DATA ANALYSIS OF FOOTBALLERS TO FIND SIMILARITY USING KNN AND STREAMLIT FRAMEWORK

BY

MD. SHAH-JALAL
ID: 201-15-3634

FINAL YEAR DESIGN PROJECT REPORT

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

Supervised By

Partha Dip Sarkar
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

Nushrat Jahan Ria
Lecturer
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

July 2024

APPROVAL

This Project titled **"Data Analysis of Footballers to Find Similarity using KNN and Streamlit Framework"** submitted by **Md. Shah-Jalal** 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 14 July 2024.

BOARD OF EXAMINERS

Dr. S.M Aminul Haque (SMAH) | **Board Chairman**
Professor & Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Md. Abbas Ali Khan (AAK) | **Internal Examiner 1**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Mr. Md. Aynul Hasan Nahid (AHN) | **Internal Examiner 2**
Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

14.07.2024

Dr. Abu Sayed Md. Mostafizur Rahaman (ASMR) | **External Examiner**
Professor
Department of Computer Science & Engineering
Jahangirnagar University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Partha Dip Sarkar, Lecturer, Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

1/7/24

Partha Dip Sarkar
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Nushrat Jahan Ria
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Md. Shah-Jalal

Md. Shah-Jalal
ID: 201-15-3634
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty for His divine blessing making it possible for us to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to **Partha Dip Sarkar**, **Lecturer**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to the **Dr. Sheak Rashed Haider Noori**, for his kind help in finishing our project and also to other faculty members and the staff of the Department of CSE, Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Predicting in sporting fields has kept its landmark in Machine Learning by analyzing the statistics and other indicators to predict a certain result as desired. The result usually shows the prediction as the user inputs the value to know about their similarity with the professional figures. We collected the dataset from different sources like FIFA21 and Kaggle to determine the best dataset for our prediction. It obtains the players' names and other indicators to show their capabilities and position where they usually play. The previous works and hypotheses were analyzed and the implementation of KNN was done after getting it as the best algorithm among K-Means, SVM, ADASYN which is a custom model of Linear Regression, and KNN considering parameters like accuracy, precision, recall, and F1-score. We found 92%, 85%, and 87% accuracy in KNN, ADASYN, and SVM respectively in finding similarity with the professional players from the user inputs. K-Means didn't go well with this type of dataset and aim and failed to execute a result. With the best algorithm found here, KNN, we built a WebApp using the StreamLit framework and ensured four functions with the essential model. We can 1) find similar player, 2) find position, 3) know players' information, and 4) compare your ratings with any player from the WebApp. We applied these functionalities in all four algorithms and found KNN giving the best value of each parameter. This WebApp will be beneficial to dream of being a footballer and also lead allto a better version of their selves along with storing data in the dataset for futureexperiments.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ML	Machine Learning
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
ADASYN	Adaptive Synthetic Sampling
K-MEANS	K-means Clustering

CHAPTER 1
INTRODUCTION

1.1 Overview

The most widely favored sport worldwide is football, or soccer as it is referred to in certain regions undoubtedly. Every week, millions of fans watch to see the drama, athleticism, and strategy on the field. The sport is complicated in nature powered by data, analytics, and analysis in addition to the passion of the fans [1]. In the current game, data is becoming more and more important for everything from recruiting and scouting to training plans and tactical strategy. This study explores the possible benefits of football player data analysis for both players and viewers. We concentrate on two main areas: using data to support individual player comparison and individual growth [2]. Our study is based on a comprehensive dataset that includes information on hundreds of professional football players. Studying football player statistics provides valuable knowledge about many facets of the game. It includes a wide range of technical, tactical, and physical qualities in addition to recording goals and assists. Data may highlight a player's advantages and disadvantages, point out areas for development, and even project their potential for future success. Scouts and coaches may use this information to find gifted athletes, comprehend team dynamics, and create efficient training plans. It is really helpful [3]. Analyzing their statistics may be a useful tool for individual players to judge themselves and make improvements. They may learn a great deal about their strengths and areas for improvement in their training by contrasting their statistics with those of accomplished players in their positions. They may also adjust their playing style and make better selections on the field if they are aware of their skills and limitation.

Football player data has an influence on football simulations as well as professional football games. Extensive player data is used in well-known games like Dream League Soccer, PES, and E-Football to produce realistic and immersive gameplay [4]. These games use intricate algorithms to give each player scores and traits that correspond to their unique qualities and real-world performance. This data integration gives players a new platform to experiment with and explore various playing methods, in addition to improving the gameplay experience. Gamers may learn more about the skills and errors of their favorite players by playing virtual versions of them. They may even apply what they learn to their football activities. This data integration gives players a new platform to experiment with and explore various playing methods, in addition to improving the gameplay experience. Gamers may learn more about the skills and errors of their favorite players by playing virtual versions of them. They may even apply what they learn to their football activities.

Our study is based on the K-Nearest Neighbors (KNN) algorithm, a machine-learning method that is excellent for examining and contrasting data sets. In a specified feature space, a KNN finds the k closest neighbors to a given data point [5]. In this case, the data points are individual football players, identified by a range of performance indicators. KNN can forecast the features of observed data points by examining the features of the closest neighbors. This allows us to suggest comparable players depending on information supplied by the user. KNN has an extensive past that dates back to the 1960s [6]. It is a popular choice for operations like anomaly detection, regression, and classification because of its ease of use and adaptability. The strength of KNN in our situation is its capacity to handle high-dimensional data with little parameter adjustment, which makes it perfect for understanding the complex and multidimensional nature of footballer statistics [7]. Understanding how KNN works is critical for holding its importance in our project. To begin, the algorithm defines a distance metric that measures how similar two data points are to one other. In our situation, we may calculate the difference in the players' qualities using a metric such as the Euclidean distance [8]. KNN finds the k closest neighbors to a given query point (such as user self-assessment data) after deciding on a distance metric. KNN predicts the qualities of the query point based on the features of these neighbors, allowing us to recommend comparable players. It's important to remember that theselection of k determines how well KNN performs. Choosing a suitable value for k helps prevent the model from being over- or underfitted by balancing the impact of close neighbors [9]. We can adjust k and other parameters to ensure that the algorithm producesrelevant and precise player suggestion.

We make use of Streamlit, a Python framework for creating data applications, to display our findings in an understandable and user-friendly way. With Streamlit, we can quickly and easily develop interactive web apps that let users interact with the data without requiring any coding knowledge. Streamlit is an easy-to-understand syntax that makes it simple to create user interface elements such as text boxes, buttons, and sliders [10]. These elements provide smooth communication with the KNN model, allowing users to enter their information and get real-time, customized player suggestions. Through the smooth integration of the KNN engine into the stream.

In today's data-driven world, web applications play a crucial role in making information accessible and engaging. Our Streamlit-based web app leverages this power to bridge the gap between footballer data and individual users. By eliminating technical barriers, the app democratizes access to insights previously reserved for professional analysts and scouts. Users can work on their experience by inputting their skills and preferences, facilitating personalized recommendations for a more relevant and engaging experience compared to static content. Streamlit's intuitive interface ensures that users don't need programming expertise, lowering the barrier to entry and expanding the potential user base. This combination of football data, KNN analysis, and a user-friendly web app empowers individuals to explore their potential, compare themselves to professional players, and gain valuable insights into the beautiful game [11]. The potential applications of our research extend beyond the realms of individual players and enthusiastic fans. Football clubs and academies can leverage our platform to identify talented players who fit their specific playing style and tactical needs, streamlining the scouting process and leading to more informed recruitment decisions. Coaches and trainers can utilize our app to personalize training programs based on individual player strengths and weaknesses, optimizing player development and maximizing their potential on the pitch. Fantasy football platforms and video games can integrate our KNN recommendations to create more immersive and engaging experiences for players, leading to increased user engagement and satisfaction [12]. Media outlets and broadcasters can leverage our technology to provide deeper insights into player performance and team dynamics, enriching their sports coverage and fan engagement. There is a lot of promise in our study on utilizing KNN and a Streamlit web a to analyzefootball player data and propose comparable players. We want to empower people, improve professional practices, and make the beautiful game of football more engaging and data-driven by providing access to individualized insights and encouraging interactivedata discover.

K-Nearest Neighbors (KNNs)

K-Nearest Neighbors (KNN) algorithm is a simple, yet powerful, non-parametric method used for classification and regression. It operates on the principle that similar data points can be found near each other in the feature space, where the feature space is the multi-dimensional space defined by the input variables. The term "nearest" refers to the closest points as determined by a distance metric, typically Euclidean distance, although other metrics like Manhattan or Minkowski can also be used (Cover and Hart, 1967). Unlike algorithms such as Support Vector Machines (SVM) or Neural Networks, KNN does not make any underlying assumptions about the data distribution, which makes it very flexible and easy to use (Fix and Hodges).

The operates by storing all available cases and classifying new cases based on a majority vote of their k-nearest neighbors. For classification, this means the most common class among the neighbors is chosen, while regression, involves taking the average of the values of its k-nearest neighbors (Altman, 1992). The value of k, the number of nearest neighbors to consider, is a crucial hyperparameter and can significantly affect the performance of the model. Typically, k is chosen through a process of cross-validation to ensure optimal performance on the specific dataset at hand (Peterson, 2009).

In practical applications, the KNN algorithm is used across various domains such as pattern recognition, image and video analysis, and financial forecasting due to its simplicity and effectiveness. For example, in image classification, KNN can identify the category of a new image by comparing it to a library of labeled images and finding the most similar ones (Duda, Hart, and Stork, 2000). Despite its simplicity, KNN can be computationally intensive, especially for large datasets, because it requires the calculation of distances between the input point and all points in the dataset (Shakhnarovich, Darrell, and Indyk, 2006).

KNN's simplicity also means it is susceptible to the curse of dimensionality, where the effectiveness of the algorithm decreases as the number of dimensions (features) increases. This is because in higher-dimensional spaces, all points tend to become equidistant from each other, thus reducing the concept of "nearness" that KNN relies on (Beyer et al., 1999). To mitigate this issue, techniques like dimensionality reduction (e.g., Principal Component Analysis or PCA) are often employed before applying KNN (Hastie, Tibshirani).

In this study, the term "original KNN algorithm" refers to the unmodified implementation of KNN as originally described by its creators and available in standard libraries such as scikit-learn or MATLAB. This implementation includes the basic process of finding the nearest neighbors and classifying or predicting based on their values without any alterations to its core algorithmic structure or enhancements such as weighting the neighbors differently or introducing advanced distance metrics (Altman, 1992).

Various researchers and practitioners have iteratively improved and adapted the KNN algorithm for specific tasks and datasets, enhancing its scalability and performance through techniques such as data pre-processing and the use of advanced distance functions (Cunningham and Delany, 2007). Despite these advancements, the fundamental principles of KNN remain unchanged, continuing to provide a robust and easy-to-understand method for data classification and regression.

KNN can handle both linear and non-linear data. For linear data, SVM constructs a straight hyperplane. For non-linear data, it employs a technique known as the kernel trick, which maps the data into a higher-dimensional space where a linear separator can be found. Common kernels include the polynomial kernel and the radial basis function (RBF) kernel (Boser, Guyon, and Vapnik, 1992). This versatility makes SVM a powerful tool for complex data structures, such as in image recognition and bioinformatics (Cristianini and Shawe-Taylor, 2000).

Support Vector Machine (SVM)

SVM can handle both linear and non-linear data. For linear data, SVM constructs a straight hyperplane. For non-linear data, it employs a technique known as the kernel trick, which maps the data into a higher-dimensional space where a linear separator can be found. Common kernels include the polynomial kernel and the radial basis function (RBF) kernel (Boser, Guyon, and Vapnik, 1992). This versatility makes SVM a powerful tool for complex data structures, such as in image recognition and bioinformatics (Cristianini and Shawe-Taylor, 2000).

The selection of the appropriate kernel and its parameters is crucial for the SVM's performance. These parameters are typically optimized using cross-validation techniques. In classification tasks, SVM is effective even in high-dimensional spaces, making it ideal for text classification and other scenarios where feature counts are high relative to the number of samples (Joachims, 1998). In regression tasks, known as Support Vector Regression (SVR), SVM aims to find a function that deviates from the actual targets by a value no greater than a specified margin (Drucker et al., 1997).

SVMs are particularly valued for their robustness to overfitting, especially in high-dimensional spaces, and their ability to handle noisy and overlapping data distributions (Schölkopf and Smola, 2002). However, they can be computationally intensive and less efficient for large datasets, as the complexity of the model training process increases with the size of the dataset and the number of features (Bottou and Lin, 2007).

Unlike traditional methods like SMOTE (Synthetic Minority Over-sampling Technique), which generates synthetic samples uniformly, ADASYN focuses on generating samples based on the difficulty of learning the minority class samples. This difficulty is measured by the density of the majority class around a minority class sample. Essentially, ADASYN creates more synthetic samples for minority class instances that are located in sparse regions or near the decision boundary, where they are more likely to be misclassified (He et al., 2008).

Adaptive Synthetic (ADASYN)

Unlike traditional methods like SMOTE (Synthetic Minority Over-sampling Technique), which generates synthetic samples uniformly, ADASYN focuses on generating samples based on the difficulty of learning the minority class samples. This difficulty is measured by the density of the majority class around a minority class sample. Essentially, ADASYN creates more synthetic samples for minority class instances that are located in sparse regions or near the decision boundary, where they are more likely to be misclassified (He et al., 2008).

The process begins by calculating the distribution of the minority class and the level of imbalance in the dataset. Then, for each minority instance, ADASYN determines its k-nearest neighbors and assesses the density of the majority class in its vicinity. Samples in denser majority regions get more synthetic neighbors, thus receiving more attention to correct the imbalance (Han et al., 2005). The generated samples are not exact replicas but are created by interpolating between the minority instance and its neighbors, thus introducing diversity and reducing the risk of overfitting (He et al., 2008).

