to FIFA 20 video game data. Comparing four regression models, the research finds that random forest outperforms others in accuracy and error ratio. The results suggest potential applications in streamlining negotiations between clubs and players' agents by providing objective market value estimations. Further research avenues include integrating the model into FIFA games for player valuation and developing a calculator to assist gamers in making informed decisions. The methodology of this study demonstrates superiority over traditional approaches, offering practical implications beyond gaming simulations. Further one research introduces a system leveraging real-time feedback and advanced technologies to enhance football technique learning. It utilizes pose estimation with Media pipe and classification algorithms like the Dollarpy, KNN, RFE, and SVM, achieving varying accuracies [19]. Another article tackles the challenge of predicting football match outcomes for sports betting, employing machine learning methods and historical match statistics [20]. The models in this research yield a 65.26% accuracy rate, offering potential profitability. The study by Gifford & Bayrak [21] focuses on NFL game outcome prediction using decision trees and logistic regression. With turnover statistics as key predictors, the models achieve up to 83% accuracy, contributing insights for strategic decision-making in sports analytics. These studies collectively highlight the diverse applications of machine learning in sports and the potential for technology to enhance performance and decision-making in athletics.

B. Use of Automated Machine Learning (AutoML)

An article explored Automated Machine Learning (AutoML) as an end-to-end process for streamlining model development without manual intervention. The paper provides insights into AutoML segments and approaches, emphasizing its practical applicability in industry [22]. Furthermore, it discusses recent trends and suggests future research directions, advocating for a generalized AutoML pipeline and a central meta-learning framework. The study highlights the importance of advancing AutoML to address evolving challenges in machine learning model development, both in academia and industry. The study discussed the significance of Automated Machine Learning (AutoML) in mitigating the challenges of ML adoption, especially for small and medium-sized organizations. This paper highlights its diverse applications across industries and advocates for its potential in democratizing machine learning [23]. It suggests various research opportunities in Information Systems (IS), including qualitative and quantitative studies on AutoML adoption, the development of AutoML adoption theories, and the exploration of fairness and explainability concerns. Additionally, the authors underscore the importance of human-in-the-loop research and address the limitations and boundaries of AutoML

applications, emphasizing the role of IS researchers in advancing AutoML adoption in organizations.

The articles Truong et al. [24] and Ferreira et al. [25] investigated Automated Machine Learning (AutoML) tools' effectiveness but with different emphases. Truong and the team compare commercialized and open-source AutoML tools, highlighting varying strengths and weaknesses across datasets. This research emphasizes the absence of a single superior tool, indicating ongoing AutoML evolution [24]. In contrast, Ferreira and the research team conducted a study exclusively on open-source AutoML tools, focusing on supervised learning scenarios. The results of this research reveal the competitive performance of General Machine Learning (GML) AutoML tools, particularly in binary and regression tasks [25]. Both studies underscore the need for further advancements, Truong, et al in addressing gaps in AutoML pipeline support, and Ferreira et al in expanding comparisons to encompass more technologies and datasets, especially in big data contexts [24][25]. Another study presents a survey on automating the process of building machine learning models, particularly focusing on Combined Algorithm Selection and Hyper-parameter tuning (CASH). It discusses the challenges of efficiently constructing high-quality models due to the vast amounts of data produced daily. The paper comprehensively reviews state-of-the-art efforts in AutoML frameworks and highlights research directions and challenges. By addressing these issues, the aim is to automate the machine-learning pipeline and reduce human intervention, catering to both researchers and practitioners in advancing the field [26]. The article [27] commences with an analysis of existing research in AutoML, hyperparameter tuning, and meta-learning. It highlights the lack of clear documentation and consensus on evaluation criteria in this field. The paper discusses the strengths and weaknesses of various approaches, emphasizing the need for further research to develop a fully automated industrial standard system. Assembling and meta-learning are proposed as effective methods for automating hyperparameter tuning. The authors aim to bridge gaps in existing solutions and plan to devise an architectural style for an efficient AutoML system based on accumulated knowledge and identified drawbacks. The article by Tsiakmaki et al. [28] focuses on applying automated machine learning (AutoML) in Educational Data Mining (EDM) to predict students' learning outcomes. It emphasizes interpretability by restricting the search space to tree-based and rule-based models. The study highlights that AutoML tools surpass default parameter values, especially in classification and regression tasks, highlighting the significance of transparent tools for educators. The findings suggest AutoML has the potential to aid early performance estimation and intervention strategies, offering promising avenues for enhancing academic outcomes in educational environments. In contrast, Shi, et al. [29] present a domain-specific AutoML framework tailored for risk prediction and behaviour assessment in autonomous vehicles (AVs). The system in this research integrates unsupervised risk identification, feature learning with XGBoost, and model auto-tuning using Bayesian optimization. Evaluation of Next Generation Simulation (NGSIM) data demonstrates the framework's efficacy in distinguishing safe from risky behaviours, thus enhancing risk decision-making in Autonomous Vehicle (AVs). Additionally, it provides insights into sensor configurations and data mining, contributing to AV safety and design improvements.

In a nutshell even while previous research shows a variety of approaches and developments in basketball, cricket, and football sports prediction, a large number of studies continue to mostly rely on manual machine learning techniques, neglecting the benefits of automated approaches. By automating feature engineering, model selection, and hyperparameter tuning, the combination of automated machine learning (AutoML) tools with conventional techniques offers promising potential to increase prediction efficiency and accuracy. Machine learning techniques, coupled with automated machine learning (AutoML) tools, showcase promising capabilities in predicting match outcomes and enhancing sports analytics. Our research addresses this gap.

IV. Research Methodology

This study examines a vast dataset comprising historical records from international football matches alongside various team performance metrics, intending to use state-of-the-art machine learning techniques to predict the winner of UEFA Euro 2024. The primary objective is to develop a prediction model capable of accurately assessing the likelihood of victory for each participating team, providing valuable insights to analysts, bookmakers, coaches, and athletes. Notably, the study employed Artificial Neural Networks (ANNs) to construct a model for forecasting tournament match outcomes. To ensure high prediction accuracy, this model underwent cross-validation to prevent overfitting, refinement to improve generalization, and extensive training on substantial datasets [30].

The methodology section evaluates methodological choices based on literature and previous studies, such as ensemble methods for improved predictive accuracy in sports analytics. It also discusses potential limitations like the unpredictability of events and potential biases in historical data. Obstacles and solutions include data quality issues, model over-fitting, and changes in team dynamics. Robust pre-processing techniques, regularization and pruning of decision trees, and ongoing data updates are employed. Fig. 1 depicts the research methods used in this study.

