APPLIED RESEARCH PROJECT COVER SHEET

NAME: KARAN KOUNDINYA JANAKIRAM

STUDENT NUMBER: 10602768

COURSE TITLE: MASTER OF SCIENCE IN DATA ANALYTICS

PROJECT NAME: PLAYER POTENTIAL PREDICTION SYSTEM USING MACHINE LEARNING TECHNIQUES

DISSERTATION COORDINATOR NAME: DR. ANDREW BROWNE

SUPERVISOR NAME: MR PIERPAOLO DONDIO

MODULE/SUBJECT TITLE: B9DA113 APPLIED RESEARCH PROJECT (B9DA113_2223_TMD3)

BY SUBMITTING THIS PROJECT, I CONFIRM THAT I AM AWARE OF DBS'S POLICY REGARDING CHEATING, PLAGIARISM AND ALL OTHER FORMS OF ACADEMIC IMPROPRIETY. THE COURSEWORK SUBMITTED IS MY OWN WORK, AND ALL OTHER SOURCES CONSULTED HAVE BEEN APPROPRIATELY ACKNOWLEDGED.I AM AWARE THAT IN THE CASE OF DOUBT AN INVESTIGATION WILL BE HELD

NAME: KARAN KOUNDINYA JANAKIRAM **DATE:** 29/08/2023

ABSTRACT

Aim of the Thesis: This thesis aims to harness the power of data and machine learning to predict the potential of football players using the FIFA dataset. Among five predictive models, the best-performing model is selected, and data analysis is conducted to uncover essential performance indicators and data relationships.

The Five data models selected are Linear Regression, K Nearest Neighbours, Support Vector Regression, Gradient Boosting Regression and Random Forest regression

Among the five predictive models examined, the Random Forest Regression model emerged as the best performer with an impressive R-squared value of 0.967, indicating an excellent fit to the data. It demonstrates a remarkable ability to capture complex relationships between player attributes and future potential. Its low Mean Absolute Error (MAE) of 0.68 and Mean Squared Error (MSE) of 1.23 underscore its precision in predicting player ratings

The questions which are used to create the KPI's and establish the relationship between the data are

Question 1: Player BMI and Physical Attributes

Question 2: Continental Comparison of Physical Attributes

Question 3: Relationship between International Reputation and Player Attributes

Question 4: Relationship between Age and Player Attributes

Question 5: Future Value Prediction and Club/Country Analysis

The model, equipped with newly created metrics, showcases meaningful patterns and data relationships, providing valuable insights into the factors driving player potential. The synthesis of data analysis and KPIs facilitates data-driven decision-making for player development and team management in professional football, offering quantifiable evidence of performance and the potential for future success.

ACKNOWLEDGEMENT

I would like to extend my heartfelt gratitude to Dublin Business School for providing me with the invaluable opportunity to pursue my Master's in Data Analytics. This journey has been a transformative experience, and I am deeply appreciative of the support, resources, and conducive academic environment that DBS has offered throughout my educational endeavour.

I would like to express my sincere appreciation to Dr. Andrew Browne, my Dissertation Coordinator, for his invaluable guidance, insightful feedback, and unwavering support during the research process. Dr. Browne's expertise and dedication were instrumental in shaping the direction of my thesis, and I am truly grateful for his mentorship.

Furthermore, I extend my profound thanks to Mr. Pierpaolo Dondio, my Thesis Supervisor, for his exceptional guidance and mentorship. Mr. Dondio's expertise in the field of machine learning and data analytics was pivotal in the successful completion of my project, "Player Potential Prediction System Using Machine Learning Techniques." His insightful feedback, encouragement, and dedication to my academic growth have been instrumental in this achievement.

I would also like to express my appreciation to the faculty and staff at Dublin Business School for their support, encouragement, and commitment to fostering a stimulating academic environment.

Lastly, I am thankful to my family and friends for their unwavering encouragement and understanding during this demanding academic journey. Your support has been my driving force.