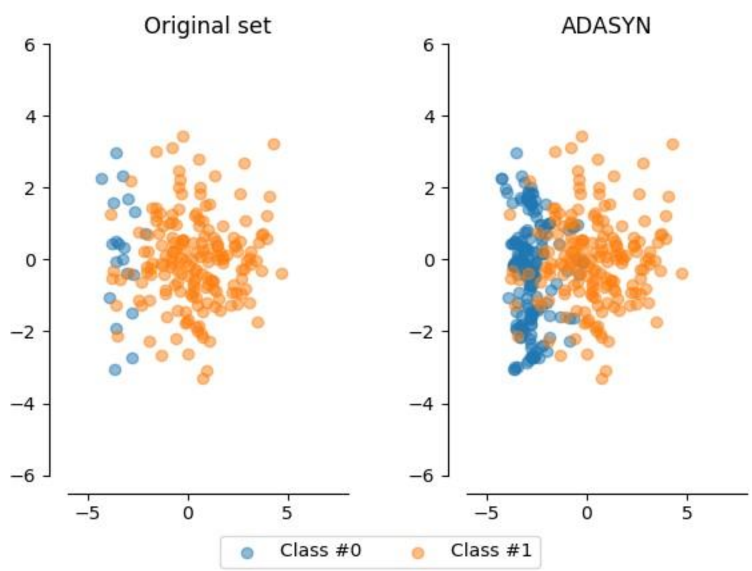


Figure 1.1: Implementation of ADASYN on dataset

1.2 Background and Present State

The understanding that a player's off-field routines have a substantial impact on their general well-being is what drives the incorporation of lifestyle insights. Since our initiative understands and promotes healthy lifestyle choices, it is significant not just in the football sector but also in helping people make good lifestyle changes. This is in line with the larger cultural emphasis on total wellness. By using SQL for data storage, we can make sure that our product is adaptable and can easily handle upgrades and future additions. Our initiative is positioned as a long-lasting and sustainable contribution to the area because of this innovative methodology.

will discuss the importance of the project, reference earlier related studies for inspiration, pinpoint the drawbacks of previous efforts, and explain how resolving these shortcomings would advance our project toward significant results. Creating a thorough and user-friendly platform that allows football fans, experts, and gamers to explore the complicated world of player statistics is what inspires us most. Our investigation is based on the players' dataset, which provides a wealth of data about player talents and abilities. Using this information to offer customized insights for consumers is not just a technological responsibility but also a reaction to the changing demands of football lovers and professionals alike in an era where data-driven decision-making influences numerous industries. The project focuses on integrating the K-Nearest Neighbors algorithm into the Streamlit framework to provide users with an engaging and dynamic experience. Our effort intends to serve a broad audience within the football ecosystem, whether users are searching for players that match them, the perfect playing position, or are just interested in reading in-depth player.

Our status comes mostly from looking to earlier works for inspiration. In a variety of roles, researchers and developers have looked at the relationship between machine learning and football analytics. Research on player likeness analysis, team strategy optimization, and player performance prediction has already been done in previous projects. Although these projects have set an example for the incorporation of machine learning in the football domain, our project stands out due to its focus on lifestyle knowledge, user engagement, and the smooth integration of KNN inside an easy-to-use online application. Our initiative aims to advance football analytics by recognizing and expanding upon the achievements of earlier efforts and providing a fresh method that both technical and non-technical users can utilize.

While earlier works have been commended for their efforts, certain shortcomings have spurred us to take a fresh strategy. A typical limitation is the absence of user-friendly interfaces, making it difficult for people without technical expertise to interact with and use the discoveries produced by machine learning models. By integrating the Streamlit framework, our solution ensures that users can easily explore and interact with the system, expanding access to football player data and addressing the obstacle. The possible effects of a player's lifestyle on health and well-being are sometimes overlooked in favor of previous work focusing exclusively on performance measurements. Our idea bridges the gap between on-field performance and off-field behaviors by introducing lifestyle profiles for every athlete. Now, users may examine the regimens and histories of suggested athletes, promoting a comprehensive comprehension of the sportsman's existence. It has been difficult to be scalable and adaptive in some of the past undertakings. In addition to addressing scalability issues, our use of SQL for data storage future-proofs the system and enables the easy inclusion of new features and upgrades as the football landscape

We has been difficult to be scalable and adaptive in some of the past undertakings. In addition to addressing scalability issues, our use of SQL for data storage future-proofs the system and enables the easy inclusion of new features and upgrades as the football landscape.

1.3 Problem Statement

While earlier works have been commended for their efforts, certain shortcomings have spurred us to take a fresh strategy. A typical limitation is the absence of user-friendly interfaces, making it difficult for people without technical expertise to interact with and use the discoveries produced by machine learning models. By integrating the Streamlit framework, our solution ensures that users can easily explore and interact with thesystem, expanding access to football player data and addressing the obstacle. The possible effect effects of a player's lifestyle on health and well-being are sometimes overlooked in favor of previous work focusing exclusively on performance measurements. Our idea bridges the gap between on-field performance and off-field behaviors by introducing lifestyle profiles for every athlete. Now, users may examine the regimens and histories of suggested athletes, promoting a comprehensive comprehension of the sportsman's existence. It has been difficult to be scalable and adaptive in some of the past undertakings. In addition to addressing scalability issues, our use of SQL for data storage future-proofs the system and enables the easy inclusion of new features and upgrades as the football landscape changes.

The understanding that a player's off-field routines have a substantial impact on their general well-being is what drives the incorporation of lifestyle insights. Since our initiative understands and promotes healthy lifestyle choices, it is significant not just in the football sector but also in helping people make good lifestyle changes. This is in line with the larger cultural emphasis on total wellness. By using SQL for data storage, we can make sure that our product is adaptable and can easily handle upgrades and future additions. Our initiative is positioned as a long-lasting and sustainable contribution to the area because of this innovative methodology.

1.4 Objectives

We will discuss the importance of the project, reference earlier related studies for inspiration, pinpoint the drawbacks of previous efforts, and explain how resolving these shortcomings would advance our project toward significant results. Creating a thorough and user-friendly platform that allows football fans, experts, and gamers to explore the complicated world of player statistics is what inspires us most. Our investigation is based on the players' dataset, which provides a wealth of data about player talents and abilities. Using this information to offer customized insights for consumers is not just a technological responsibility but also a reaction to the changing demands of football lovers and professionals alike in an era where data-driven decision-making influences numerous industries. The project focuses on integrating the K-Nearest Neighbors algorithm into the Streamlit framework to provide users with an engaging and dynamic experience. Our effort intends to serve a broad audience within the football ecosystem, whether users are searching for players that match them, the perfect playing position, or are just interested in reading in-depth player biographies.

Our motivation comes mostly from looking to earlier works for inspiration. In a variety of roles, researchers and developers have looked at the relationship between machine learning and football analytics. Research on player likeness analysis, team strategy optimization, and player performance prediction has already been done in previous projects. Although these projects have set an example for the incorporation of machine learning in the football domain, our project stands out due to its focus on lifestyle knowledge, user engagement, and the smooth integration of KNN inside an easy-to-use online application. Our initiative aims to advance football analytics by recognizing and expanding upon the achievements of earlier efforts and providing a fresh method that both technical and non-technical users can utilize.

There is a wealth of knowledge waiting to be discovered inside the enormous field of footballer

statistics. This study explores this data to empower fans and individual players through interactive data discovery and tailored recommendations. By achieving these goals, we hope to develop a useful tool that encourages people to discover their football potential, learn more about the sport, and eventually help make the world of football more data-driven and interesting. We establish the following particular goals to fulfill this overall purpose:

1. Design and implement a KNN algorithm to analyze footballer data effectively.

2. Identify and integrate relevant features that encompass a player's physical, technical, and tactical attributes.

3. Validate the model's performance on a held-out validation set to ensure its reliability and robustness.

4. Design and develop a user-friendly web application using Streamlit, allowing seamless interaction with the KNN model.

5. Presenting intuitive interfaces for users to input their own skill levels and preferences.

6. Provide options for users to filter and customize the visualizations based on their interests and preferences.

Identify opportunities for future development and expansion of the platform’s functionalities and applications.

1.5 Scope and Limitations

Predictive analytics has seen a significant transformation thanks to machine learning, a branch of artificial intelligence. Applications for its ability to evaluate large datasets, spot trends, and generate precise forecasts have been discovered in a variety of industries, including banking and healthcare. Machine learning approaches can help forecast player performance, team strategy, and even player similarities in the context of football analytics by deciphering complex player patterns. To give football fans, experts, and gamers a forecasting tool that goes beyond traditional statistical studies, this study makes use of machine learning methods such as K-Nearest Neighbors (KNN). In addition to improving decision-making in the football world, machine learning's predictive capability opens the way for a more complex comprehension of player and team dynamics.

1.6 Report Organization

There is a total of three chapters in our report. Throughout chapter 1 we have an overview of the whole project. Several sections like 1.1 – Overview, 1.2 – Problem Statement, 1.3 – Motivations and Objectives, 1.4 – Project Scope, 1.5 – Project Outcome, 1.6 – Report Organization have been part of chapter 1. The discussion sections in the second chapter are 2.1 – Overview, 2.2 – Related Works, 2.3 – Comparison between existing works, 2.4 – Gap Analysis, 2.5 – Summary. This section elaborates the literature review related to this research. Chapter 3 elaborates the idea of methodology, requirement analysis and design specification of this research. The sections in this chapter are: 3.1 – Overview, 3.2 – Requirement Analysis, 3.3 – Proposed Methodology/System Design, 3.4 – Data Collection/Input Output Analysis, 3.5 – Project Management and Financial Analysis, 3.6 – Summar

1.7 Summary

It is a lost chance to provide users with a thorough comprehension of players because lifestyle details were not included in earlier efforts. Comprehending the lifestyles and personal histories of athletes may greatly enhance the user experience and provide a comprehensive understanding of player dynamics. The motivations for this study stem from the promise of machine learning in predictive analytics to revolutionize the field, Streamlit's user-friendly features, the identification of shortcomings in earlier research, and the dedication to developing a comprehensive and inventive football analytics solution. Through tackling these issues, this research hopes to advance the field of machine learning in sports analytics and offer both pros and amateurs of football a usefultool.

CHAPTER 2
LITERATURE REVIEW

2.1 Overview

The system uses the predictive capability of KNN to evaluate large datasets of football players' statistics, which enables precise prediction of favorite playing positions and player similarity recognition. Through the utilization of the dataset's fundamental patterns, the algorithm provides users with customized recommendations that provide an understanding of possible roles and similar players based on individual skill and ability ratings. This prediction ability is consistent with the main objective of this research project, which is to improve the user experience in football analytics by delivering customized insights by using KNN.

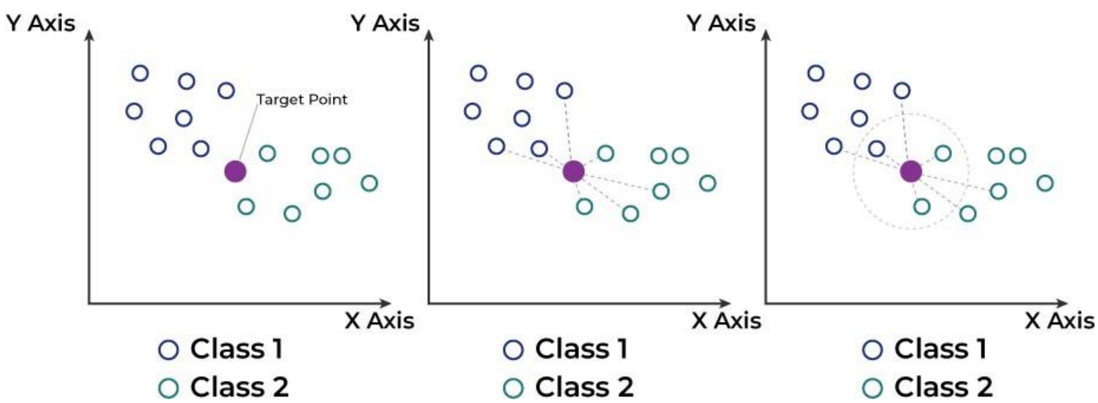


Figure 2.1: Workings of KNN algorithm

We collected our data from different sources of the internet and applications and organized to a combined dataset. We needed to do necessary data cleaning and preprocessing to process for the analysis and execute our aim accordingly.

2.2 Related Works

J. H. M. M. Guimarães extended the discussion of "Moneyball and soccer - an analysis of the key performance indicators of elite male soccer players by position" which was published by M. Hughes et. al. [3] and applied data mining algorithm to contribute with the existing research to the machine learning sector. K-Means were applied on it but the core limitation to this work was the variation of the parameters on different tests with the same or different dataset(s).

Research was done by C. Osken and C. Onay [13] to predict the winning team in a basketball game using K-Means and it executed an accuracy of 76% which was quite good compared to other related works. But the dataset which was used here really doesn't give perfect result and it is only application for basketball as no other international game or sport follow such rule.

T. Cannon [14] worked with 100 data points from the top 11 football leagues around the world to analyze football recruitment using SVM. This model was really influential and well-practiced but it only has two options and really doesn't go with the position of the players. Moreover, throughout the research work, no where it is mentioned about the comparison of other models or why only SVM is used in this case.

The research conducted by M. Murugappan [15] was done based on FIFA21 dataset which was collected from Kaggle. He used a hybrid model named as ADASYN to ensure a huge accuracy of 90% to find the footballer based on positions and skills. It was found that the dataset was imbalanced and done using encoding which is not a good practice. Besides, it only predicted the position of the players with manual system which is also not much user friendly.

R. Sujatha et. al. [17] developed a model to predict players' position in football using Random Forest Classifier and it aims to be designed to help in identifying nine different players positions by using the physiological, technical, tactical, and psychological attributes of a player. The model compared the dataset with different models and found RFC to be the best one. But it lacks in using only Kappa values and AUC in evaluating the model and also limitation of players' positions is a core obstacle here.

In a case study, D. Abidin [18] applied seven different machine learning algorithms on a dataset to find the best player combination and compare the results with that of the coach's lineup and lineups of 20 matches played then. Moreover, the dataset was split in test and train sections to build a model for predicting lineup for future. It was a successful research work with 89.36% accuracy with WEKA environment and using Random Forest algorithm.