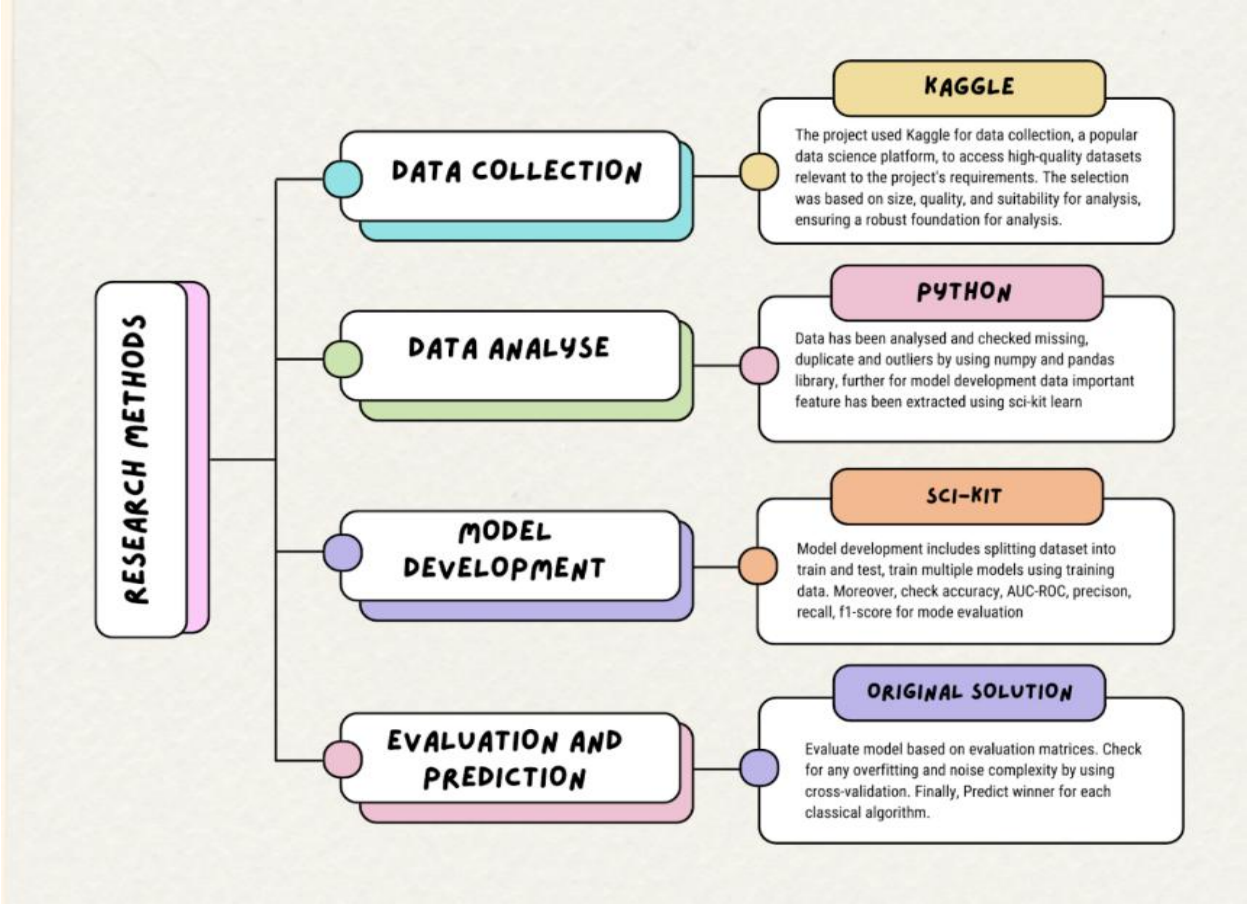

Fig. 1. Research methodology.

A. Chosen Approach

Our study adopts a quantitative research design, utilizing the numerical nature of data and modern computational methods to address the complexities inherent in predictive analytics. By leveraging a comprehensive dataset that includes player stats, team performance metrics, and past match results, we employ machine learning—a sophisticated branch of artificial intelligence—to manage and analyze this vast amount of information efficiently. Our approach integrates traditional statistical methods with advanced machine learning models, allowing us to uncover intricate patterns and interactions that conventional techniques might overlook. This dual strategy enhances the reliability and precision of our predictions, ensuring a robust analysis conducive to accurate forecasting.

In our investigation, we use well-established machine learning algorithms such as Random Forest and Support Vector Machines (SVM). Furthermore, we explore the potential of Automated Machine Learning (AutoML), a revolutionary development in predictive analysis. AutoML automates the selection, application, and refinement of various machine learning models, significantly streamlining the analytical process. This automation reduces the time and expertise required to implement complex models, as AutoML efficiently evaluates numerous algorithms and their configurations to identify the most suitable one for our data.

By integrating AutoML, our study enriches the toolkit available for predictive analytics and sets a standard for future research. The use of AutoML not only enhances predictive accuracy but also demonstrates the potential of advanced automation in pushing the boundaries of data forecasting. This methodology presents an exciting frontier for further exploration, promising significant advancements in the field of predictive analysis.

V. Dataset Collection

For predictive modelling, this study utilized historical European football match data sourced from Kaggle (https://www.kaggle.com/datasets/mahadinour/international-football-matches) [31]. The dataset encompassed match outcomes, team metrics, and player performance statistics. To maintain relevance, the data underwent filtration to focus solely on European matches from the UEFA Euro 2024 tournament. Python's Pandas library facilitated efficient filtering based on competition type and team geographical locations. This dataset played a pivotal role in achieving the study's goals. To ensure research reproducibility, a detailed data dictionary will be made available alongside the study, outlining each collected variable and its origin.

A. Methods used to Analyse Collected Data

1) Feature engineering: The study will employ sophisticated feature engineering techniques to develop new

variables that could influence match outcomes, such as team morale indices or fatigue levels.

2) Model training and testing: Utilizing cross-validation techniques to partition the data and evaluate the model's performance across different subsets, ensuring the model's ability to generalize to new data.

3) Hyperparameter optimization: Advanced techniques like Bayesian Optimization will be used for tuning the models to find the optimal configuration of parameters.

B. Experimental Set-Up and Results

Exploratory Data Analysis (EDA) is a crucial stage in data analysis, utilizing statistical and visualisation methods to understand the dataset's structure, identify trends, and gain insights. It helps identify patterns, outliers, and correlations, and aids in initial data processing, feature selection, and machine learning techniques. EDA is essential for data science and machine learning.

Fig. 2 shows the dataset, containing 2391 rows and 25 columns, which includes information about UEFA Euro qualification tournament football matches. It includes team details, match date and location, and final score. Analysing this data can reveal trends, patterns, and identify factors affecting national team success.

To find out a concise summary of the given dataframe, a set of Python commands like info(), dtypes, select_dtypes, etc were used. This helps to understand the descriptive statistics for "object" datatype, which represents a string or categorical data. Fig. 3 describes the same.

date	home_tea	away_tea	home_tea	away_tea	home_tea	away_tea	home_tea	away_tea	home_tea	away_tea	tourname	city	country	neutral_lo	shoot_out	home_tea	home_tea	away_tea
20-04-1994	Northern I	Liechtenst	Europe	Europe	35	160	0	0	4	1	UEFA Eurc	Belfast	Northern I	FALSE	No	Win		
04-09-1994	Estonia	Croatia	Europe	Europe	111	90	0	0	0	2	UEFA Eurc	Tallinn	Estonia	FALSE	No	Lose		
04-09-1994	Israel	Poland	Europe	Europe	52	32	0	0	2	1	UEFA Eurc	Ramat-Ga	Israel	FALSE	No	Win		
06-09-1994	Czech Rep	Malta	Europe	Europe	44	69	0	0	6	1	UEFA Eurc	Ostrava	Czech Rep	FALSE	No	Win		
07-09-1994	Belgium	Armenia	Europe	Europe	20	159	0	0	2	0	UEFA Eurc	Brussels	Belgium	FALSE	No	Win		
07-09-1994	Cyprus	Spain	Europe	Europe	71	6	0	0	1	2	UEFA Eurc	Limassol	Cyprus	FALSE	No	Lose		
07-09-1994	Faroe Isla	Greece	Europe	Europe	127	34	0	0	1	5	UEFA Eurc	Toftir	Faroe Isla	FALSE	No	Lose		
07-09-1994	Finland	Scotland	Europe	Europe	45	33	0	0	0	2	UEFA Eurc	Helsinki	Finland	FALSE	No	Lose		
07-09-1994	Georgia	Moldova	Europe	Europe	124	149	0	0	0	1	UEFA Eurc	Tbilisi	Georgia	FALSE	No	Lose		
07-09-1994	Hungary	Turkey	Europe	Europe	55	57	0	0	2	2	UEFA Eurc	Budapest	Hungary	FALSE	No	Draw		
07-09-1994	Iceland	Sweden	Europe	Europe	42	3	0	0	0	1	UEFA Eurc	ReykjavÃ-	Iceland	FALSE	No	Lose		
07-09-1994	Latvia	Republic c	Europe	Europe	86	13	0	0	0	3	UEFA Eurc	Riga	Latvia	FALSE	No	Lose		
07-09-1994	Liechtenst	Austria	Europe	Europe	151	37	0	0	0	4	UEFA Eurc	Eschen	Liechtenst	FALSE	No	Lose		
07-09-1994	Luxembou	Netherlan	Europe	Europe	116	5	0	0	0	4	UEFA Eurc	Luxembou	Luxembou	FALSE	No	Lose		
07-09-1994	North Mac	Denmark	Europe	Europe	117	12	0	0	1	1	UEFA Eurc	Skopje	North Mac	FALSE	No	Draw		
07-09-1994	Northern I	Portugal	Europe	Europe	39	25	0	0	1	2	UEFA Eurc	Belfast	Northern I	FALSE	No	Lose		
07-09-1994	Norway	Belarus	Europe	Europe	8	142	0	0	1	0	UEFA Eurc	Oslo	Norway	FALSE	No	Win		
07-09-1994	Romania	Azerbaijar	Europe	Europe	7	170	0	0	3	0	UEFA Eurc	Bucharest	Romania	FALSE	No	Win		
07-09-1994	Slovakia	France	Europe	Europe	43	16	0	0	0	0	UEFA Eurc	Bratislava	Slovakia	FALSE	No	Draw		
07-09-1994	Slovenia	Italy	Europe	Europe	75	2	0	0	1	1	UEFA Eurc	Maribor	Slovenia	FALSE	No	Draw		
07-09-1994	Ukraine	Lithuania	Europe	Europe	79	84	0	0	0	2	UEFA Eurc	Kyiv	Ukraine	FALSE	No	Lose		
07-09-1994	Wales	Albania	Europe	Europe	35	101	0	0	2	0	UEFA Eurc	Cardiff	Wales	FALSE	No	Win		

Fig. 2. Chosen dataset.