This thesis would not have been possible without the contributions and support of these individuals and institutions. Their dedication to my academic and professional growth has been invaluable, and I am truly grateful for their guidance and encouragement.

Thank you.

Karan Koundinya Janakiram

TABLE OF CONTENTS

GENERAL INFO REGARDING FILES FOLDER

- The raw dataset, sourced from Kaggle, serves as the foundational data for this study.

- Within this repository, you will find a PowerPoint presentation file that encapsulates the key findings and insights derived from our research.

- A comprehensive project report is available here, providing a detailed account of the research methodology, results, and conclusions.

- The coding segment is separated into two distinct files:

1. **"MODEL_IMPLEMENTATION_10602768.ipynb"** comprises the implementation of the project model.

2. Furthermore, we have the file titled **"METRICS&KPI_DATA_ANALYSIS_10602768.ipynb,"** which is dedicated to housing the questions used to construct our Key Performance Indicators (KPIs) and establish the crucial data relationships central to this research effort.

INTRODUCTION

In the rapidly changing landscape of Football, a combination of enthusiasm and data-based knowledge has initiated a transformation. This transformation is revolutionizing the way in which clubs and organizations conduct their operations, particularly in terms of player and team development. At the core of this transformation is the need to accurately forecast the potential of football players. In the midst of the excitement and enthusiasm of the sport, data analytics has become a key factor in gaining a competitive advantage.

This thesis is embarked on a noble endeavour, motivated by a single goal: to uncover the complex network of potential players in football. This is a journey illuminated by the power of data and the wisdom of machine intelligence. Five predictive models are at the forefront of this endeavour, each attempting to uncover the mysteries of the dataset. Like scouts searching for the hidden gems, these models venture into the depths of data to predict the future of aspiring stars

RESEARCH QUESTION

To what extent can machine learning models reliably anticipate the potential of football players by leveraging their attributes within the FIFA dataset? Furthermore, how do the five different prediction models perform, and what variations exist among them within this context

Among five predictive models, select the best-performing model on the basis of MSE, MAE and R Squared Values

The insights that can be gleaned from key performance indicators (KPIs) and data relationships is on the basis of these questions

Question 1: How does the Body Mass Index (BMI) of football players correlate with their physical attributes, and how does it vary across different countries? This investigation entails calculating and visualizing the average BMI of players for each country, identifying the ten countries with both the highest and lowest average BMI scores. Furthermore, it seeks to pinpoint the player with the lowest and the player with the highest BMI in the dataset. The analysis aims to reveal any discernible patterns or relationships between player BMI and their physical characteristics within the realm of football

Question 2: How do physical attributes such as physicality, stamina, strength, and acceleration vary across different continents in football, and can we determine which continent exhibits superiority in these aspects? This analysis seeks to quantify and compare the performance metrics of players from various continents, shedding light on the disparities in physical attributes and identifying potential strengths among continents in the realm of football.

Question 3: What is the connection between a player's international reputation and key attributes such as wage, market value, player count, and age? This investigation aims to delve into the intricate relationship between a player's standing on the international stage and these fundamental attributes, seeking to analyze and quantify the interplay between reputation and player characteristics within the realm of football

Question 4: What is the correlation between the age of football players and crucial attributes such as wage, potential, market value, and sprint speed? This inquiry aims to investigate and quantify the associations between player age and these key characteristics, facilitating a comprehensive analysis of how age influences player attributes within the context of football**.**

QUESTION 5: This research endeavours to construct a forward-looking metric, denoted as 'Value Next Year,' designed to predict a player's expected market value in the subsequent year based on historical data and performance metrics. Subsequently, the analysis will encompass the visualization of the top 10 players projected to have the highest values in the upcoming year. Additionally, it will scrutinize the dataset to identify countries and clubs with the highest representation among players. These analyses aim to offer insights into player valuation trends and global football dynamics, thereby contributing to a more comprehensive understanding of market dynamics and the prevalence of players from specific regions and clubs within the FIFA dataset

LITERATURE REVIEW

The paper titled "Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques" by Mustafa A. Al-Asadi and Sakir Tasdemir(2022) is a research study that investigates the use of FIFA video game data and machine learning techniques to predict the value of football players.