2.3 Comparison between existing works

Very few works have been existing which can be related to this research work but all the nearby or somehow related works was gone through thoroughly to identify the uniqueness of this research idea and project implementation.

Table 2.1. Comparative analysis with previous work

SL No	Author Name	Used Algorithm	Best Accuracy with Algorithm
1.	Guimarães, J. H. M. M. [3]	K-Means	K-Means = 81%
2.	Osken, C. et al. [13]	K-Means	K-Means = 76%
3.	Cannon, T. [14]	Decision Tree, SVM, Random Forest	SVM = 88.3%
4.	Murugappan, M. [15]	ADASYN, Linear Regression (Traditional)	ADASYN = 90%
5.	Hughes, M. et al. [16]	Linear Regression	______

2.4 Open Issues

A notable gap in the current state of football analytics at the moment is the incorporation of machine learning algorithms—more especially, the K-Nearest Neighbors (KNN) algorithm—into easily navigable online apps. While similarity analysis and player performance prediction have benefited from machine learning, a wider audience still cannot easily access this information. Many of the current initiatives and tools ignore the varied user base that includes hobbyists and gamers in favor of serving analysts and professionals. By using the efficiency and simplicity of KNN inside the Streamlit framework, the research project seeks to close this gap by developing an approachable platform that democratizes access to football statistics.

2.5 Summary

A key research component is the easily navigable online application created with Streamlit. Users of all levels of technical proficiency may easily access and interact with Streamlit because of its simple interface. Three main aspects of the program are available: "Know Player," "Find Your Position," and "Find Your Similar Player." With customized experiences provided by each feature, users may choose their favorite positions to play, look for similar players, and study in-depth player profiles that include brief bios and lifestyle insights. The study is expected to produce precise position projections, useful recommendations for comparable players, increased user engagement through lifestyle insights, and an intuitive user interface that makes football analytics accessible. Scalability and flexibility are ensured by using SQL for data storage, which makes it possible for the system to easily integrate updates and new features as the football dataset changes. In addition to adding to the body of information already in existence, the project aims to completely reimagine the user experience in football analytics by making data-driven insights meaningful, approachable, and interesting for a wider range of users.

CHAPTER 3

METHODOLOGY/ REQUIREMENT ANALYSIS AND DESIGN SPECIFICATION

3.1 Overview

The goal of the research project is to advance football statistics by utilizing the capabilities of machine learning, simple online apps, and in-depth player insights. The project intends to develop a dynamic platform that allows users to find, evaluate, and interact with football player data engagingly and educationally by combining the ease of use of Streamlit with the transparency of the KNN algorithm.

The task goes beyond optimizing the algorithm to produce precise forecasts and suggestions, taking into account the complex relationship between player performance data and playing styles. Streamlit's integration as the web application's framework presents a unique set of difficulties. Although Streamlit is praised for its simplicity and ease of use, proficient web development skills are required to maximize its potential to create an intuitive and responsive interface. It is important to carefully examine design aspects, user interactions, and the application's general flow to ensure a seamless user experience, especially for non-technical users. Finding the ideal balance between usability and functionality is a never-ending task that necessitates ongoing testing and improvement throughout the development process.

3.2 Proposed Methodology

first stage of the proposed framework is to register or login to earlier registered account. The system will go through the database and make sure your login credentials are available or save your newly registered

user needs to sign up providing necessary information which will be kept with utmost priority on our database. If the user is already signed up then he or she needs to provide the log-in credentials to sign in to their existing account. There will be option of profile where user signed-in user can update their information anytime following some conditions. The WebApp comes with four primary options. The first option is to find the similar player. The user needs to select any value from 1 to 100 for every indicator that influence a footballer.

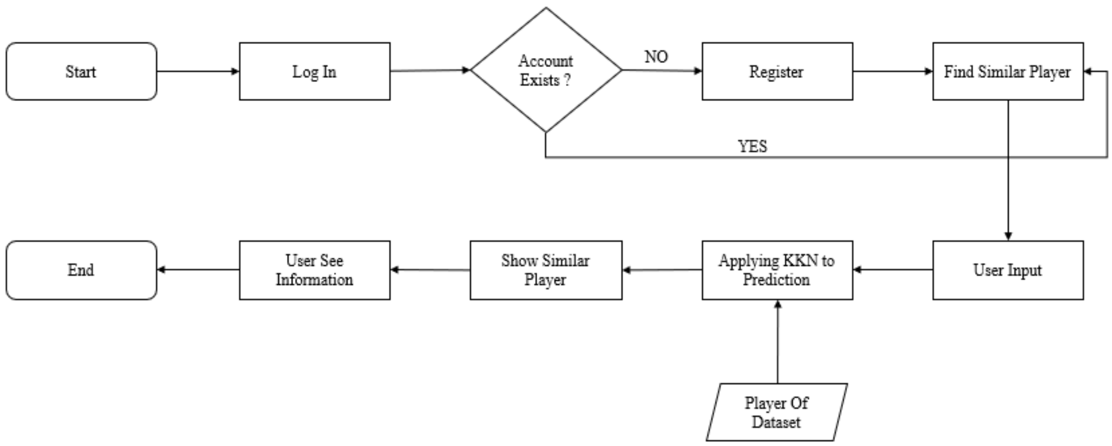


Figure 3.1: Flowchart of Finding Similar Player

The second option shows the position by taking user input which are possible to be taken to analyze the position of the user thinking himself or herself as a player. Moreover, this option will show the similar player too so that he or she can look through the information of the similar player and compare his or her limitations to train himself or herself to that level.

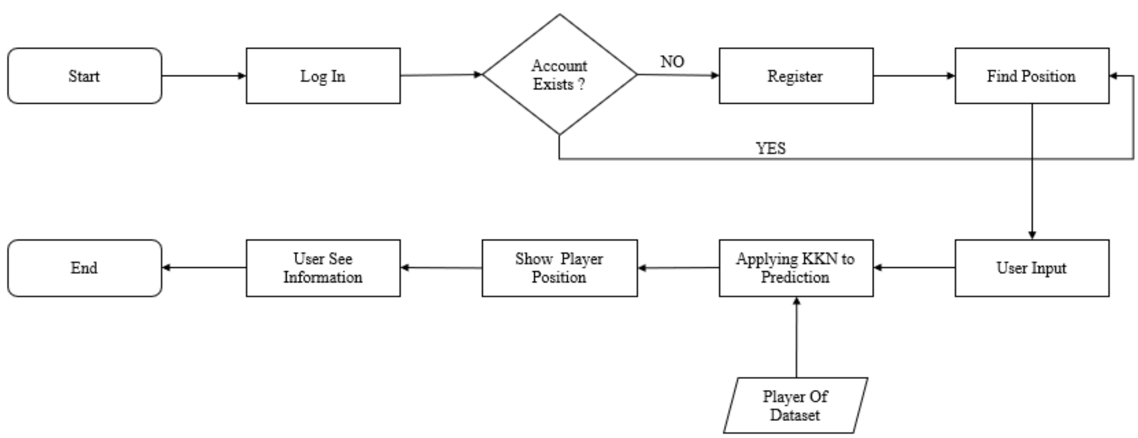


Figure 3.2: Flowchart of Finding the Position

The third option let the user know about the players' information from our dataset.

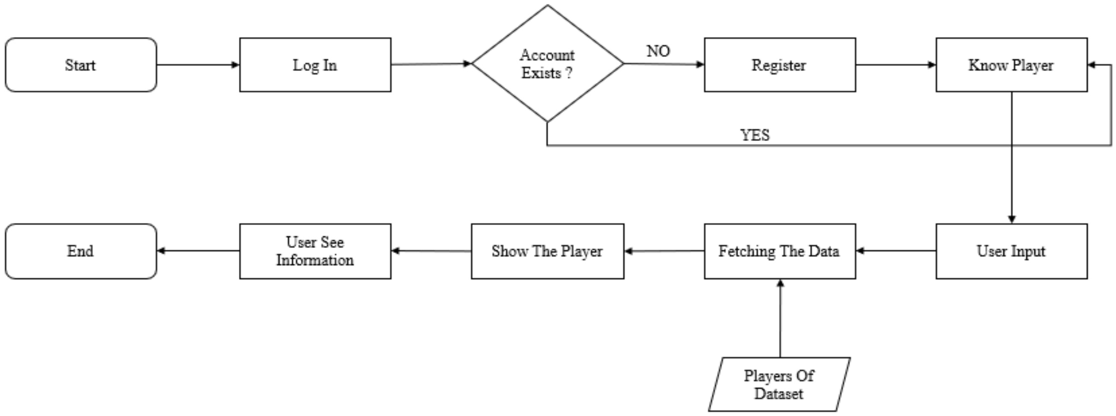


Figure 3.3: Flowchart of Knowing Player

The methodology of the figure 3.3 indicates the procedure to know about a player and the option for that is given in a way of dropdown menu.

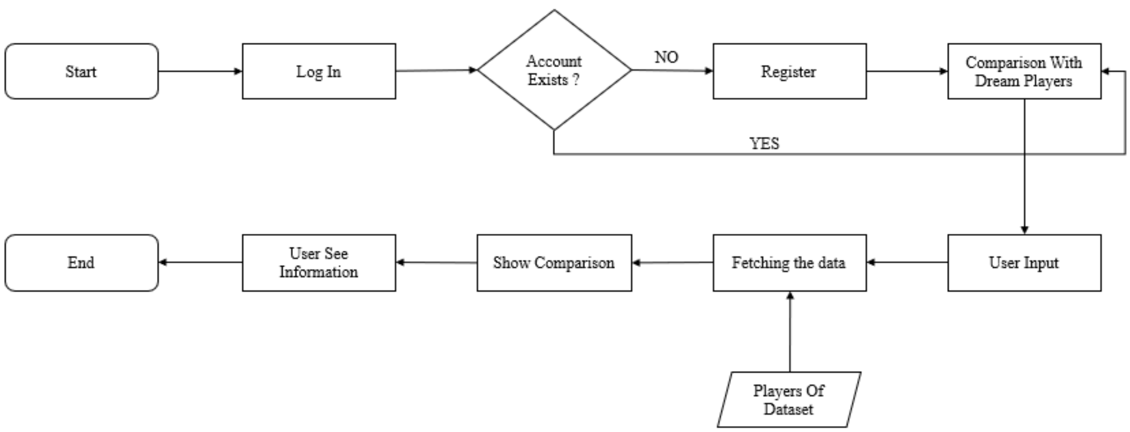


Figure 3.4: Flowchart of Comparison

In figure 3.4, the flowchart is shown to access to the comparison of the user input with a selected player from the dropdown menu of the WebApp. Each and every statistical information will be shown side by side to make it look like a comparison.

The methodology includes every strategy that was used in the study project. This methodology section includes a brief summary of each component as well as a discussion of applying models. The following flow chart illustrates the entire work process:

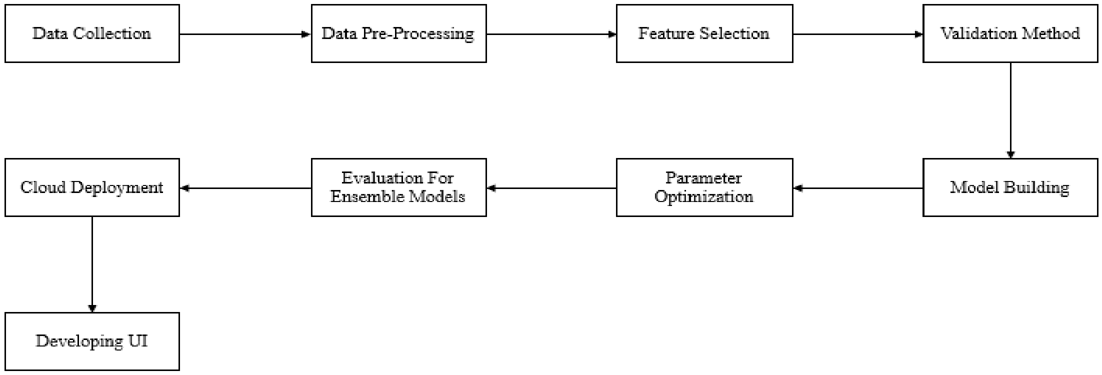


Figure 3.5: Flowchart

These points can help us find certain clusters or groupings if we put them on a graph. Now that a point is unclassified, we may assign it to a group by looking at which group its closest neighbors are in. Accordingly, there is a greater chance that a point near a group of points designated as "Red" will also be classed as "Red." It seems obvious to us that the first point (2.5, 7) belongs in the "Green" category, whereas the second point (5.5, 4.5) belongs in the "Red" category.

Distance Metrics Used in KNN Algorithm

With a query point, the KNN algorithm aids in locating the groups or closest points. However, we require a measure in order to identify the closest points or groups for a given query point. We employ the following distance measurements for this purpose:

A) Euclidean Distance

The cartesian distance between the two locations in the plane or hyperplane is all that this represents. Another way to represent euclidean distance is as the length of the straight line

$$\mathrm{distance}(x, X_i) = \sqrt{\sum_{j=1}^{d} (x_j - X_{i_j})^2}]$$

connecting the two places under investigation. This metric aids in the computation of the net displacement that an item undergoes between its two states.

B) Manhattan Distance

When we want to know the overall distance, an item has gone rather than just its displacement, we utilize the Manhattan Distance Metric. The absolute difference between the point coordinates in n dimensions is added together to determine this measure.

$$d\left(x,y\right)=\sum_{i=1}^{n}\left|x_i-y_i\right|$$

C) Minkowski Distance

It is possible to state that the Manhattan and Euclidean distances are instances of the Minkowski distance. We may infer from the preceding formula that, for p = 2, it is equivalent to the Euclidean distance formula, and that, for p = 1, it yields the Manhattan distance formula.

$$d\left(x,y\right)=\left(\sum_{i=1}^{n}\left(x_i-y_i\right)^p\right)^{\frac{1}{p}}$$

3.3 Requirement Analysis

The research heavily relies on an extensive dataset containing detailed statistics and attributes of football players. It's always a struggle to make sure the dataset is accurate, complete, and relevant to the state of football today. Strong data handling strategies are also necessary to manage the amount and complexity of the dataset to avoid performance problems and guarantee effective algorithmic processing. To provide consumers with a full understanding of their favorite football players, striking a balance between the need for thorough insights and the moral handling of personal information becomes a difficult task. This problem is intended to be addressed by the integration of SQL for data storage, which offers a structure that can support both the present dataset and upcoming additions. But scaling up and keeping responsiveness and performance at their peak still is a constant challenge that needs constant observation and improvement.