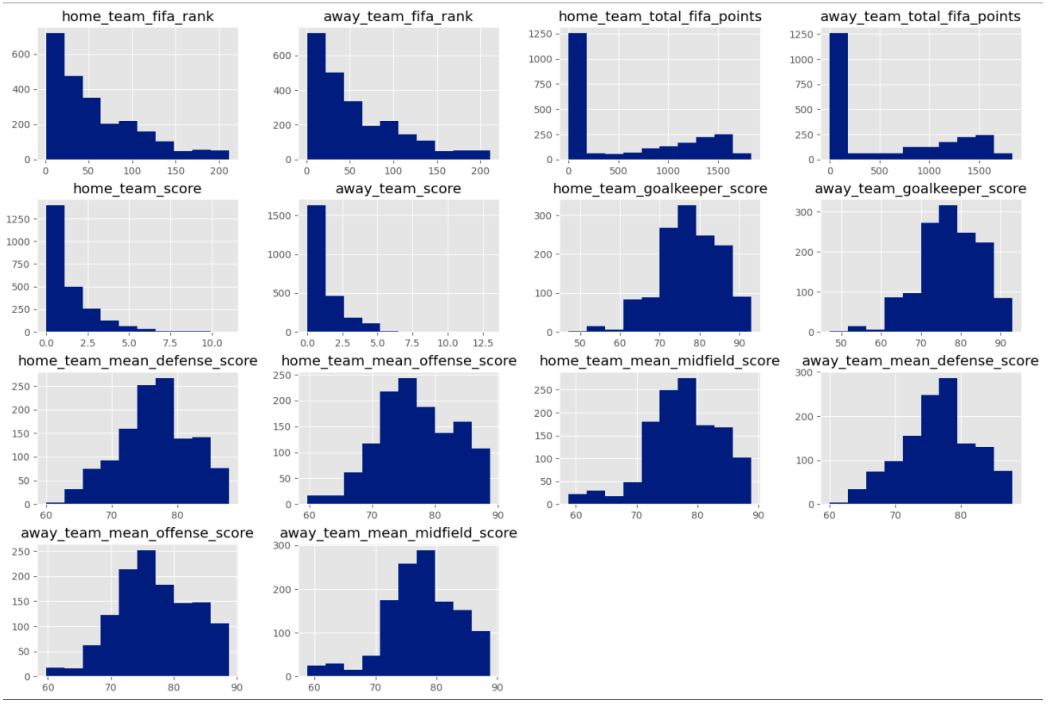

Fig. 3. Histogram plots.

Fig. 4 shows histograms plotted on various columns in a dataset, revealing potential patterns and frequency distributions within the variables.

C. Data Pre-Processing

Data pre-processing involves various procedures like transformation, cleansing, reduction, normalisation, and integration. It involves handling NaN data, managing noisy data, and error fixation. Data transformation converts data into a more intelligible format, while data integration combines data from multiple sources. Data normalisation scales data for similar distribution. Data pre-processing is crucial for machine learning (ML) algorithms to ensure suitable and high-quality data.

1) Visualizing NaN: Fig. 4 illustrate the findings done as part of a null check for the given dataset. Each heatmap corresponds to a data cell, and the colour of each heatmap cell indicates whether or not NaN values are present. The density of NaN values in a given column or row is larger when the colour is darker. It is noted that the following columns in the dataset contain more NaN values than the other columns –

a) home_team_goalkeeper_score
b) away_team_goalkeeper_score
c) home_team_mean_defense_score
d) away_team_mean_defense_score
e) home_team_mean_offense_score
f) away_team_mean_offense_score
g) home_team_mean_midfield_score
h) away_team_mean_midfield_score

2) Visualizing missing data across different years: Fig. 5 displays the trend of missing data across different years which would help to identify years with higher proportion of missing data (data quality assessment). This is a scatter plot and each data point on the plot corresponds to the proportion of missing values for a specific year.

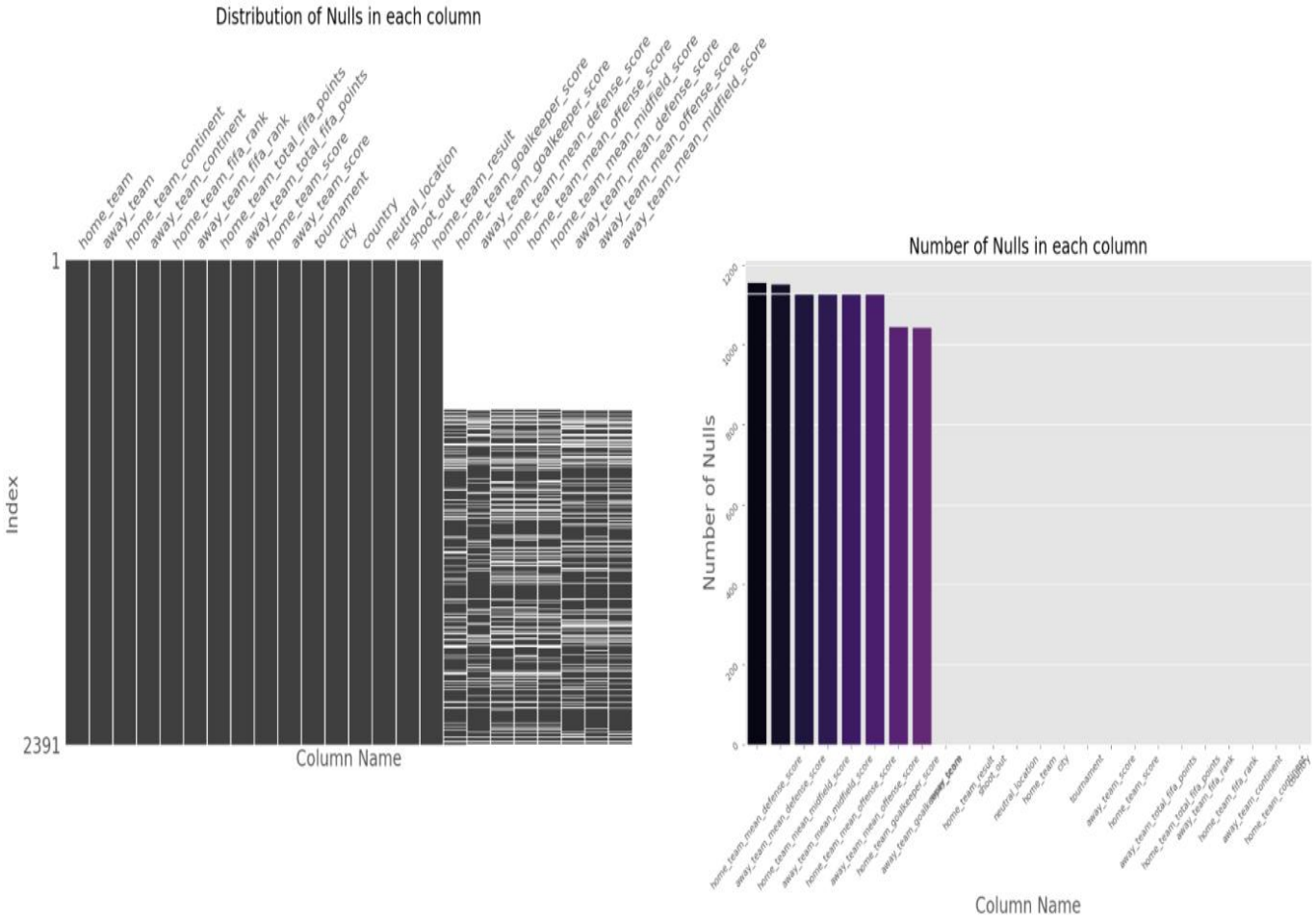

Fig. 4. NaN visualization using HeatMap.