The paper begins by discussing the importance of player valuation in the football industry, and the challenges associated with traditional methods of player valuation. The authors then introduce the concept of using video game data, specifically data from the FIFA video game series, as a source of information for player valuation.

The paper then goes on to explain the methodology used in the study, which involves collecting data on the in-game attributes of football players in FIFA, such as speed, dribbling, and shooting accuracy, as well as their market values in the real-world transfer market.

Machine learning takes centre stage in the research's analytical journey. The authors apply a suite of techniques, including Linear Regression, Decision Tree, Random Forest, Support Vector Regression (SVR), and Neural Networks. These algorithms are utilized to construct predictive models capable of estimating player values based on the amalgamation of in-game attributes

The implications of the research are both profound and multifaceted. The fusion of gaming data and machine learning not only presents an innovative approach to player valuation but also unveils a broader intersection between virtual experiences and real-world sports dynamics. The predictive models, if refined and expanded, could be a game-changer in sports analytics, aiding club decisions, scouting processes, and talent identification. Moreover, the study resonates with the contemporary emphasis on data-driven decision-making, offering a pioneering application in the world of football

Overall, this paper provides a novel approach to player valuation in the football industry, and highlights the potential of using video game data and machine learning techniques to improve the accuracy of player valuation models. The authors demonstrate the feasibility of their approach, and suggest that it could be used in combination with traditional methods of player valuation to provide a more comprehensive and accurate picture of player value.

The article "Predicting Market Value of FIFA Soccer Players with Regression" on Towards Data Science is a technical write-up that discusses the application of machine learning (ML) models to predict the market value of football players in the video game FIFA. The author, Alvin uses a dataset containing player attributes and market value data from the game to develop regression models that can predict a player's market value based on several factors such as age, overall rating, and specific attributes like pace, dribbling, and shooting (Alvin, T.P. 2022)

The methodology employed in the article comprises data preprocessing, exploratory data analysis (EDA), and regression modelling. Data preprocessing involves cleansing and organizing the dataset to ensure its quality and compatibility for analysis. EDA, a crucial step, offers insights into the data's distribution, trends, and potential correlations between attributes and market values. This process sets the stage for selecting relevant features and understanding their significance in predicting player valuations.

The article provides a detailed explanation of how to pre-process and analyse the dataset, including techniques for handling missing data, encoding categorical variables, and feature scaling. The author then compares several regression models, including linear regression, decision tree regression, and random forest regression, and evaluates their performance using various metrics such as the mean squared error (MSE) and the coefficient of determination (R-squared).

Overall, this article is useful for my literature review as it provides a practical example of how to use regression models to predict player ranking in a video game.

The article "Using a Multivariable Linear Regression Model to Predict the Sprint Speed of Players in FIFA 19" on Hackernoon is a technical write-up that discusses the application of machine learning (ML) models to predict player ranking in the video game FIFA 19. The author uses a dataset containing player attributes and performance data from the game to develop a multivariable linear regression model that can predict a player's sprint speed based on several factors such as age, strength, agility, and stamina. (Sibanda, E. 2019)

The article as the article above guides us on how to prepare the data, clean it, and analyse it using various statistical techniques. The author also explains how to build, train, and test the ML model and how to evaluate its performance using different metrics such as (MAE) and (R-squared) as in the above-mentioned article

The pivotal element of the analysis is the application of a multivariable linear regression model. This model aims to establish a mathematical relationship between player attributes and sprint speed. By identifying the attribute coefficients that significantly influence sprint speed, the model provides insights into the relative importance of different in-game characteristics in shaping players' virtual sprint performance.

In conclusion, E. Sibanda's study "Using a Multivariable Linear Regression Model to Predict the Sprint Speed of Players in FIFA 19" provides a glimpse into the burgeoning realm of sports analytics within virtual gaming environments. By harnessing multivariable linear regression, the study showcases the potential of data-driven insights to predict virtual player performance. It underscores the convergence of technology and sports by illustrating how virtual attributes can be leveraged to make predictive inferences about gameplay outcomes. While acknowledging its limitations, the study contributes to the broader discourse on the intersection of gaming, data science, and sports analytics, paving the way for future explorations in this evolving landscape.