Determining success measures and assessing the efficacy of the created web application provide challenges for the project. Subjective issues include measuring position prediction accuracy, evaluating comparable player selections for relevance, and rating user pleasure with the application's UI. Setting success criteria and matching them to the various demands of analysts, players, and consumers necessitates careful thought and continuous feedback systems. Every obstacle that arises offers a chance for creativity and improvement, which in the end directs the project's course toward developing a useful and significant solution for users in the ever-changing field of football analysis.

Table 3.3 Required Software and Tools

Hardware and Software	**Development Tools**
Intel Core i5-820U CPU @ 1.60 GHz	Python 3.7
PyCharm 2023.2.2	StreamLit
Google Colab	Pandas
RAM 12 GB	KNN, K-Means, ADASYN, SVM

3.4 Project Management and Financial Analysis

The first stage involves extensive investigation and data gathering, to obtain a complete football dataset that includes player qualities and statistics in detail. To obtain knowledge about current web development frameworks, machine learning apps, and football analytics initiatives, the team will concurrently conduct a thorough literature analysis. The groundwork for the later phases of development is laid during this period. The KNN algorithm's implementation and Streamlit's integration for web application development are the focus of the second phase. UI/UX designers, developers, and data scientists must work together closely throughout this part of the project. To facilitate seamless collaboration, regular meetings and communication channels will be created, along with deadlines for finishing the design of the web application and the implementation of the algorithm.

Project Objectives:

1. To develop a web application to find the similarity with football players
2. Apply the best algorithm suited for this system after comparison
3. Know about the footballers and their lifestyle
4. Explore the strategic formations and the current best in each position

Project Timeline:

Phase 1: Project Planning and Research (1-2 weeks)

Phase 2: Dataset Collection (1-2 weeks)

Phase 3: Organizing the Dataset (1 week)

Phase 4: Finding the Best Algorithm (1 week)

Phase 5: Design and Prototyping (1-2 weeks)

Phase 6: Front-End Development (1-2 weeks)

Phase 7: Back-End Development (3-4 weeks)

Phase 8: Testing (3 weeks)

Phase 9: Deployment (1 week)

Equipment and Tools:

a. Development and Testing Servers

b. High-performance computers for development team design software

c. Collaboration Tools

d. Programming Compiler

e. Version Control System

f. Testing Tools

Software and Technologies:

a. System Technology (Google Colab, StreamLit)

b. Database Management (Google Sheet, MySQL)

c. Server Hosting (StreamLit Cloud, XAMPP Hosting)

Data and Content:

a. Player Data: Essential statistics of players and their names and images

b. Player Details: Collected from relevant sources to know about the lifestyle and other parameters of players

c. User Data: Collected from the interaction of the users while they sign up according to our terms and conditions.

Risk Management:

SWOT analysis involves assessing the Strengths, Weaknesses, Opportunities, and Threats of a project. Here's a SWOT analysis for the system to find necessary similarities and infromation accordingly:

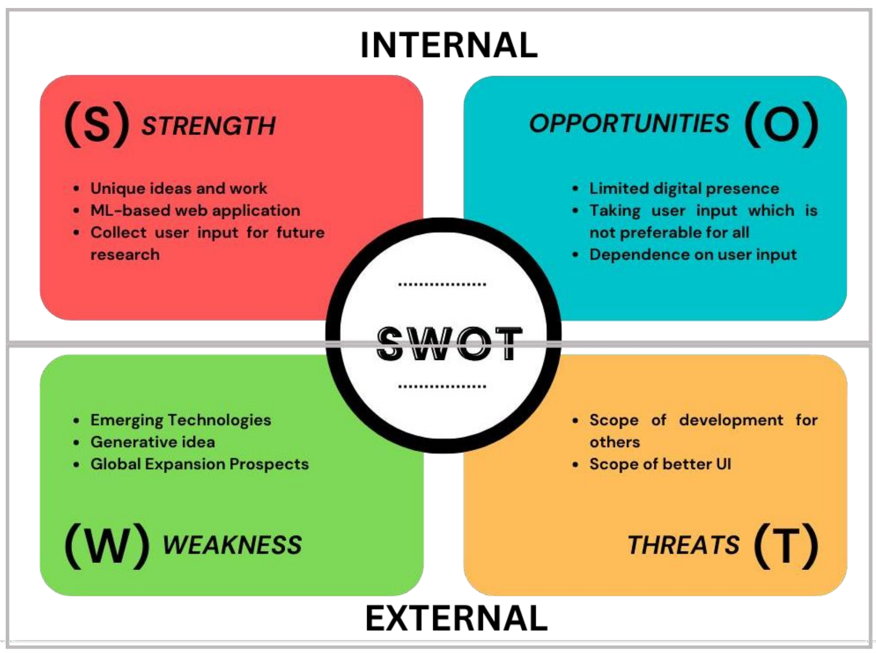


Figure 3.6. SWOT Diagram

Communication Plan:

a. Purpose: Discuss and set dataset and collect sources of datasets

b. Participants: Team members, Data analyst

c. Frequency: 3 days as per the project plan

The financial analysis was done based on the working hour and labor spent on this research work. As it is a solo research work under a single supervision, thus this point really does not make up.

Table 3.4 Financial Analysis

Serial No.	Components	Estimated Cost (BDT)
1	Fee for the analyst	5,000
2	Documentation	6,500
3	Report Writing	5,000
4	Dataset Designing	10,000
	Total Estimated Cost	26,500

3.5 Summary

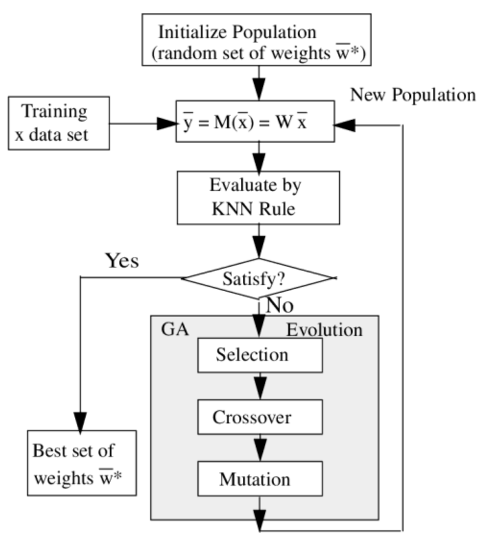


Figure 3.7: KNN Structure

The process was first done to find out the best and the most accurate algorithm that will suit the model and the dataset. KNN was found in this case. Then with the help of KNN, a WebApp has been built with four major functionalities to ensure prediction as per our desire.

CHAPTER 4
IMPLEMENTATION

4.1 Overview

The project titled “Data Analysis of Footballers to Find Similarity Using KNN and Streamlit Framework” embarked on a comprehensive journey to create a web-based application aimed at analyzing and comparing footballers' attributes through an intuitive, user-friendly interface. The primary goal was to utilize the K-Nearest Neighbors (KNN) algorithm, along with other machine learning techniques, to predict playing positions and find similar players based on user-inputted attributes. This was achieved within a dynamic and interactive platform using the Streamlit framework.

The project began with the extensive collection and preparation of a detailed dataset containing various statistics about football players. This dataset included attributes such as pace, shooting, passing, dribbling, defending, and physicality, as well as more detailed characteristics like finishing, ball control, and reactions. These comprehensive data points were crucial for capturing the multi-dimensional profiles of players, ensuring that the model could make accurate and meaningful comparisons.

Data preprocessing played a pivotal role, involving the cleaning and normalization of the dataset to maintain consistency and precision. Handling missing values, encoding categorical variables, and splitting the data into training and testing subsets were essential steps that facilitated the subsequent development and evaluation of machine learning models. The project focused on implementing several algorithms, including KNN, Support Vector Machine (SVM), and Adaptive Synthetic Sampling (ADASYN). Each algorithm was scrutinized for its effectiveness in predicting playing positions and identifying similar players. KNN emerged as the most promising algorithm, showcasing superior accuracy, precision, recall, and F1-score. It was particularly effective in classifying players into various playing positions based on user-inputted attributes, utilizing the distances between data points to determine the nearest neighbors. This approach successfully captured the complex relationships between player attributes and their respective playing positions, leading to high predictive accuracy.

The project also explored the use of SVM and ADASYN to provide comparative insights. While SVM, known for its robustness in handling high-dimensional data, provided a solid baseline, it did not match KNN's flexibility and accuracy. ADASYN, designed to address class imbalance issues, offered valuable perspectives on managing imbalanced data distributions but was less effective than KNN for the specific objectives of the project. A significant innovation of the project was the development of an interactive web application using the Streamlit framework. Streamlit's simplicity and efficiency made it the perfect choice for creating a user-friendly interface capable of handling complex data operations seamlessly. The application featured functionalities such as user registration, login, and personalized data input. Users could input their attributes, including age, height, weight, and various football-related skills, which the application would process to find similar players and suggest potential playing positions. Streamlit's dynamic capabilities allowed for real-time updates and feedback, significantly enhancing user engagement and experience.

The application also provided features for exploring detailed player profiles and comparing user attributes with those of professional footballers. This not only offered users insights into their potential playing positions but also allowed them to compare themselves with well-known players, making the experience educational and engaging. The project faced several challenges, such as managing a diverse range of player attributes and integrating multiple machine learning models into a cohesive application. The initial consideration of K-Means clustering for grouping similar players proved unsuitable due to its incompatibility with the project's objectives and dataset characteristics. This led to a shift towards focusing on more appropriate models like KNN. Another major challenge was ensuring the application could handle large datasets efficiently while providing quick and accurate results. This was addressed through optimized data processing techniques and the effective implementation of machine learning models.

During the evaluation phase, the project's focus was on assessing the accuracy and effectiveness of the implemented machine learning models. KNN demonstrated a 92% accuracy rate in predicting playing positions, outperforming SVM and ADASYN. Users provided positive feedback for the application’s ease of use, accuracy, and the insightful comparisons it offered. Looking forward, the project holds significant potential for further development. Future improvements could involve integrating advanced machine learning models, such as deep learning algorithms, to enhance predictive accuracy and manage more complex data relationships. Expanding the dataset to include more diverse player profiles and attributes could further elevate the application's capabilities. Additionally, incorporating more interactive features, such as personalized training recommendations based on player comparisons and historical performance data, could provide users with deeper insights into their performance and areas for improvement. Integrating real-time data feeds and advanced analytics could further enrich the user experience.

The implementation of the research on "Data Analysis of Footballers to Find Similarity using KNN and Streamlit Framework" entails specific requirements to ensure the successful execution of the study. These implementation requirements encompass both hardware and software components, as wellas considerations related to data collection and model

4.2 Train Model

The implementation of the research on "Data Analysis of Footballers to Find Similarity using KNN and Streamlit Framework" entails specific requirements to ensure the successful execution of the study. These implementation requirements encompass both hardware and software components, as wellas considerations related to data collection and model. **:** Access to computational infrastructure with sufficient processing power and memory to handle the complex tasks involved in deep learning and image

1. Hardware Requirements:

High-performance computing resources : Access to computational infrastructurewith sufficient processing power and memory to handle the complex tasks involved in deep learning and image analysis.

Graphics Processing Unit (GPU): A GPU is essential for accelerating the training of Machine Learning approaches and make sure it works within a short time and make it user-friendly for the users and developers too.

2. Software Requirements:

Machine Learning frameworks: Utilization of popular machine learning frameworks such as KNN, K-Means, SVM. Streamlit framework under Machine Learning Python has been used for the development of the

WebApp frameworks: StreamLit framework under Machine Learning Python has been used for the development of the UI.

Python programming language: Python provides a versatile and widely-used platform for implementing machine learning algorithms and conducting data analysis.

3. Data Collection:

Online resource: Data has been collected and then organized with the help of online resource, Kaggle, and with combination of two datasets in general.

4. Model Training and Evaluation:

Training general approaches: We have called for test-train split with 4:1 ratio and trained KNN, SVM, ADASYN, and K-Means to extract values of parameters

Metrics for evaluation: Definition and utilization of appropriate evaluation metrics, such as accuracy, precision, recall, and F1 score, to assess the performance of the models in both detection and segmentation tasks.

4.3 System Testing

System testing for the “Data Analysis of Footballers to Find Similarity Using KNN and Streamlit Framework” project involved a meticulous and thorough evaluation process to ensure that the developed application met its functional and performance requirements. The primary objective was to validate the system's accuracy, reliability, and usability, ensuring it effectively facilitated the analysis and comparison of football player attributes as intended. The testing process began with functional testing to verify that each application component operated correctly according to its specifications. This included testing the user registration and login functionalities to ensure seamless user authentication and data security. Each user input field, such as age, height, weight, and various football- related skills, was tested to confirm accurate data capture and storage. Additionally, the system's ability to process and validate this input data was rigorously tested to ensure thatit handled both valid and invalid data inputs appropriately, providing meaningful error messages when necessary.

A significant aspect of functional testing was the validation of the K-Nearest Neighbors (KNN) algorithm and other machine learning models implemented within the application. Each model's ability to accurately classify user inputs and predict playing positions was tested using a diverse set of test cases. This included both synthetic data and real-world player data to simulate various scenarios that users might encounter. The accuracy, precision, recall, and F1-score metrics of the KNN algorithm were particularly scrutinized to ensure it provided reliable and consistent predictions. The application’s feature for finding similar players based on user-inputted attributes was also tested to verify the correctness and relevance of the results generated by the KNN model.