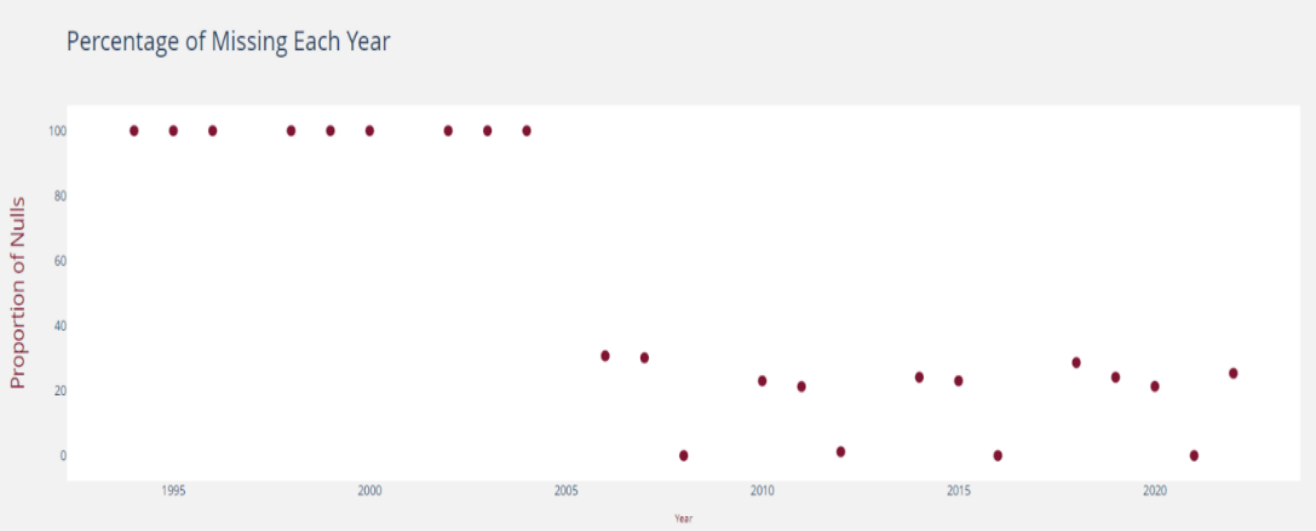

Fig. 5. Visualizing missing data across different years (Scatter plot).

3) Checking for duplicates: This is a check to identify the duplicates in the given dataset. As per the result, we did not find any duplicates in the dataframe.

4) Outliers detection: Fig. 6 represents a parallel coordinate plot. This plot helps us visualize multivariate data by showing multiple variables or attributes as parallel vertical axes. An individual data point is depicted by each polyline by connecting its values across different variables. This technique helps to identify similarities, patterns or relationships between different variables in the given dataset.

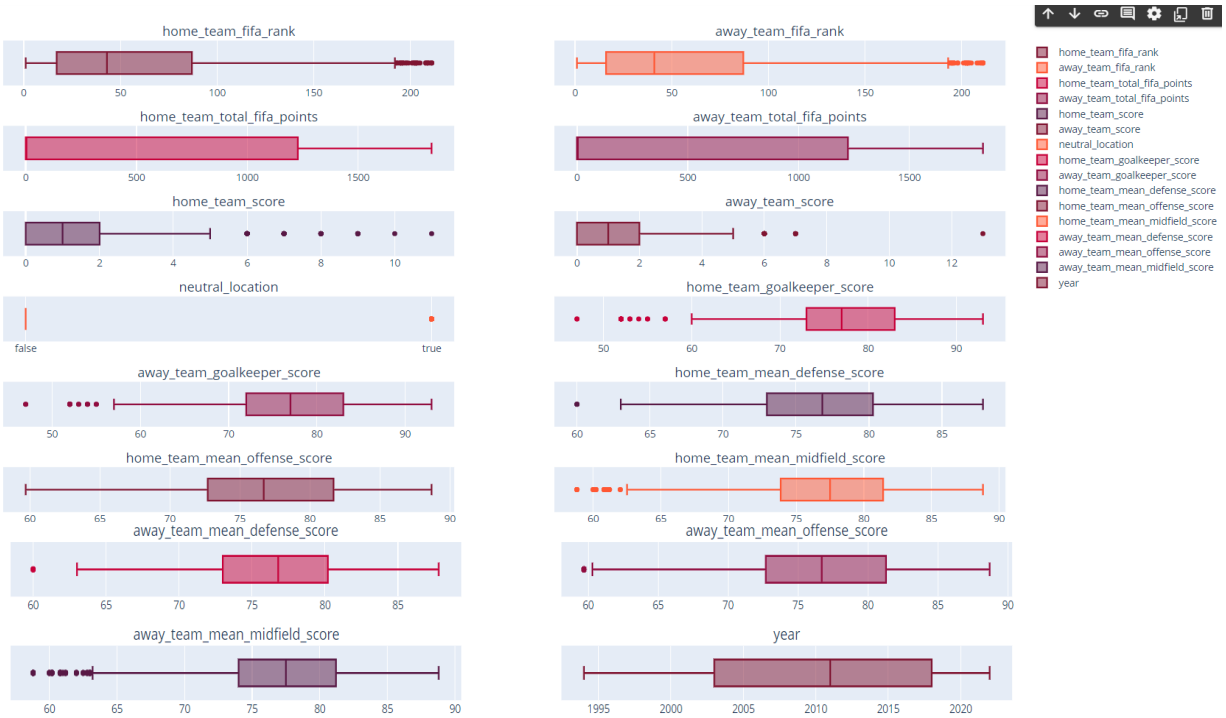

Fig. 6. Identifying similarities, patterns or relationships between different variables.

Fig. 7. Team points distribution with time.

Multiple visualizations between team points distribution with time is illustrated in Fig. 7. Overall point distribution for home and away teams are depicted in top histograms and the variation in average points per year for home and away teams are depicted in the bottom histograms.

5) Data analysis and transformation: To make the dataset more useful for analysis, we added a few new columns to summarize the performance of the home team. These columns—home_win, home_draw, and home_lose—were created based on the results of each match. This made it easier to see whether the home team won, drew, or lost a game, simplifying the dataset for comparison across different teams.

We also added several other columns to provide a richer, team-level view of the data, including:

a) Total points for home and away teams: This is to capture the sum of FIFA points for each team across all matches.

b) Average FIFA points per team: By averaging the FIFA points of both the home and away teams, we got a clearer idea of the overall strength of each team.

c) Team rankings and performance metrics: We introduced columns like the median FIFA rank, home and away goal scores, and goals conceded to offer more detailed insights into how each team performed.

Additionally, we included information on the continent each home team comes from, allowing us to identify any geographical trends in performance. A custom function helped us find the most frequent (mode) continent for each team.

To address missing data, we used the backward fill (bfill) method, which ensured that any gaps were filled with the most recent available data. This was especially useful in cases like carrying forward a goalkeeper's score for the next match if the data was incomplete.

VI. Comparative Analysis of Manual ML and AutoML

A. Case 1 – Using Manual Machine Learning

1) Model selection and training: Google Colaboratory, a free cloud-based environment, is the platform used for writing and running the Python code, inclusive of machine learning models. The technique used here is One-Hot encoding. This technique or program is made more scalable by creating a generic function that will fit the data and forecast the result based on the chosen methods. The accuracy levels and models' correctness are also specified for each model instance.

Algorithm 1 – Random Forest Classifier

Supervised machine learning uses ensemble technique to improve model performance by combining multiple classifiers. This approach boosts forecasting accuracy by using multiple decision trees on different datasets, utilizing feature randomization and bagging [32]. For Random Forest, data pre-processed with One-Hot Encoding yields an accuracy of 70%, as illustrated in Fig. 8.

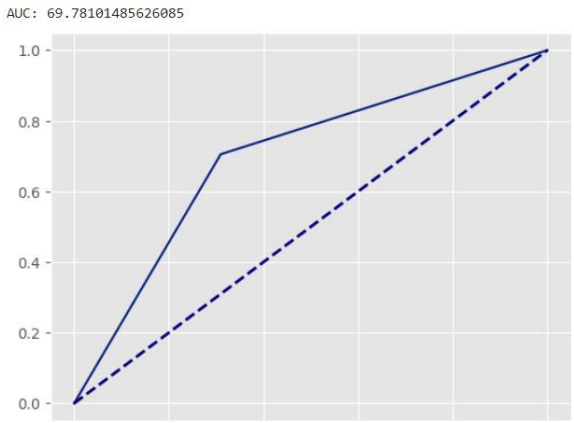

Fig. 8. Model evaluation for random forest.

Algorithm 2 – XG Boost Classifier

Sequential decision trees use XGBoost, assigning weights to independent variables. The second decision tree is used after adjusting the weight of miscalculated components, allowing faster training of large datasets through parallel processing [33]. As illustrated in Fig. 9, XGBoost reported 70% accuracy for data that has undergone One Hot Encoding pre-processing.

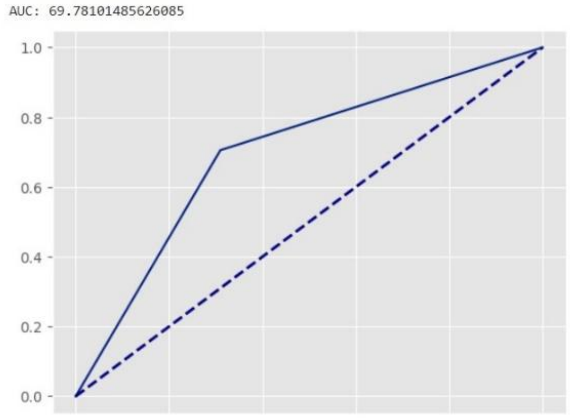

Fig. 9. Model evaluation for XG boost.