RESEARCH DESIGN METHODOLOGY.1

I have opted for CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology for this research for multiple reasons

Structured Approach to Data Preparation: Our research method, beginning with data collection and preparation, is aligned with CRISP-DM.

Modelling and Evaluation: To achieve our aim of training several machine learning models and choosing the best-performing one, CRISP-DM contains defined steps for modelling and evaluation.

Data analysis and insight production: It is really important for our research's secondary goals, and CRISP-DM enables their implementation

Practical Application: Our research intends to give enterprises a prediction model for player potential, making it useful in the fields of football and sports analytics. The systematic approach of CRISP-DM supports our objective of minimizing the impact of biases and subjective opinions by ensuring that the study is practical and applicable.

The different stages in **CRISP DM** methodology in our research are

Business Understanding: In the ever-evolving realm of football, data-driven insights have become instrumental in gaining a competitive edge. This research, driven by five predictive models, aims to predict players' future potential, offering clubs a streamlined approach to player rating updates and the ability to leverage data analytics for improved decision-making and player development.

Data Understanding: Data sources include the FIFA 23 Players Dataset from Kaggle as the primary source and secondary sources such as academic articles, books, and online resources related to football and video game rating systems. The FIFA 23 Players Dataset contains player attributes and performance statistics.

Data Preparation: Data preprocessing includes cleaning, normalization, and feature selection to prepare the dataset for regression modelling, The dataset will be divided into training and testing sets for model development and evaluation.

Modelling: Multiple machine learning algorithms, including Linear Regression, Random Forest Regression, K Nearest Neighbors Regression, Support Vector Regression, and Gradient Boosting Regression, will be used for modelling. One model will be selected as the best-performing model.

Evaluation: Model performance will be evaluated using standard metrics such as mean squared error, root mean squared error, and R-squared.

Data Analysis: Five questions have been formulated to analyse the dataset, leveraging data headers to create new metrics and understand patterns and performance. Insights will be generated to indicate key performance indicators and correlations between data points

Feasibility: This research is feasible to answer within the timeframe and practical constraints. The dataset is readily available, and the research methodology is well-established

Specificity: The research aims to identify the most significant factors that affect player ratings in the FIFA dataset. It will use multiple regression analysis to determine the relationship between player attributes and ratings

Complexity: The research is complex enough to develop the answer over the space of a paper or thesis, The well-established nature of the research methodology enhances the potential for effective communication and collaboration with peers, mentors, and advisors

Relevance: This research is relevant to the field of study and society more broadly. The findings will provide insights into how player rating prediction systems can be built. The research will also provide insights into the factors that affect player ratings in football, which can be used by football clubs and coaches to identify and recruit players

DESIGN AND METHODOLOGY:

Definition of Machine Learning

By means of application of statistical methods, machine learning can be defined as a branch of artificial intelligence that enables computers to learn and make decisions without being explicitly programmed. It is based on the notion that computers can make judgments, identify patterns in data, and learn from it all without a great deal of assistance from humans.

It falls within the category of artificial intelligence. By providing robots the capacity to learn and create their own programming, it aims to make them more human-like in their actions and decision-making.

The computers are supplied high-quality data, and various methods are employed to create ML models to train the machines on this data. The type of data available and the sort of action that has to be automated determine the algorithm to be used. (Advani, V. 2020)

Types of Machine Learning

Supervised Learning

A class of issues known as supervised learning use a model to learn the mapping between the input and target variables. The term "supervised learning tasks" refers to applications that include training data detailing the various input variables and the goal variable.

Let (x) represent the set of input variables and (y) represent the goal variable. An algorithm for supervised learning attempts to train a fictitious function that is a mapping denoted by the formula y=f(x), which is a function of x.