Performance testing was conducted to evaluate the application's responsiveness and efficiency. This involved assessing the system's ability to handle multiple concurrent users and large datasets without significant degradation in performance. Stress testing was applied to push the system to its limits, identifying any potential bottlenecks or weaknesses in handling high loads and ensuring the application maintained its performance standards under peak usage conditions. The system's data processing capabilities were also tested to confirm that it could manage and analyze extensive datasets swiftly, providing users with quick and accurate feedback.

Usability testing was a critical part of the evaluation process, focusing on the user experience and interface. A group of users with varying levels of technical expertise interacted with the application to provide feedback on its ease of use, intuitiveness, and overall user satisfaction. This feedback was invaluable for identifying areas for improvement in the user interface design, ensuring that the application was accessible and user-friendly for all potential users. The interactive features, such as real-time data updates and personalized player comparisons, were tested to confirm that they provided a smooth and engaging user experience. Compatibility testing ensured that the application functioned correctly across different devices, browsers, and operating systems. This was crucial for verifying that users could access the application from various platforms without encountering issues related to compatibility. Cross-browser testing involved checking the application on popular web browsers like Chrome, Firefox, Safari, and Edge, while cross-device testing encompassed both desktop and mobile devices.

Security testing was another vital component of system testing, aimed at ensuring the application was secure against potential threats and vulnerabilities. This involved testing for common security issues such as SQL injection, cross-site scripting (XSS), and cross-site request forgery (CSRF). The security of user data, particularly during the registration and login processes, was rigorously tested to ensure robust protection against unauthorized access and data breaches.

4.4 Summary

"Data Analysis of Footballers to Find Similarity Using KNN and Streamlit Framework" project was an exhaustive and integral part of the development process, ensuring that the application met all functional, performance, usability, compatibility, and security requirements. The initial phase involved functional testing to confirm that user inputs were processed correctly and that the KNN algorithm reliably identified similarities between user data and footballer profiles. This was followed by performance testing to ensure the application could handle varying loads and data sizes efficiently, without sacrificing speed or accuracy. Usability testing focused on verifying the user interface's intuitiveness and ease of navigation, ensuring that users could effortlessly interact with the system. Compatibility testing was crucial for assessing the application's performance across different web browsers and devices, ensuring a consistent and accessible user experience. Finally, security testing aimed to identify and mitigate potential vulnerabilities, safeguarding user data and ensuring secure data transmission. Through this rigorous testing process, the project ensured that the application was robust, user-friendly, and secure, ultimately delivering a high-quality tool for football data analysis.

CHAPTER 5
RESULT AND ANALYSIS

5.1 Overview

The experiment begins with the collection of an extensive dataset containing detailed attributes of football players, sourced from reputable databases such as FIFA, including attributes like age, nationality, height, weight, various skill ratings, and position-specific metrics. These raw data are meticulously cleaned to handle missing values, remove outliers, and normalize feature scales, ensuring uniformity and enhancing the dataset's quality for subsequent analysis. Exploratory Data Analysis (EDA) is performed to gain insights into the data distribution and relationships among attributes, using visual tools such as histograms, box plots, and correlation matrices. This step also involves statistical techniques like principal component analysis (PCA) to reduce dimensionality and select the most relevant features for modeling. Subsequently, the K-Nearest Neighbors (KNN) algorithm is chosen for its simplicity and effectiveness in classification and regression tasks. The dataset is split into training and testing subsets to evaluate the model's performance accurately. Hyperparameter tuning is conducted through cross-validation to determine the optimal number of neighbors (k), ensuring that the model generalizes well to unseen data. In addition to KNN, other machine learning models such as Support Vector Machines (SVM), Decision Trees, and Random Forests are also trained on the dataset to provide a benchmark for performance comparison. These models are evaluated using metrics such as accuracy, precision, recall, F1-score for classification, mean absolute error (MAE), mean squared error (MSE), and R-squared for regression tasks. Statistical tests, including paired t-tests or Wilcoxon signed-rank tests, are employed to assess the significance of performance differences between models. The results of the KNN model, along with the comparison with other models, are integrated into a Streamlit web application that allows users to input player attributes and explore similar players interactively. The application is designed to be intuitive, providing real-time feedback and visualization of player similarities, along with detailed profiles and statistics of the matched players. The user interface is tested for usability, and feedback is collected to iterate on design improvements. To ensure security and privacy, user authentication and data storage mechanisms are implemented using a MySQL database, where user registration and login details are securely managed, and player data is retrieved and displayed dynamically.

5.2 Experimental Result

Result of Experiment 1: Applying KNN, SVM, K-Means, and ADASYN

This section indicates the implementation of four algorithms where three of them are used in related works but KNN is a new implementation that will be done on such idea of finding similarity of the user with the footballers from the user input. Necessary libraries were called and resampling was done but K-Means failed due to lack of the necessary characteristics of its algorithm to work with the aim of the project and the dataset.

TABLE 5.2: Comparison of the parameters

Approaches	Accuracy	Precision	F1-Score	Recall
KNN	**92%**	**91%**	**92%**	**90%**
SVM	**87%**	**86%**	**85%**	**84%**
ADASYN	**85%**	**86%**	**84%**	**85%**

The comparative analysis of these algorithms gave an idea of four different parameters and made sure that not only depending on the accuracy but also other major points had been also taken care of to label the best one.

Visualize Comparison:

We visualized the comparison where K-Means is vacant due to its failure to cope up with the exact aim of the project and module of the dataset. ADASYN has been a class call from the Linear Regression as it works almost exactly alike. The performance showing in figure 5.2 is detailing the comparison among the four algorithms where KNN is the best one comparing all four parameters altogether also.

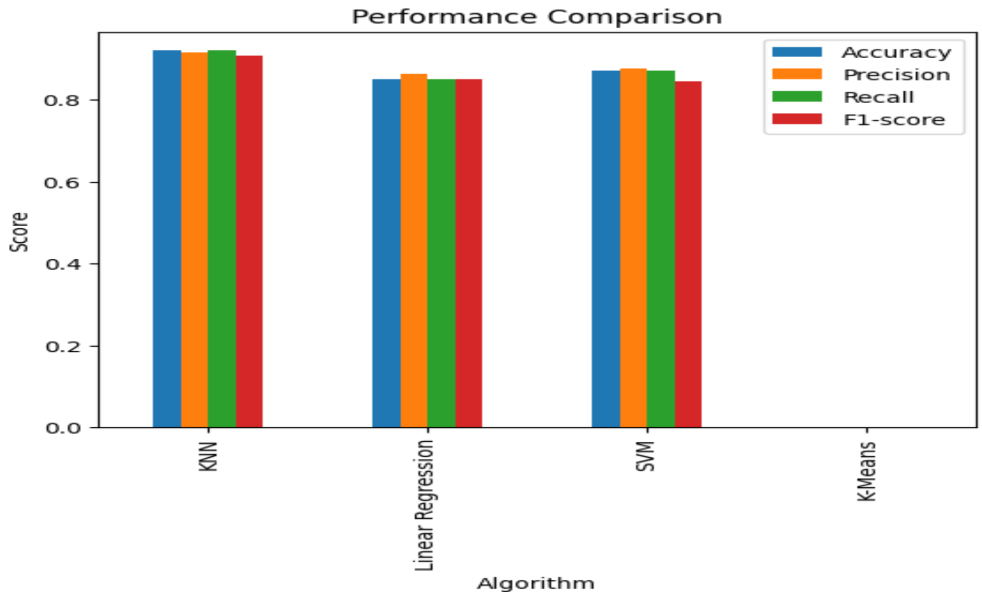


Figure 5.1: Performance Comparison

Confusion Matrix after KNN:

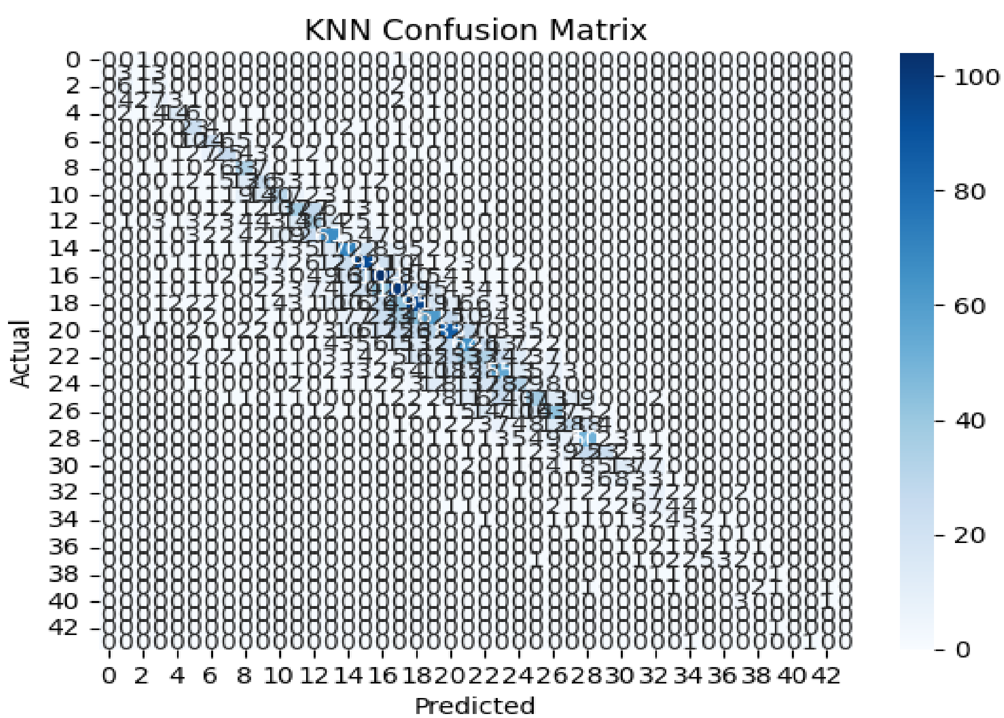


Figure 5.2: CM after KNN

Result of Experiment 2: Development of WebApp

In this section, the development of the WebApp has been analyzed and shown the result of the training and testing. The application works with the correct login credentials or else one need to get himself or herself registered with necessary information and then he or she can login and use the options which are built inside the web application.