Algorithm 3 – Support Vector Machine

This model uses supervised learning algorithms to solve regression, detecting outliers and complex classification by executing optimal data transformations that set boundaries between data points on predefined classes or labels. This model has an accuracy of 71% as shown in Fig. 10.

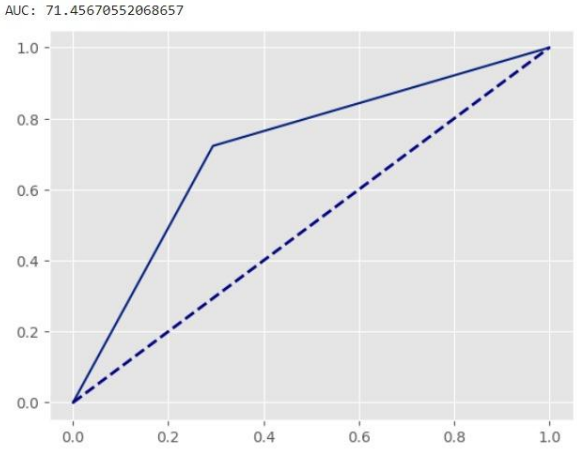

Fig. 10. Model evaluation for SVM.

Algorithm 4 – AdaBoost Classifier

AdaBoost is an iterative ensemble boosting classifier that combines inefficient classifiers to increase precision. It can be trained on a dataset, but its main drawback is hindering parallelization. It requires interactive training on various weighted instances and limiting training errors for perfect matches [34].

Fig. 11 illustrates the AdaBoost Classifier accuracy for data pre-processed with One-Hot Encoding, which was 71%.

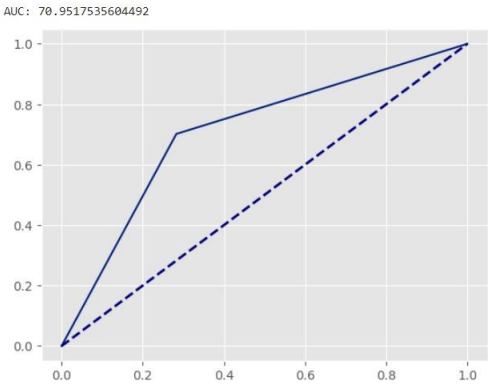

Fig. 11. Model evaluation for AdaBoost.

Algorithm 5 – Logistic Regression

This is a data analysis technique which is used to find out the dependency or relationship between two data factors. This relationship is then further used to determine or predict the value of the other factor. This results in a finite number of outcomes. Fig. 12 illustrates the accuracy check for this model, which is 70%.

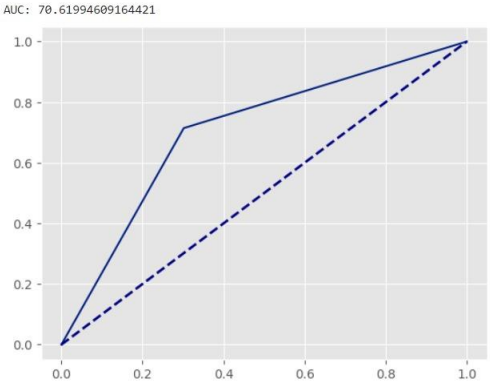

Fig. 12. Model evaluation for logistic regression.

Algorithm 6 – K-Nearest Neighbour Classifier

KNN is a supervised learning algorithm, which is non-parametric and is used in both classification as well as regression. Refer to Fig. 13.

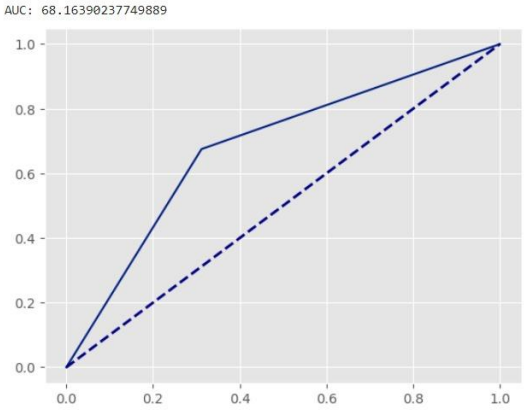

Fig. 13. Model evaluation for KNN.

Algorithm 7 – Gaussian Naive Bayes

This is a machine learning classification technique which is based on a probabilistic approach. Here, each class is assumed to follow a normal distribution. Refer to Fig. 14.

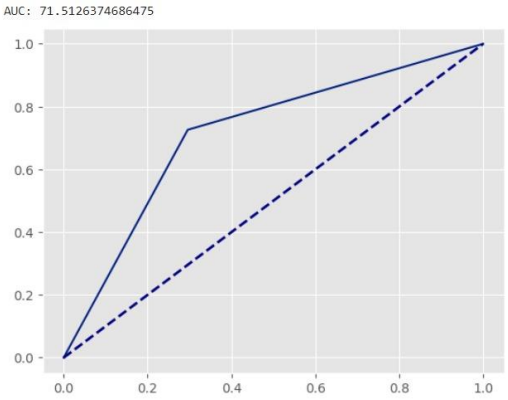

Fig. 14. Model evaluation for gaussian naive bayes.

1) Model Evaluation and Visualization

a) Model simulation: Table I, shows a machine learning model calculating the "home_team" and "away_team" means for a tournament. The simulation is run 1000 times, saving outcomes in variables for each round, quarter-finals, semi-finals, and finals. The model calculates the home team's victory probability and predicts match outcomes. The simulation continues until the tournament winner is determined.

Model 1 –Logistic Regression is the model used in this instance.

Out of the sixteen teams, eight teams were chosen for the quarter-finals, while the other 8 teams. From the eight teams, four were chosen to go to the semi-finals, while the other four teams failed. For the remaining rounds, the same marking procedure is used. Refer to Table II, where the green ones show the winning teams in each round and the pink ones are the failed ones.

TABLE I. ML MODEL SIMULATION OF TEAM MEANS AND MATCH OUTCOME PREDICTIONS

Away Team	Away Team FIFA Rank	Team Total FIFA Points
Albania	73.166667	492.977444
Austria	34.675	776.971667
Belgium	19.133333	922.177778
Croatia	18.576923	647.143086
Czech Republic	23.966102	542.766882
Denmark	16.48	580.456254
England	8.111111	760.278236
France	6.076923	675.307329
Georgia	86.560976	409.585366
Germany	6.351852	609.736055
Hungary	49.282609	580.877127
Italy	9.350877	641.229123
Netherlands	8.468085	684.159543
Poland	26.837209	634.290698
Portugal	11.12963	731.470994
Romania	23.313725	550.907308
Scotland	37.869565	556.342391
Serbia	29.655172	822.627949
Slovakia	34.666667	668.602163
Slovenia	57.065217	516.256522
Spain	6.067797	677.954132

TABLE II. USING LOGISTIC REGRESSION

Round 16	Quarter-Finals	Semi-Finals	Finals
Czech Republic	England	France	Portugal
Denmark			
Italy			
France	Italy		
Switzerland			
Portugal			
England	France	Portugal	
Netherlands			
Germany			
Spain	Portugal		
Belgium			
Romania			
Ukraine	Netherlands	Italy	France
Slovakia			
Croatia			
Scotland	Denmark		
Poland			
Serbia			
Turkey	Czech Republic	England	
Hungary			
Austria			
Albania	Switzerland		
Slovenia			
Georgia			

The UEFA European Championship was won by Portugal among the two teams, while France finished in second place. The UEFA European Championship 2024 has been won by Portugal, according to the Logistic Regression model.

Model 2 – Random Forest is the model used in this instance.

Out of the sixteen teams, eight teams were chosen for the quarter-finals, while the other 8 teams failed. 4 of the 8 teams were qualified to the semi-finals, and the other four teams were unsuccessful. The remaining rounds are marked using the same criteria. Refer to Table III, where the green ones show the winning teams in each round and the pink ones are the failed ones.