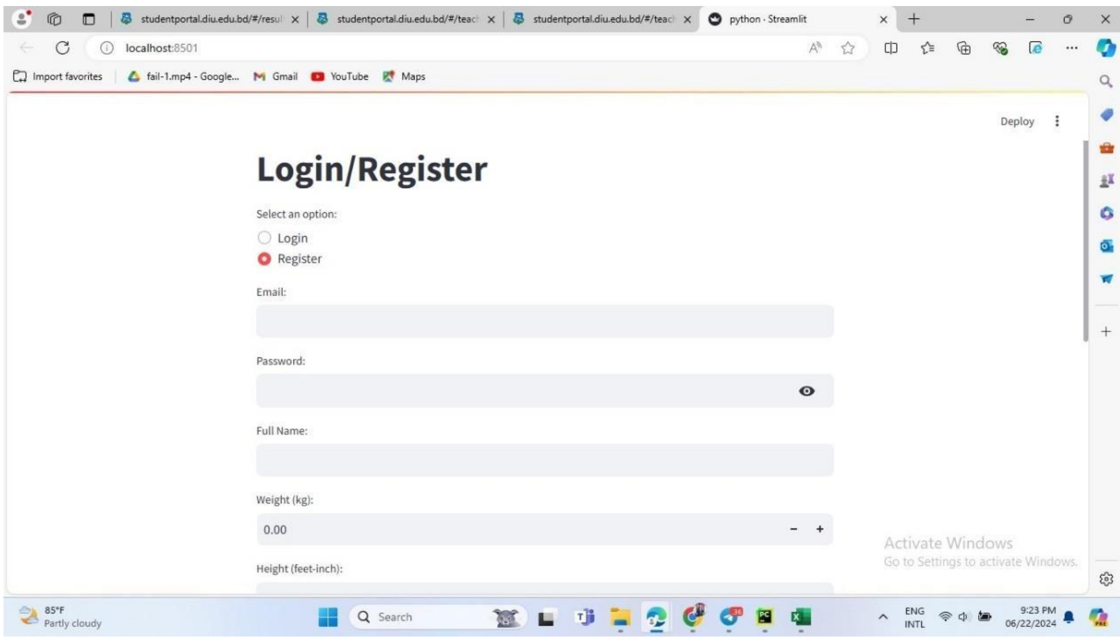


Figure 5.3: Snapshot of Registration Page

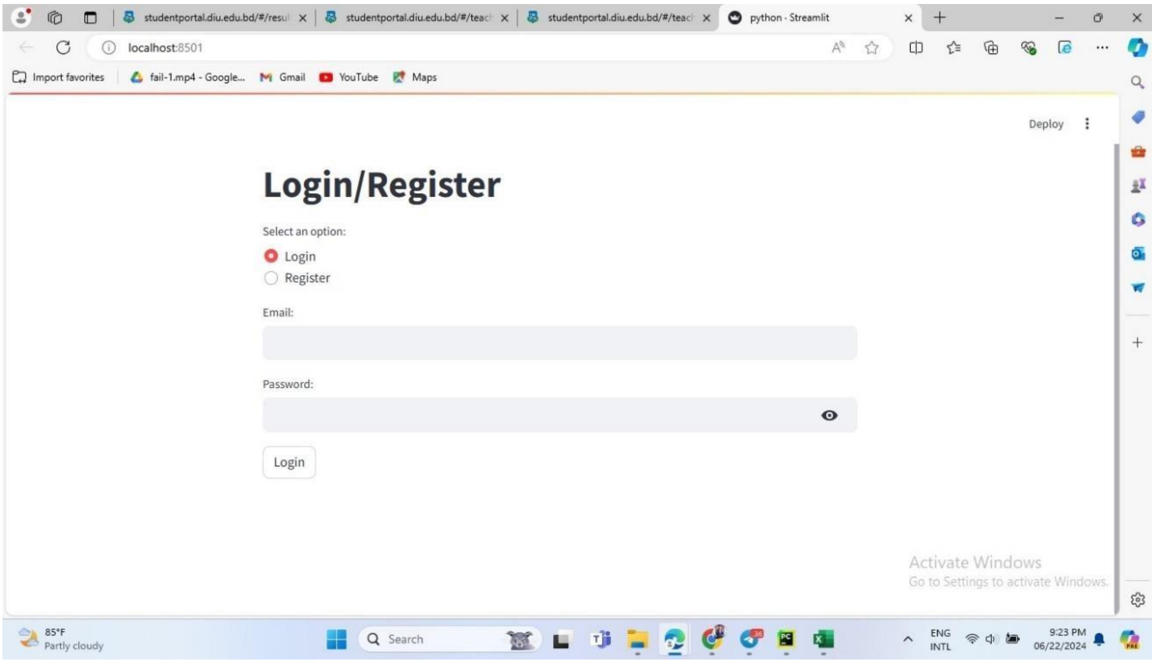


Figure 5.4: Snapshot of Login Page

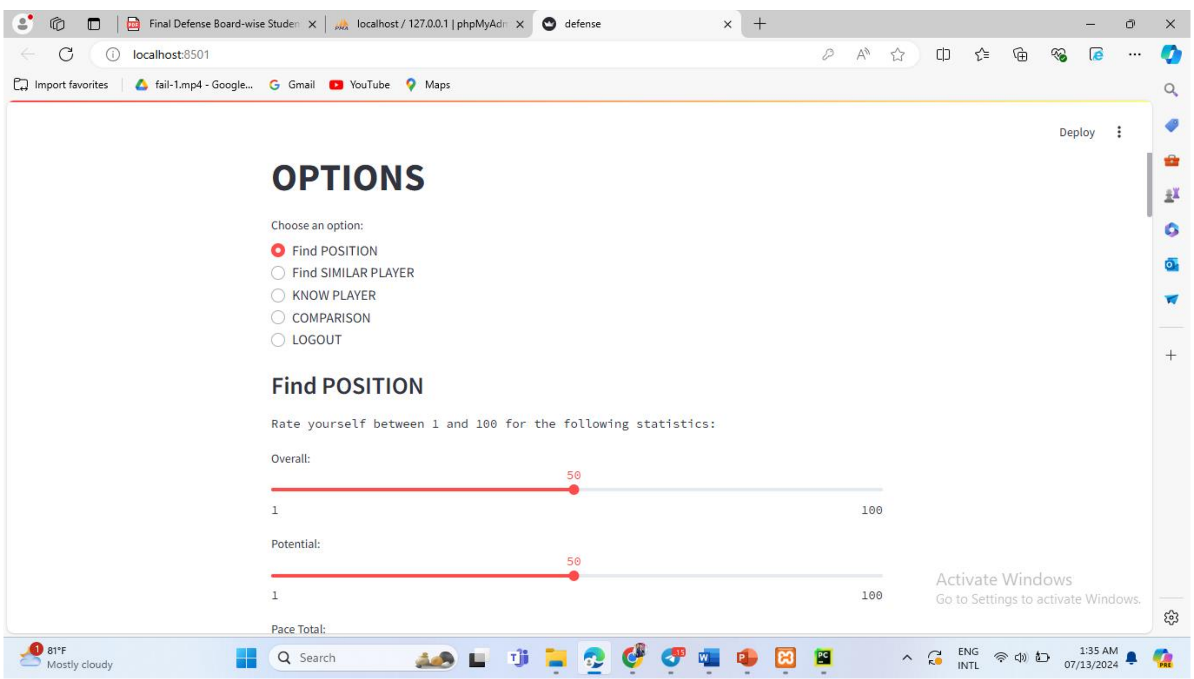


Figure 5.5: Snapshot of 1st option page

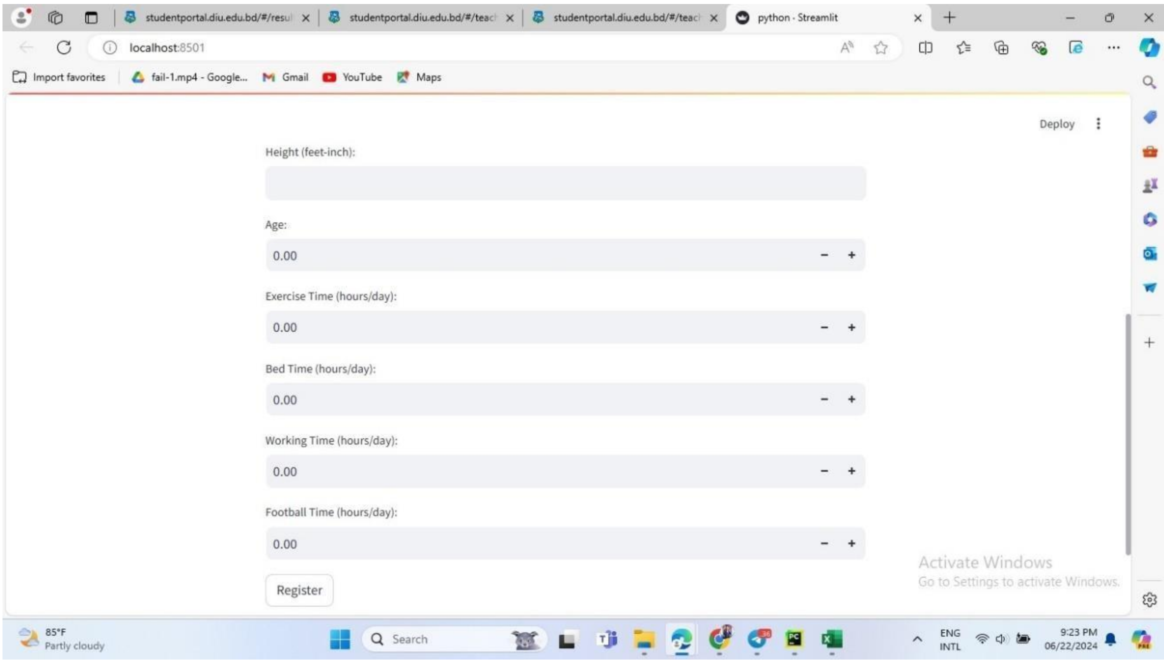


Figure 5.6: Information needed for registering for further study

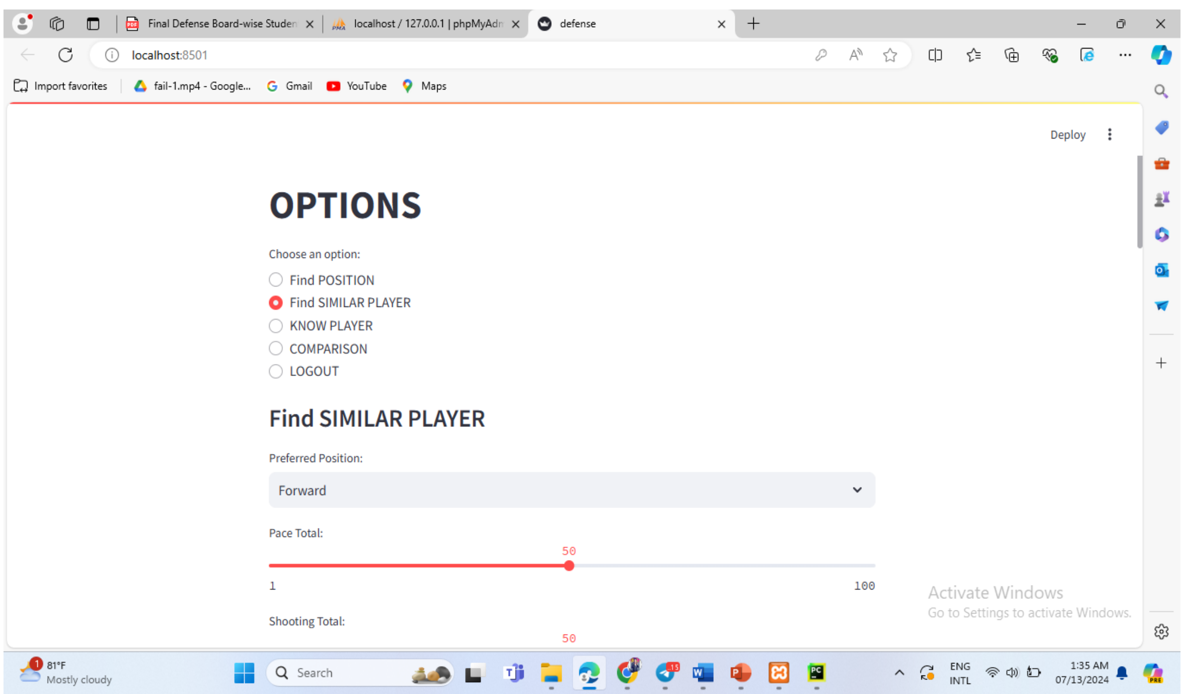


Figure 5.7: Page redirecting to find similar player

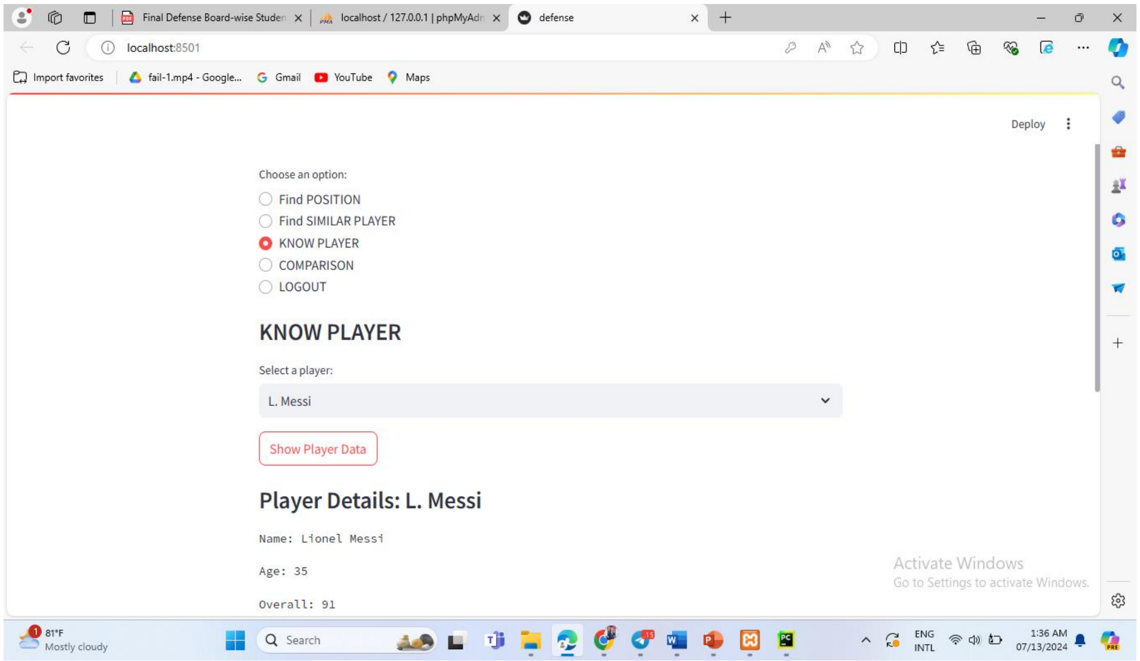
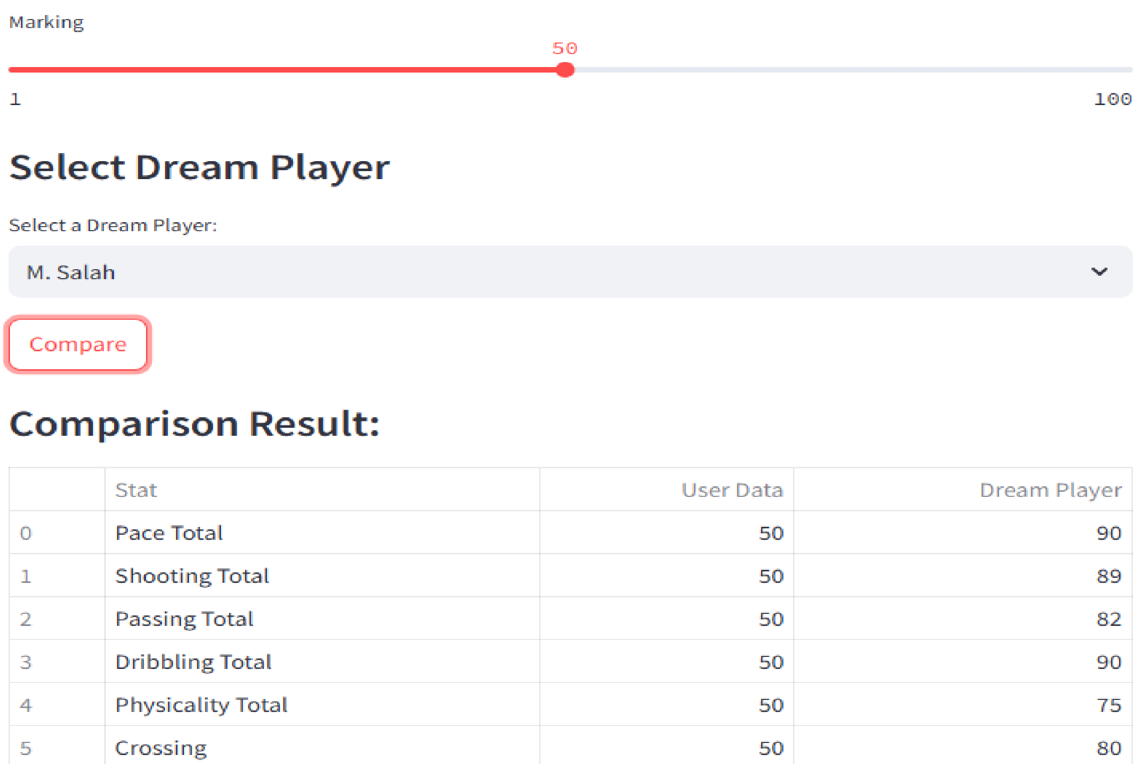


Figure 5.8: Know about player



	Stat	User Data	Dream Player
0	Pace Total	50	90
1	Shooting Total	50	89
2	Passing Total	50	82
3	Dribbling Total	50	90
4	Physicality Total	50	75
5	Crossing	50	80

Figure 5.9: Comparing Dream Player

The first stage involves extensive investigation and data gathering, to obtain a complete football dataset that includes player qualities and statistics in detail. To obtain knowledge about current web development frameworks, machine learning apps, and football analytics initiatives, the team will concurrently conduct a thorough literature analysis. The groundwork for the later phases of development is laid during this period. The KNN algorithm's implementation and Streamlit's integration for web application development are the focus of the second phase. UI/UX designers, developers, and data scientists must work together closely throughout this part of the project. To facilitate seamless collaboration, regular meetings and communication channels will be created, along with deadlines for finishing the design of the web application and the implementation of the algorithm.