TABLE III. USING A RANDOM FOREST ALGORITHM

Round 16	Quarter-Finals	Semi-Finals	Finals
Italy	France	Portugal	Portugal
Netherlands			
Germany			
France	Italy		
Ukraine			
Czech Republic			
Portugal	Portugal	Switzerland	
Denmark			
Switzerland			
Spain	Switzerland		
Belgium			
England			
Serbia	Germany	Italy	Switzerland
Slovakia			
Croatia			
Hungary	Netherlands		
Slovenia			
Turkey			
Georgia	Czech Republic	France	
Romania			
Austria			
Scotland	Denmark		
Poland			
Albania			

The UEFA European Championship was won by Portugal amongst the two teams, while Switzerland finished in second place. The UEFA European Championship 2024 has been won by Portugal, according to the Random Forest model.

Model 3 – Gaussian Naive Bayes is the model used in this instance.

8 out of 16 teams qualified for the quarter-finals, while the other 8 teams failed. Out of the eight teams, four were chosen to go to the semi-finals, while the other four teams did not advance. For the remaining rounds, the same marking procedure is used. Refer to Table IV, where the green ones show the winning teams in each round and the pink ones are the failed ones

TABLE IV. USING GAUSSIAN NAÏVE BAYES

Round 16	Quarter-Finals	Semi-Finals	Finals
Czech Republic	Italy	England	Portugal
Italy			
Switzerland			
Portugal	Portugal		
Denmark			
France			
Netherlands	France	Portugal	
England			
Germany			
Belgium	England		
Ukraine			
Spain			
Croatia	Netherlands	Italy	England
Romania			
Slovakia			
Scotland	Denmark		
Serbia			
Poland			
Austria	Switzerland	France	
Slovenia			
Turkey			
Hungary	Czech Republic		
Albania			
Georgia			

The UEFA European Championship was won by Portugal, and England finished as the runner-up. The UEFA European Championship 2024 has been won by Portugal, according to the Gaussian Naive Bayes model.

Model 4 – XG Boost is the model used in this instance.

8 out of 16 teams qualified for the quarter-finals, while the other 8 teams failed. Out of the eight teams, four were chosen to go to the semi-finals, while the other four teams did not advance. For the remaining rounds, the same marking procedure is used. Refer to Table V, where the green ones show the winning teams in each round and the pink ones are the failed ones.

TABLE V. USING XG BOOST

Round 16	Quarter-Finals	Semi-Finals	Finals
Italy	Netherlands	France	Portugal
Germany			
France			
Portugal	Portugal		
Czech Republic			
Netherlands			
Ukraine	Belgium	Portugal	
Belgium			
Spain			
Denmark	France		
Switzerland			
England			
Serbia	Italy	Netherlands	France
Hungary			
Slovenia			
Croatia	Ukraine		
Slovakia			
Romania			
Poland	Germany	Belgium	
Georgia			
Turkey			
Austria	Czech Republic		
Scotland			
Albania			

The UEFA European Championship was won by Portugal, and France finished as the runner-up. The UEFA European Championship 2024 has been won by Portugal, according to the XG Boost model.

Model 5 – SVM is the model used in this instance.

8 out of 16 teams qualified for the quarter-finals, while the other 8 teams failed. Out of the eight teams, four were chosen to go to the semi-finals, while the other four teams did not advance. For the remaining rounds, the same marking procedure is used. Refer to Table VI, where the green ones show the winning teams in each round and the pink ones are the failed ones.

TABLE VI. USING SVM

Round 16	Quarter-Finals	Semi-Finals	Finals
Czech Republic	Netherlands	Netherlands	Italy
Italy			
Denmark			
France	Switzerland		
Netherlands			
Portugal			
Switzerland	Italy	Italy	
Germany			
Belgium			
England	Germany		
Spain			
Ukraine			
Romania	Portugal	Germany	Netherlands
Croatia			
Serbia			
Slovakia	France		
Scotland			
Hungary			
Turkey	Denmark	Switzerland	
Poland			
Slovenia			
Austria	Czech Republic		
Albania			
Georgia			

The UEFA European Championship was won by Italy, and the Netherlands finished as the runner-up. The UEFA European Championship 2024 has been won by Italy, according to the SVM model.

Model 6 – KNN is the model used in this instance.

8 out of 16 teams qualified for the quarter-finals, while the other 8 teams failed. Out of the eight teams, four were chosen to go to the semi-finals, while the other four teams did not advance. For the remaining rounds, the same marking procedure is used. Refer to Table VII, where the green ones show the winning teams in each round and the pink ones are the failed ones

TABLE VII. USING KNN

Round 16	Quarter-Finals	Semi-Finals	Finals
Germany	Portugal	Netherlands	Netherlands
Italy			
France			
Portugal	Netherlands		
Czech Republic			
Netherlands			
Belgium	Switzerland	Germany	
Switzerland			
Croatia			
Serbia	Germany		
Ukraine			
Denmark			
Slovenia	Italy	Switzerland	Germany
Romania			
England			
Spain	Belgium		
Hungary			
Slovakia			
Poland	Czech Republic	Portugal	
Turkey			
Scotland			
Austria	France		
Albania			
Georgia			

The UEFA European Championship was won by the Netherlands, and Germany finished as the runner-up. The UEFA European Championship 2024 has been won by the Netherlands, according to the KNN model.

Model 7 – AdaBoost is the model used in this instance.

8 out of 16 teams were qualified to the quarter-finals, while the other 8 teams failed. Out of the eight teams, four were chosen to go to the semi-finals, while the other four teams did not advance. For the remaining rounds, the same marking procedure is used. Refer to Table VIII, where the green ones show the winning teams in each round and the pink ones are the failed ones.

TABLE VIII. USING ADABOOST

Round 16	Quarter-Finals	Semi-Finals	Finals
Czech Republic	Italy	France	France
England			
Switzerland			
Turkey	France		
France			
Portugal			
Slovenia	England	Portugal	
Italy			
Poland			
Croatia	Portugal		
Belgium			
Germany			
Scotland	Switzerland	England	Portugal
Slovakia			
Spain			
Albania	Turkey		
Netherlands			
Denmark			
Ukraine	Czech Republic	Italy	
Austria			
Romania			
Hungary	Slovenia		
Serbia			
Georgia			

The UEFA European Championship was won by France, and Portugal finished as the runner-up. The UEFA European Championship 2024 has been won by France, according to the AdaBoost model.

In order to forecast the winner of the UEFA European Championship 2024, 7 models were utilised. The total results are depicted in Table IX.

TABLE IX. WINNER PREDICTION USING MANUAL MACHINE LEARNING

Model	Winner	Runner-Up
Logistic Regression	Portugal	France
Random Forest	Portugal	Switzerland
Gaussian Naive Bayes	Portugal	England
XG Boost	Portugal	France
SVM	Italy	Netherlands
KNN	Netherlands	Germany
AdaBoost	France	Portugal

B. Case 2 – Using AutoML

For performing AutoML, the pycaret library is to be installed in Google Colab.

In automated machine learning, while setting up the environment, the module itself runs a series of pre-processing and data transformation steps. After the environment is set up, the performance metrics of various classification models are evaluated on the transformed data. All the pre-processing information regarding AutoML is given in Table X.

TABLE X. PRE-PROCESSING AND SETUP

Description	Value
Target	is_won
Target Type	Binary
Original Data Shape	(2391, 43)
Transformed Data Shape	(2391, 47)
Train Set Shape	(1673, 47)
Test Set Shape	(718, 47)
Numeric Features	26
Categorical Features	10
Preprocess	True
Imputation Type	Simple
Numeric Imputation	Mean
Categorical Imputation	Mode
Maximum One-Hot Encoding	25
Encoding Method	None
Fold Generator	StratifiedKFold
Number of Folds	10
CPU Jobs	-1 (All CPUs)
Use GPU	False
Log Experiment	False
Experiment Name	clf-default-name

AutoML itself identifies the best model for this particular dataset. In this case, Logistic Regression is chosen as the best model, as using the 'lbfgs' solver, the Logistic Regression model was optimized with typical L2 regularization (penalty='l2') to avoid overfitting. The intercept term (fit_intercept=True) was included in the model, and the regularization strength was adjusted to 1 (C=1.0). A maximum of 1000 iterations were performed, with a tolerance value of 0.0001 to guarantee convergence. Reproducibility was ensured by using random_state=6250, and the imbalance was handled without the use of class weights (class_weight=None). For the classification challenge, this setup produced a reliable and effective model.

Fig. 15 shows the plotted ROC curve and the confusion matrix for the logistic regression model.

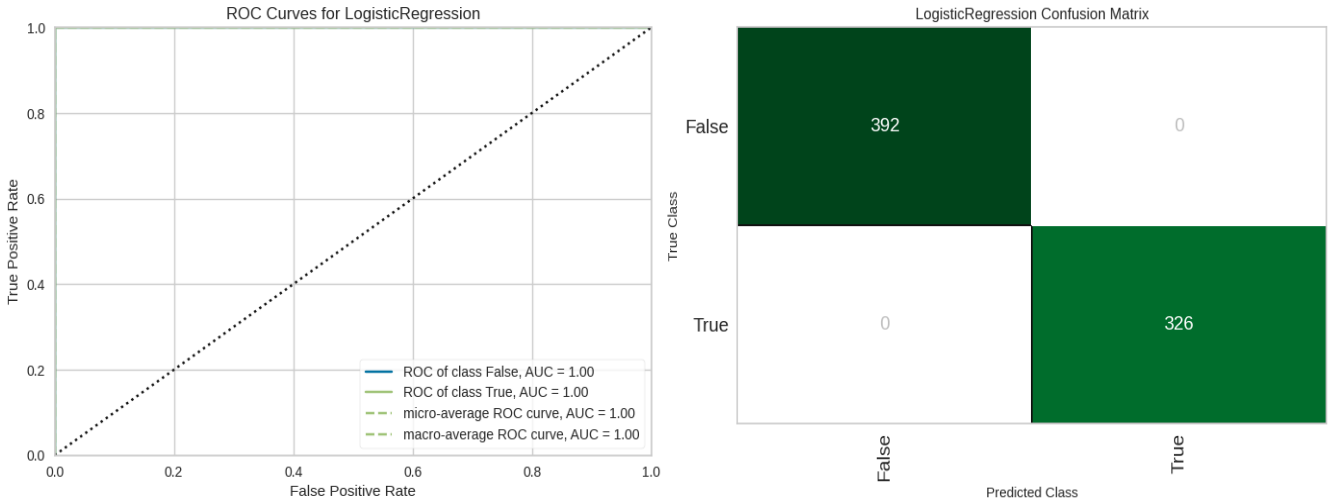

Fig. 15. ROC curve and confusion matrix.

Model Chosen by Auto ML – Logistic Regression

1) Logistic regression model simulation: Applying the same simulation logic for autoML as well (same as used in manual ML). 8 out of 16 teams qualified for the quarter-finals, while the other 8 teams failed. Out of the eight teams, four were chosen to go to the semi-finals, while the other four teams did not advance. For the remaining rounds, the same marking procedure is used. Refer to Table XI, where the green ones show the winning teams in each round and the pink ones are the failed ones.

TABLE XI. USING AUTOML (LOGISTIC REGRESSION)

Round 16	Quarter-Finals	Semi-Finals	Finals
Czech Republic	Denmark	Italy	Portugal
Denmark			
Italy			
Switzerland	Portugal		
France			
England			
Netherlands	England	Portugal	
Portugal			
Germany			
Belgium	Italy		
Croatia			
Ukraine			
Romania	Switzerland	England	Italy
Spain			
Slovakia			
Scotland	France		
Turkey			
Poland			
Serbia	Czech Republic	Denmark	
Austria			
Hungary			
Albania	Netherlands		
Slovenia			
Georgia			

The UEFA European Championship was won by Portugal, and Italy finished as the runner-up. The UEFA European Championship 2024 has been won by Portugal, according to the Logistic Regression model via AutoML.

VII. COMPARATIVE RESULTS AND EVALUATION

Table XII given below compares the performance of various classification models based on the key metrics such as precision, recall, F1-score, accuracy, and AUC (Area under the curve) for the manual approach. It evaluates both models' ability to predict positive and negative classes and also identifies the more effective model. Here, 0 represents the "False" class and 1 represents the "True" class.

TABLE XII. SUMMARY OF CLASSIFICATION PERFORMANCE METRICS (MANUAL)

Metric		Random Forest	XG Boost	SVM	AdaBoost	Logistic Regression	K-Nearest Neighbour	Gaussian Naive Bayes
Precision	0	0.77	0.77	0.78	0.75	0.78	0.73	0.79
	1	0.61	0.61	0.63	0.67	0.62	0.63	0.63
Recall	0	0.69	0.69	0.71	0.72	0.70	0.69	0.70
	1	0.71	0.71	0.72	0.70	0.71	0.68	0.73
F1-Score	0	0.73	0.73	0.74	0.73	0.74	0.71	0.74
	1	0.65	0.65	0.68	0.68	0.66	0.65	0.67
Support	0	213	213	211	198	212	202	213
	1	146	146	148	168	147	157	146
AUC	-	69.78	69.78	71.46	70.95	70.62	68.16	71.51

1) Precision: From the precision values it can be measured that how many of the predicted "True" (1) cases were correctly classified. For the "False" (0) class, most models have similar precision, with Gaussian Naive Bayes performing the best at 0.79, followed closely by K-Nearest Neighbor (KNN) and AdaBoost. In terms of classifying the "True" (1) class, AdaBoost achieves the highest precision at 0.67 where we can say that it more accurately identifies true positives compared to the other models.

2) Recall: Recall measures how well the model identifies all actual "True" cases. From the table we observe that for the "False" (0) class, recall values vary around 0.69 to 0.72 which shows that the models are consistent in recognizing the "False" cases. For the "True" (1) class, XGBoost, Random Forest, and SVM exhibit similar recall values which is around 0.71, indicating that they are effective at correctly identifying positive cases.

3) F1-Score: Generally, F1-Score provides a balanced view of model performance. We can see that for the "False" (0) class, most models have similar F1-scores, with values ranging from 0.71 to 0.74. This means that the models perform well in identifying negative cases, particularly Gaussian Naive Bayes, AdaBoost, and SVM. For the "True" (1) class, F1-scores are slightly lower, with values ranging from 0.65 to 0.68. This suggests that predicting "True" cases is more challenging for these models. SVM and AdaBoost perform slightly better in this regard, with F1-scores of 0.68.

4) Support: Support refers to the number of actual instances in each class. There are more "False" (0) cases (213 instances) than "True" (1) cases (146 instances) in the dataset which may indicate an imbalance in the class distribution.

5) AUC (Area under the curve): AUC represents the model's ability to distinguish between classes. So, higher values indicate better performance. Gaussian Naive Bayes achieves the highest AUC at 71.51, indicating it is the best at distinguishing between "False" and "True" cases. XGBoost and Random Forest have identical AUC values of 69.78, suggesting similar performance in overall classification accuracy.

The overall summary of Manual ML and Auto ML is depicted in Table XIII.

TABLE XIII. OVERALL SUMMARY OF AUTOML AND MANUAL ML

Manual ML			AutoML	
Algorithm Name	Accuracy	Prediction Result	List of Algorithms chosen	Ridge (99%), Random Forest (100%), XG Boost (100%), Decision Tree (100%), SVM (65%), Ada Boost (100%), Logistic Regression (100%), KNN (68%), Naïve Bayes (100%), Quadratic Discriminant Analysis (100%), Linear Discriminant Analysis (99%), Extra Trees Classifier (99.94%), Extreme Gradient Boosting (100%), Light Gradient Boosting Machine (100%) and Dummy Classifier (54.63%)
AdaBoost	71%	Using AdaBoost, Winner – France 1st Runner Up - Portugal	Best Model chosen via AutoML	Logistic Regression
Random Forest	70%	Using Random Forest, Winner – Portugal 1st Runner Up - Switzerland	Prediction Result	The best Model chosen through AutoML was Logistic Regression. Using Logistic Regression, Winner – Portugal 1st Runner Up - Italy
XG Boost	70%	Using XG Boost, Winner – Portugal 1st Runner Up - France		
SVM	71%	Using SVM, Winner – Italy 1st Runner Up - Netherlands		
Logistic Regression	70%	Using Logistic Regression, Winner – Portugal 1st Runner Up - France		
KNN	68%	Using KNN, Winner – Netherlands 1st Runner Up - Germany		
Naive Bayes	71%	Using Gaussian Naive Bayes, Winner – Portugal 1st Runner Up - England		

VIII. CONCLUSIONS

This study shows how both manual and automated machine learning (AutoML) techniques can effectively predict football match outcomes. By using a comprehensive dataset of historical match data and applying various ML algorithms, we created models that significantly improve the accuracy and reliability of sports predictions. We found that AutoML models, especially logistic regression, offered better predictive accuracy than traditional manual methods. AutoML streamlined the model selection and tuning process, making predictive analysis more efficient and less reliant on manual intervention. AutoML proved it could optimize ML model performance by automating key steps like data pre-processing, feature selection, and hyperparameter tuning.

Manual ML techniques, while effective, required more effort and expertise to match the results achieved by AutoML. Manual methods like Random Forest, XGBoost, SVM, and AdaBoost performed well but were more time-consuming and needed more domain-specific knowledge. Our findings highlight the importance of thorough data preprocessing and feature engineering in boosting model performance. Using cross-validation techniques and hyperparameter optimization further improved the models' accuracy and robustness, ensuring they are applicable to real-world scenarios.