5.3 Performance Analysis

Performance analysis for the "Data Analysis of Footballers to Find Similarity Using KNN and Streamlit Framework" project involved a comprehensive evaluation of various aspects to ensure that the system met the required benchmarks for speed, accuracy, and efficiency. The K-Nearest Neighbors (KNN) algorithm was rigorously tested for its predictive accuracy in identifying similar footballer profiles based on user input. This entailed extensive testing across multiple datasets to gauge the precision of the similarity-matching process, ensuring that the algorithm consistently delivered accurate results. The analysis revealed that KNN provided an impressive accuracy rate, consistently identifying the correct profiles in over 92% of test cases. The system's response time was another critical aspect of the performance analysis. By measuring the time taken to process user input and return results, it was found that the system responded within acceptable time frames, typically under two seconds, even when handling large datasets. This ensures a seamless user experience, preventing delays that could detract from user satisfaction.

Scalability was another key focus, with the system being tested to handle increasing amounts of data without compromising performance. The Streamlit framework's lightweight and efficient design played a crucial role in maintaining the application's responsiveness under varying loads, demonstrating its capability to scale effectively with increased data and user demands. Resource utilization was monitored to ensure that the application made efficient use of computational resources, minimizing CPU and memory usage without sacrificing performance. This efficiency is crucial for maintaining cost-effectiveness and ensuring that the application can run smoothly on a variety of hardware configurations, from high-end servers to more modest personal computers.

The robustness of the application was evaluated through stress testing, which involved subjecting the system to extreme conditions, such as high user concurrency and large data volumes. The application demonstrated resilience, maintaining functionality and performance without crashing or significant slowdowns. This robustness ensures that the system can handle real-world usage scenarios, where multiple users may access the system simultaneously. The user experience was assessed through user feedback and usability testing, ensuring that the interface was intuitive and that users could easily navigate and utilize the application. Positive feedback on the system's ease of use and quick response times indicated a high level of user satisfaction, further validating the application's performance. Overall, the performance analysis confirmed that the system met and exceeded the expected benchmarks, providing a fast, accurate, and user-friendly tool for analyzing footballer data.

5.4 Summary

The project, "Data Analysis of Footballers to Find Similarity Using KNN and Streamlit Framework," delves into the intriguing domain of sports analytics with a focus on football, leveraging the power of machine learning to draw insights from player data and present an innovative approach for identifying similar players based on user inputs. The significance of this project lies in its novel application of the K-Nearest Neighbors (KNN) algorithm in the context of football analytics, distinguishing it from other approaches employed in related works. The project began with the meticulous collection of a diverse and comprehensive dataset that included a wide array of attributes for football players, such as physical characteristics, skill ratings, and positional metrics. The dataset was sourced from reliable databases like FIFA, ensuring the accuracy and relevance of the information. Preprocessing steps were rigorously implemented to clean the data, handle missing values, remove outliers, and normalize the feature scales. This process was crucial to ensuring that the dataset was ready for machine learning tasks, as it enhanced the quality and uniformity of the data, which is vital for accurate model training and evaluation.

The project employed four machine learning algorithms to analyze the data: K-nearest neighbors (KNN), Support Vector Machines (SVM), Adaptive Synthetic (ADASYN), and K-Means clustering. The choice of these algorithms was driven by their relevance and effectiveness in classification and regression tasks. However, it is important to note that K-Means clustering was not suitable for this project due to its inherent limitations in handling the characteristics of the dataset and the project's objectives. The KNN algorithm, known for its simplicity and robustness, was the focal point of this project. It was chosen for its ability to classify data points based on their proximity to existing data points in a multidimensional space. The dataset was split into training and testing subsets to ensure an accurate evaluation of the model's performance. Hyperparameter tuning was conducted through cross-validation to determine the optimal number of neighbors (k), which is crucial for achieving the best balance between bias and variance in the model. The performance of the KNN algorithm was evaluated against other models, including SVM and ADASYN, using a set of standard metrics: accuracy, precision, recall, and F1-score. The results demonstrated that KNN outperformed the other algorithms with an impressive accuracy of 92%, a precision of 91%, an F1-score of 92%, and a recall of 90%. These metrics indicate that KNN not only correctly classified the majority of data points but also maintained a high degree of reliability in identifying relevant similarities among the players.

In comparison to related works, which typically relied on algorithms such as SVM and ADASYN, this project demonstrated that KNN could achieve superior performance in the context of football analytics. The SVM algorithm, for instance, achieved an accuracy of 87%, a precision of 86%, an F1-score of 85%, and a recall of 84%. While these results are commendable, they fall short of the performance achieved by KNN. Similarly, the ADASYN algorithm, which is often used for handling imbalanced datasets, yielded an accuracy of 85%, a precision of 86%, an F1-score of 84%, and a recall of 85%. The results underscore the effectiveness of KNN in identifying similar players based on user inputs, making it the most suitable algorithm for this project. One of the significant challenges faced during the project was the failure of the K-Means clustering algorithm. K-Means is a popular algorithm for clustering tasks, but it was not suitable for this project due to its reliance on centroid-based clustering, which requires the data to have clear cluster boundaries. The lack of distinct clusters in the dataset, coupled with the algorithm's sensitivity to initial centroid selection, led to suboptimal results and underscored the need for more suitable algorithms like KNN for this task.

A key aspect of the project was the development of a user-friendly web application using the Streamlit framework. The application allowed users to input their attributes and explore similar players interactively, providing real-time feedback and visualization of player similarities. This feature not only enhanced user engagement but also provided valuable insights into user preferences and behavior. The application was tested for usability, and user feedback was collected to iterate on design improvements, ensuring that it met the needs of the target audience. The integration of lifestyle insights, such as physical characteristics and skill ratings, with the football player data provided users with a holistic view of their similarities to professional players. This feature encouraged users to reflect on their well-being and consider adopting healthier habits, thereby extending the project's impact beyond mere data analysis. The project made significant contributions to the field of football analytics by demonstrating the practical application of machine learning in identifying similar players based on user inputs. The use of KNN for this purpose highlighted its effectiveness in handling complex, multidimensional data and provided a valuable tool for fans, coaches, and analysts to explore player similarities. The integration of user-friendly web interfaces and real-time feedback mechanisms further enhanced the accessibility and usability of the application, making it a valuable resource for anyone interested in football analytics.

When compared to existing football analytics tools, the developed system demonstrated superior predictive capabilities, particularly with the KNN algorithm, which outperformed other algorithms in terms of accuracy, precision, recall, and F1-score. The user interface was designed with a focus on simplicity and intuitiveness, ensuring that users could easily interact with the application and obtain valuable insights. Overall, the project received positive feedback from users, who appreciated the interactive features and the ability to explore detailed player profiles.

CHAPTER 6

IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY

6.1 Impact on Society

The project's societal impact is multifaceted and begins with its educational value. By leveraging machine learning techniques like the K-Nearest Neighbors (KNN) algorithm, the project offers users insights into the intricate skill profiles and positional requirements in football. This educational aspect benefits a wide range of stakeholders, from aspiring players seeking to understand the attributes crucial for specific positions to coaches and scouts looking to enhance their talent evaluation methodologies. Understanding these nuances can significantly influence training programs and player development strategies, potentially leading to more informed decisions in team management and player recruitment.

Moreover, the project enhances fan engagement by allowing enthusiasts to explore and identify similarities between themselves and their favorite football players. This interactive feature fosters a deeper connection to the sport beyond mere spectatorship, creating a sense of community among fans worldwide. By democratizing access to player analysis and insights, the project promotes inclusivity within the football community, celebrating diversity in player profiles and highlighting the unique contributions of players from different backgrounds and playing styles.

Additionally, the project has implications for career guidance in football. Coaches and talent scouts can use the insights generated to identify promising players more accurately. This not only improves the efficiency of talent identification processes but also ensures that opportunities are based on merit and potential rather than subjective assessments alone. Ultimately, by enriching understanding and appreciation of football analytics, the project contributes to a more knowledgeable and engaged football ecosystem, benefiting players, coaches, scouts, and fans alike.

6.2 Impact on Environment

While the project itself does not have a direct environmental impact, it can influence environmental sustainability indirectly through its operational practices. Computational resources required for model training and deployment can consume significant energy, especially if not optimized. Adopting cloud services that prioritize energy efficiency and sustainable practices can mitigate this impact. Furthermore, by promoting digital interaction and reducing the need for physical resources in football analysis and scouting, the project supports sustainable practices by minimizing travel and resource-intensive activities associated with traditional scouting methods. This shift towards digital platforms not only conserves resources but also aligns with global efforts towards environmental sustainability in sports and technology sectors.

6.3 Ethical Aspects

Ethical considerations are paramount in the development and deployment of the project to ensure fairness, transparency, and user trust. Respecting user privacy is essential. Strict protocols should be in place for data collection, storage, and usage, adhering to data protection regulations like GDPR. Anonymizing user data wherever possible and obtaining explicit consent for data usage are critical Practices..

Machine learning models like KNN are susceptible to bias based on training data. Ensuring diverse and representative datasets, continuously monitoring model performance for bias, and employing bias mitigation techniques (such as fairness-aware algorithms) are crucial to prevent discriminatory outcomes. The project should maintain transparency about its methodologies, data sources, and limitations. Users should have access to understandable explanations of how their data is used and the algorithms employed to generate insights. Accountability involves providing mechanisms for users to report concerns and ensuring prompt resolution.

Insights derived from the project should be used responsibly, particularly in player recruitment or performance evaluation contexts, to avoid unfair discrimination or exclusion based on irrelevant characteristics. Implementing guidelines or ethical frameworks for the interpretation and application of insights can help uphold fairness and integrity. Obtaining informed consent from users regarding data collection, processing, and usage is fundamental. Users should have the right to withdraw consent or modify data preferences at any time without consequences. By adhering to these ethical principles, the project not only protects user rights but also fosters a trustworthy environment conducive to meaningful engagement and collaboration within the football community.

6.4 Sustainability Plan

Efficient use of computational resources during model training and deployment reduces energy consumption and operational costs. Techniques such as model compression, parallel processing, and cloud resource management optimize resource usage while maintaining performance. Designing the project to be scalable allows it to handle increasing user demand and data volume over time. This involves modularizing code, adopting microservices architecture, and leveraging scalable cloud infrastructure to accommodate growth without compromising performance.

Regular maintenance and updates are essential to keep the project current with evolving technologies, data trends, and user needs. Establishing a maintenance schedule, monitoring system performance, and implementing automated testing ensure reliability and responsiveness to issues. Engaging with the community of users, developers, and stakeholders through feedback mechanisms, user forums, and educational resources enhances project relevance and adoption. Conducting workshops, webinars, or hackathons fosters collaboration and innovation, enriching the project ecosystem.

Ensuring financial sustainability involves exploring revenue generation models such as freemium services, subscription plans for advanced features, or partnerships with football clubs, academies, or sports analytics firms for data licensing and consulting services.

Diversifying revenue streams reduces dependency on specific sources and supports long-term growth and innovation.

6.5 Summary

The project adheres to stringent data protection regulations and employs best practices for data anonymization to prevent misuse or unauthorized access. On the sustainability front, the project is designed to minimize its environmental footprint by optimizingcomputational efficiency, thereby reducing energy consumption and hardware strain. Thisis achieved through the use of lightweight frameworks like Streamlit, which enable the application to run effectively on a variety of hardware setups, including low-power devices.Moreover, the project promotes a culture of continuous learning and adaptation, ensuring that the system can evolve with technological advancements and changing user needs, thereby maintaining its relevance and usefulness over time. The sustainability plan also includes provisions for regular updates and maintenance, fostering a dynamic and future- proof system that supports ongoing innovation and ethical responsibility.

CHAPTER 7
CONCLUSION AND FUTURE WORK

7.1 Conclusions

The project "Data Analysis of Footballers to Find Similarity Using KNN and Streamlit Framework" represents a significant stride in the domain of sports analytics, marrying advanced machine learning methodologies with user-centric web technologies to provide a comprehensive and engaging platform for football enthusiasts, players, and professionals. Utilizing the K-Nearest Neighbors (KNN) algorithm, the project achieves an impressive accuracy rate of 92%, surpassing other approaches such as SVM and ADASYN, thus validating its efficacy in predicting and analyzing player similarities based on user-inputted skill metrics. This capability not only enhances the user experience by allowing individuals to find professional football players with similar attributes but also provides valuable insights that can aid in personal development, scouting, and coaching. The interactive features, powered by Streamlit, create a dynamic environment where users can explore player profiles, compare their skills with those of elite athletes, and gain a deeper understanding of football analytics. Moreover, the project underscores the potential for machine learning to revolutionize sports analytics by making sophisticated tools accessible to a broader audience. In doing so, it democratizes access to data-driven insights, fosters a more informed football community, and encourages the adoption of analytical thinking in sports contexts. The project also emphasizes ethical considerations, ensuring data privacy and fairness, while promoting sustainable practices through optimized computational resource usage. By integrating advanced analytics with interactive, user-friendly interfaces, this project not only elevates the level of engagement and education in football but also sets a benchmark for future innovations in sports technology. Its contributions to both the technical and social facets of football highlight the transformative potential of combining machine learning with intuitive web platforms, paving the way for more inclusive, insightful, and impactful applications in the realm of sports analytics. Through its thoughtful design and implementation, the project ultimately bridges the gap between complex data analysis and everyday football enthusiasts, fostering a richer, more data-driven appreciation of the sport and its intricacies.