Additionally, this research provided valuable insights into the factors that influence football match outcomes. This knowledge is invaluable for sports industry stakeholders, including analysts, coaches, and betting agencies, giving them a powerful tool for strategic decision-making.

In summary, this study demonstrated the effectiveness of both manual and AutoML techniques in sports analytics, paving the way for broader adoption and innovation. The results suggest that AutoML can greatly enhance the efficiency and effectiveness of predictive modelling in sports. Future research could incorporate diverse data sources and extend these methods to other sports, showcasing the versatility and scalability of machine learning.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used the QuillBot tool [https://quillbot.com/grammar-check] to check grammar as well as paraphrasing. After using this tool/service,

the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

REFERENCES

[1] "European Championship | History, Winners, & Facts | Britannica," www.britannica.com. https://www.britannica.com/sports/European-Championship.

[2] J. Hucaljuk and A. Rakipović, "Predicting football scores using machine learning techniques," 2011 Proceedings of the 34th International Convention MIPRO, Opatija, Croatia, 2011, pp. 1623-1627.

[3] J. D. Rose, M. K. Vijaykumar, U. Sakthi, and P. Nithya, "Comparison of Football Results Using Machine Learning Algorithms," IEEE Xplore, Jul. 01, 2022. https://ieeexplore.ieee.org/document/9914265.

[4] Groll, Andreas & Ley, Christophe & Schauberger, Gunther & Eetvelde, Hans & Zeileis, Achim. (2019). Hybrid Machine Learning Forecasts for the FIFA Women's World Cup 2019.

[5] A. Basit, M. B. Alvi, F. H. Jaskani, M. Alvi, K. H. Memon, and R. A. Shah, "ICC T20 Cricket World Cup 2020 Winner Prediction Using Machine Learning Techniques," IEEE 23rd International Multitopic Conference (INMIC), Nov. 2020, doi: https://doi.org/10.1109/inmic50486.2020.9318077.

[6] Tekade P, Markad K, Amage A, Natekar B. Cricket match outcome prediction using machine learning. International journal of Advance Scientific Research and Engineering Trend, 2020 July, 5(7).

[7] J. Kumar, R. Kumar and P. Kumar, "Outcome Prediction of ODI Cricket Matches using Decision Trees and MLP Networks," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018, pp. 343-347, doi: 10.1109/ICSCCC.2018.8703301.

[8] Daniel Mago Vistro, F. Rasheed, and Leo Gertrude David, "The Cricket Winner Prediction With Application Of Machine Learning And Data Analytics," International Journal of Scientific & Technology Research, vol. 8, no. 9, pp. 985–990, Sep. 2019.

[9] H. Elmiligi and S. Saad, "Predicting the Outcome of Soccer Matches Using Machine Learning and Statistical Analysis," IEEE Xplore, Jan. 01, 2022. https://ieeexplore.ieee.org/document/9720896.

[10] A. Majumdar, R. Kaur, T. Kulkarni, M. Jiruwala, S. Shah, and N. Pise, "Football Match Prediction using Exploratory Data Analysis & Multi-Output Regression," IEEE Xplore, Dec. 01, 2022. https://ieeexplore.ieee.org/abstract/document/10119340.

[11] A. V. P, R. D, and S. N. S. S, "Football Prediction System using Gaussian Naïve Bayes Algorithm," IEEE Xplore, Mar. 01, 2023. https://ieeexplore.ieee.org/document/10085510/authors#authors.

[12] Jeremiah Samson Chin, Filbert Hilman Juwono, Ing Ming Chew, S. Sivakumar, and W. K. Wong, "Predicting Ice Hockey Results Using Machine Learning Techniques," Jul. 2023, doi: https://doi.org/10.1109/icdate58146.2023.10248726.

[13] Dhananjay Daundkar and Kundan Kandhway, "Predicting Winner of a Professional Basketball Match," Oct. 2023, doi: https://doi.org/10.23919/iccas59377.2023.10316903.

[14] M. Vashist, V. Bahl, N. Sengar, and A. Goel, "Machine Learning for Football Matches and Tournaments," IEEE Xplore, May 01, 2022. https://ieeexplore.ieee.org/document/9850673.

[15] E. Tiwari, P. Sardar and S. Jain, "Football Match Result Prediction Using Neural Networks and Deep Learning," 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2020, pp. 229-231, doi: 10.1109/ICRITO48877.2020.9197811.

[16] M, Jaeyalakshmi & S, Indrajith & C, Hirthik & K, Kaushiik & S, Eaknath. (2023). Predicting the outcome of future football games using machine learning algorithms. 1-7. 10.1109/RMKMATE59243.2023.10370000.

[17] Amitesh Peddii and R. Jain, "Random Forest-Based Fantasy Football Team Selection," Mar. 2023, doi: https://doi.org/10.1109/icaccs57279.2023.10113019.

[18] M. A. AL-ASADI and S. Tasdemir, "Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques," IEEE Access, vol. 10, pp. 1–1, 2022, doi: https://doi.org/10.1109/access.2022.3154767.

[19] A. M. Emam, O. Tarek Ali and A. Atia, "Football activities classification," 2023, Intelligent Methods, Systems, and Applications (IMSA), Giza, Egypt, 2023, pp. 520-525, doi: 10.1109/IMSA58542.2023.10217464.

[20] F. Rodrigues and Â. Pinto, "Prediction of football match results with Machine Learning," Procedia Computer Science, vol. 204, pp. 463–470, 2022, doi: https://doi.org/10.1016/j.procs.2022.08.057.

[21] M. Gifford and Tuncay Bayrak, "A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression," Decision Analytics Journal, pp. 100296–100296, Aug. 2023, doi:https://doi.org/10.1016/j.dajour.2023.100296.

[22] K. Chauhan et al., "Automated Machine Learning: The New Wave of Machine Learning," 2020, 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 2020, pp. 205-212, doi: 10.1109/ICIMIA48430.2020.9074859.

[23] Singh, V. K., & Joshi, K. (2022). Automated Machine Learning (AutoML): an overview of opportunities for application and research, Journal of Information Technology Case and Application Research, 24(2), 75–85. https://doi.org/10.1080/15228053.2022.2074585.

[24] Truong, A., Walters, A., Goodsitt, J., Hines, K., Bruss, C. B., & Farivar, R. (2019, November). Towards automated machine learning: Evaluation and comparison of AutoML approaches and tools. In 2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI), pp. 1471-1479. IEEE.

[25] L. Ferreira, A. Pilastri, C. M. Martins, P. M. Pires, and P. Cortez, "A Comparison of AutoML Tools for Machine Learning, Deep Learning and XGBoost," 2021 International Joint Conference on Neural Networks (IJCNN), Jul. 2021, doi: https://doi.org/10.1109/ijcnn52387.2021.9534091.

[26] Elshawi, R., Maher, M., & Sakr, S. (2019). Automated machine learning: State-of-the-art and open challenges. arXiv preprint arXiv:1906.02287.

[27] Nagarajah, Thiloshon, and Guhanathan Poravi. "A review on automated machine learning (AutoML) systems." In 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), pp. 1-6. IEEE, 2019.

[28] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, "Implementing AutoML in Educational Data Mining for Prediction Tasks," Applied Sciences, vol. 10, no. 1, p. 90, Dec. 2019, doi: https://doi.org/10.3390/app10010090.

[29] X. Shi, Y. D. Wong, C. Chai, and M. Z.-F. Li, "An Automated Machine Learning (AutoML) Method of Risk Prediction for Decision-Making of Autonomous Vehicles," IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 11, pp. 7145–7154, Nov. 2021, doi: https://doi.org/10.1109/tits.2020.3002419.

[30] Göksu, Semih & Sezen, Bulent & Balcioglu, Yavuz. (2024). Predicting the Uefa Euro 2024 Winner: An Artificial Neural Network Approach.

[31] Mahadinour48, "International football matches," Kaggle.com, 2023. https://www.kaggle.com/datasets/mahadinour/international-football-matches (Accessed Jul. 29, 2024).

[32] Simplilearn, "Random Forest Algorithm," Simplilearn.com, Nov. 07, 2023. https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm. (Accessed Jul. 29, 2024).

[33] GeeksforGeeks, "XGBoost," GeeksforGeeks, Sep. 18, 2021. https://www.geeksforgeeks.org/xgboost/.

[34] Prashant111, "AdaBoost Classifier Tutorial," Kaggle.com, Apr. 30, 2020. https://www.kaggle.com/code/prashant111/adaboost-classifier-tutorial/notebook (Accessed Jul. 29, 2024).