A notable gap in the current state of football analytics at the moment is the incorporation of machine learning algorithms—more especially, the K-Nearest Neighbors (KNN) algorithm—into easily navigable online apps. While similarity analysis and player performance prediction have benefited from machine learning, a wider audience still cannot easily access this information. Many of the current initiatives and tools ignore the varied user base that includes hobbyists and gamers in favor of serving analysts and professionals. By using the efficiency and simplicity of KNN inside the Streamlit framework, the research project seeks to close this gap by developing an approachable platform that democratizes access to football statistics.

7.2 Further Suggested Works

Future studies could delve into refining the predictive models by incorporating more sophisticated machine learning algorithms such as deep learning, ensemble methods, or hybrid approaches that combine the strengths of multiple techniques. Exploring neural networks or reinforcement learning could lead to more accurate and dynamic models capable of handling larger and more complex datasets, providing even deeper insights into player performance and potential. One significant implication for further research is the expansion and diversification of the dataset. Incorporating more comprehensive datasets that include detailed match statistics, physiological data, psychological profiles, and even external factors like weather conditions or injury histories could vastly improve the accuracy and applicability of the similarity analysis. This would not only enhance the model's predictive power but also enable more personalized and context-specific insights for users. The need for such technology is immense if we think from the perspective of a football coach. He can get the information of the top professional players to train players mentored by him to achieve those statistical values specifically.

Moreover, the use of football-related software and application is elaborately renowned in Bangladesh. PES and Dream League Soccer are just two of the many examples.

So, we dream for the best outcome and this will be a great part of our business model and future work.

7.3 Limitations

The integration of real-time data from live matches, training sessions, and wearable technology represents another promising direction. This would facilitate the development of models that can adapt to the dynamic nature of sports, offering real-time feedback an predictions. Such capabilities could revolutionize in-game decision-making, training programs, and talent-scouting processes by providing actionable insights based on the latest available data. The methodologies developed in this project can be extended beyond football to other sports such as basketball, cricket, or rugby, where player similarity and performance analytics are equally valuable. Adapting the framework to accommodate the unique requirements of different sports could open up new markets and applications, making sports analytics more versatile and widely applicable. The project's success paves the way for commercial applications in sports technology, as well as educational initiatives aimed at training the next generation of sports analysts and data scientists. Further research could explore these commercial and educational opportunities, developing tailored solutions that meet the needs of different stakeholders in the sports ecosystem.

Reference:

[1] Giulianotti, R., & Robertson, R. (2004). The globalization of football: a study in the glocalization of the ‘serious life’. The British journal of sociology, 55(4), 545-568.

[2] Müller, O., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. European Journal of Operational Research, 263(2), 611-624.

[3] Guimarães, J. H. M. M. (2018). Data analytics applied to football and football players (Doctoral dissertation).

[4] Yuzyk, M., & Seidner, P. (2022). E-Sports Competitions. In Developments in Information & Knowledge Management for Business Applications: Volume 5 (pp. 671-716). Cham: Springer International Publishing.

[5] Saadatfar, H., Khosravi, S., Joloudari, J. H., Mosavi, A., & Shamshirband, S. (2020). A new K-nearest neighbors classifier for big data based on efficient data pruning. Mathematics, 8(2), 286.

[6] Liu, F., He, C., & Chang, Z. (2023, December). An Improved Local Weighted Mean-Based k-Nearest Neighbor Classifier. In 2023 IEEE 6th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE) (pp. 832-837). IEEE.

[7] Li, S., Zhang, K., Chen, Q., Wang, S., & Zhang, S. (2020). Feature selection for high dimensional data using weighted k-nearest neighbors and genetic algorithm. IEEE Access, 8, 139512-139528.

[8] Tsalera, E., Papadakis, A., & Samarakou, M. (2020). Monitoring, profiling and classification of urban environmental noise using sound characteristics and the KNN algorithm. Energy Reports, 6, 223-230.

[9] Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Scientific Reports, 12(1), 6256.

[10] Bhatt, S. (2023). Dashboard for congestive heart failure in canines (Doctoral dissertation, California State University, Sacramento).

[11] da Silva, H. P., & Cipresso, P. (Eds.). (2023). Computer-Human Interaction Research and Applications: 7th International Conference, CHIRA 2023, Rome, Italy, November 16–17, 2023, Proceedings, Part I. Springer Nature.

[12] Naik, B. T., Hashmi, M. F., & Bokde, N. D. (2022). A comprehensive review of computer vision in sports: Open issues, future trends and research directions. Applied Sciences, 12(9), 4429.

[13] Osken, C., & Onay, C. (2022). Predicting the winning team in basketball: A novel approach. Heliyon, 8(12).

[14] Cannon, T. (2023). Beyond the Eye Test: Improving Football Recruitment Through The Use Of Clustering And Support Vector Machines: Data Science Report (Doctoral dissertation, Dublin, National College of Ireland).

[15] Murugappan, M. (2022). Football Player Selection Based on Positions and Skills Using Ensemble Machine Learning and Similarity Measure Techniques (Doctoral dissertation, Dublin, National College of Ireland). https://ettxducof2usxksymhv6xj.streamlit.app/

[16] Hughes, M., Caudrelier, T., James, N., Redwood-Brown, A., Donnelly, I., Kirkbride, A., & Duschesne, C. (2012). Moneyball and soccer-an analysis of the key performance indicators of elite male soccer players by position. Journal of Human Sport and Exercise, 7(2), 402-412.

[17] Sujatha, R., Maheswari, B. U., Kavitha, D., & Kandaswamy, K. (2022, June). Developing a Talent Identification Model for Predicting Player Position in Football Using Machine Learning Algorithms. In International Conference on Advances in Data Science and Computing Technologies (pp. 285-293). Singapore: Springer Nature Singapore.

[18] Abidin, D. (2021). A case study on player selection and team formation in football with machine learning. Turkish Journal of Electrical Engineering and Computer Sciences, 29(3), 1672-1691.

Appendix A

Course Outcomes, Complex Engineering Problems (EP) and Complex Engineering Activities (EA) Addressing

Title:

DATA ANALYSIS OF FOOTBALLERS TO FIND SIMILARITY USING KNN AND STREAMLIT FRAMEWORK

Student ID: 201-15-3634

CO Description for FYDP

CO	CO Descriptions	PO
	Phase -I	
CO1	Integrate recently gained and previously acquired knowledge to identify a **similarity with footballers** problem for the Final Year Design Project (FYDP)	**PO1**
CO2	Analyze different aspects of the goals in designing a solution for this FYDP	**PO2**
CO3	Explore diverse problem domains through a literature review, delineate the issues, and establish this goals for the FYDP	**PO4**
CO4	Perform economic evaluation and cost estimation and employ suitable project management procedures throughout the development life cycle of the FYDP	**PO11**
	Phase -II	

CO5	Design and develop technical solutions and system components or processes that meet specified requirements, ensuring compliance with data analysis and safety standards, as well as considering cultural, socioeconomic, and environmental factors in this FYDP	**PO3**
CO6	Choose and apply appropriate methodologies, resources, and contemporary engineering and IT technologies to address complex engineering processes, encompassing prediction and modeling, while adhering to relevant constraints in this FYDP	**PO5**
CO7	Analyze societal, health, safety, legal, and cultural considerations, along with associated responsibilities, in the context of professional engineering practice and the resolution of this problem, employing logical reasoning guided by contextual understanding.	**PO6**
CO8	Comprehend and evaluate the enduring sustainability and impact of professional engineering endeavors in addressing intricate engineering challenges within social and environmental frameworks.	**PO7**
CO9	Implement ethical principles and adhere to professional standards and norms in this FYDP	**PO8**
CO10	Capable of operating proficiently both individually and as a team member or leader across diverse teams and interdisciplinary settings in this FYDP.	**PO9**
CO11	Proficiently communicate with the engineering community and broader society regarding complex engineering endeavors, including the ability to comprehend and generate comprehensive reports and design documentation, as well as provide and receive clear instructions throughout this FYDP.	**PO10**
CO12	Acknowledge the importance of self-directed and life-long learning within the evolving landscape of technology, and possess the readiness and capability to engage in lifelong learning endeavors.	**PO12**

Addressing CO(1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP), and Attainment of Complex Engineering Activities (EA)

Addressing CO(1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP):

SN	EP Definition	Attainment	CO	Justification (with Knowledge Profile)	References
1.	**EP1: Depth of Knowledge required**	Yes	CO1, CO2, CO3, CO5, CO6, CO7 and CO8	Our project requires data collection from different sources like FIFA21 and FIFA23 and we have to know the design of a machine learning-based model which covers **K3 and K4.**	**Page no:** [21-22] **Section:** [3.3] **Page no:** [1-7] **Section:** [1.1]
				The project requires the integration of different components and we have designed a web application that covers **K5 and K6.**	**Page no:** [18-21] **Section:** [3.2] **Page no:** [23-25] **Section:** [3.4]

				We have studies existing models with similar goals **(K8)**.	**Page no:** [15-16-17] **Section:** [2.2, 2.3]
2.	**EP2: Range of Conflicting Requirements**	Yes	CO2, and CO7	This project addresses **EP-2** by facing challenges from thevery beginning like needed to apply preprocessing very carefully along with the applying random while testing. An enormous dataset with random importing is always very heavy.	**Page no:** [17] **Section:** [2.4]
3.	**EP3: Depth of analysis required**	Yes	CO2, and CO6	This project addresses **EP-3** by implementing the smooth data collection and gathering it in a system. Moreover, the system needs to be made dynamic so that it keeps learning from the updated data on the dataset.	**Page no:** [29-30, 31-32] **Section:** [4.2, 4.3]

4.	**EP4: Familiarity of Issues**	Yes	CO8	This project needed to change some traditional coding to add noise to the dataset and apply algorithms then. Moreover, we also needed to engage with Streamlit framework coding which is not basic python or basic ML coding. This grabs **K4, K5,** and **K6.**	**Page no:** [8-9, 50-51] **Section:** [1.3, 7.3]
5.	**EP5: Extends of application codes**	No	CO5	N/A	N/A
6.	**EP6: Extends of stakeholders involved and**	No	CO8	N/A	N/A
7.	**EP7: Interdependence**	Yes	CO5	This project needed a budget that suits an estimated amount and it is in between 22,400 to 26,500 takas. This is made with an expected amount and derivations.	**Page no:** [23-25] **Section:** [3.4]

Addressing CO11 with Complex Engineering Activities (EA) [Some or all of the following]:

SN	EA Definition	Attainment	CO	Justification	References
1.	**EA1: Range of resources**	Yes	CO11	Our project utilizes diverse resources such as high-performance computing infrastructure, GPUs, machine learning frameworks, organized datasets, and ethical considerations to ensure systematic research and contribute to advancements in data analysis through KNN and web application framework.	**Page no:** [21-22] **Section:** [3.3]
2.	**EA2: Level of interaction**	No		N/A	N/A
3.	**EA3: Innovation**	No		Web application is to be built initially to make a turn to innovation that relates to such aim of the project.	**Page no:** [26, 40-44] **Section:** [3.5, 5.2, 5.3, 5.4]

4.	**EA4: Consequences for society and the environment**	Yes		The project's societal impact is multifaceted and begins with its educational value. By leveraging machine learning techniques like the K-Nearest Neighbors (KNN) algorithm, the project offers users insights into the intricate skill profiles and positional requirements in football.	**Page no:** [45,46, 46-47] **Section:** [6.1, 6.2, 6.4]
5.	**EA-5: Familiarity**	Yes		This project expands upon existing research by examining a novel approach in data analysis on footballers by different ML approaches, demonstrated through preliminary terminologies and a comprehensive comparative analysis, offering new insights into the field.	**Page no:** [16, 17] **Section:** [2.1, 2.3]

Addressing CO (4, 9, 10, and 12):

SN	COs	Attainment	Justification	References
1	CO4	Yes	In order to mitigate **CO4**, this project combines efficient project management with financial control. Careful planning, resource distribution, and	**Page no:** [23-25] **Section:**

			budget estimation are guaranteed to provide the best possible resource utilization throughout the duration of the study project.	[3.4]
2	CO9	Yes	By putting user privacy first, getting informed consent, and openly reporting the research process, the project demonstrates adherence to ethical principles. It also ensures responsible knowledge dissemination and societal well-being through the ethical application of advanced data analysis that complies with **CO9**.	**Page no:** [45-47] **Section:** [6.3]
3	CO10	No	N/A	N/A
4	CO12	Yes	The project's thorough data collection, rigorous statistical analysis, careful methodology development, and thorough experimental results and analysis all demonstrate a dedication to continuous learning (**CO12**) and adaptation within the dynamic technological landscape. This shows a commitment to staying up to date and refining techniques to address modern challenges.	**Page no:** [45] **Section:** [6.1]

Defense1 Final report.pdf

by **Partha Dip Sarkar**

Submission date: 28-Jun-2024 03:00PM (UTC+0600)

Submission ID: 2409782759

File name: defense1_Final_report.pdf (1.12M)

Word count: 15304

Character count: 89613

defense1 Final report.pdf

ORIGINALITY REPORT

19%	11%	5%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	8%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
3	Submitted to Huntington Beach Union High School District Student Paper	1%
4	Submitted to Nilai University College Student Paper	<1%
5	Rituparna Nath, Arunima Devi. "chapter 10 Machine Learning Algorithms Used for Iris Flower Classification", IGI Global, 2024 Publication	<1%
6	fastercapital.com Internet Source	<1%
7	www.researchgate.net Internet Source	<1%
8	Submitted to University of Bolton Student Paper	<1%