Here, learning is being watched over or guided. The algorithm is adjusted every time it makes a forecast to improve the outcomes because we already know the outcome. The input and output variables are included in training data, which is used to fit models before they are applied to test data to produce predictions. During the test phase, only the inputs are supplied, and the model's outputs are compared to the goal variables held back to determine how well it performed.

Unsupervised Learning

In unsupervised learning problems, the model learns independently and finds patterns and extracts the relationships between the data. Just like in supervised learning, there is no teacher or supervisor to guide the model. In unsupervised learning, the model only works on the input variables, there are no target variables, and the goal is to understand the underlying patterns of the data. Unsupervised learning can be divided into two main categories:

Cluster – finding out the groups in the data

Density Estimation – consolidating the distribution of data

These operations are done to understand the patterns of the data Visualization – creating plots and graphs, Projection – reducing the dimensionality of the data. (Advani, V. 2020)

Reinforcement Learning

What is reinforcement learning?

A reinforcement problem is a type of problem in which there is an agent operating in an environment. The agent is operating in the environment according to the feedback or reward provided to it by the environment. The feedback can be positive or negative. The agent proceeds in the environment on the basis of the rewards it has received. The reinforcement agent determines the actions to be performed. There is no specific training data and the machine learns independently. A classic reinforcement problem is playing a game. Google's AlphaGo is a great example of reinforcement learning. It beat the world's number one Go player. (Advani, V. 2020)

For our thesis, we have opted for Supervised Learning as we have to model to learn the mapping between the input and target variables, the two types in supervised learning are

Classification, which involves the prediction of a class label, and Regression, which requires the prediction of a numerical value, are the two main categories of supervised issues.

For example, let's say we have a dataset containing student information from a specific university. We can use a regression algorithm to determine the height of each student based on the student's weight, sex, diet, subject major, etc. Because height is a constant quantity, there are infinite possible values for it.

In contrast, classification is used to determine whether or not an email is spam or not. The algorithm looks at the keywords in the email and then checks the sender's address to see how likely the email is to be spam.

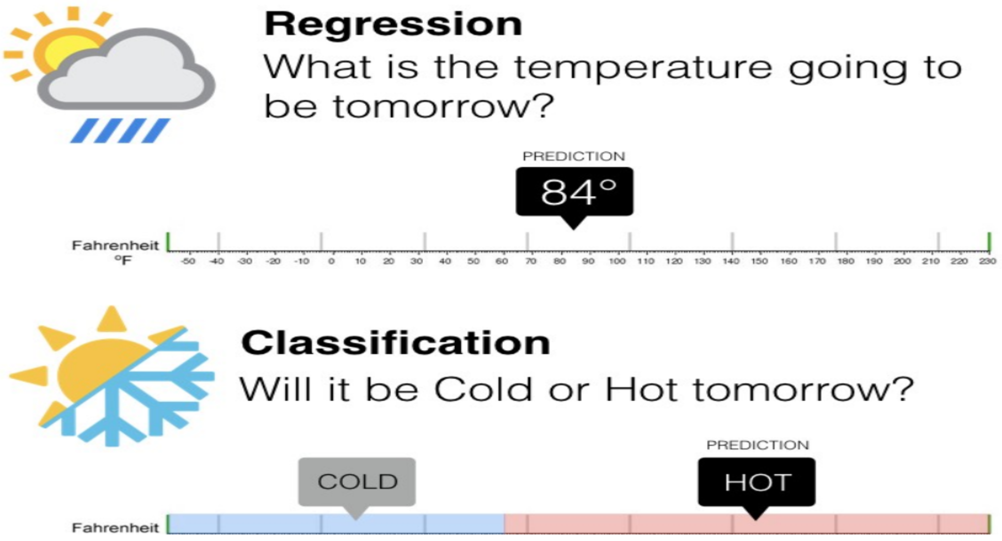


Fig 1.1

Just as a regression model is used to predict the next day's temperature, a classification algorithm is used to figure out whether or not it will be cool or hot based on the temperature values given. (Gupta, S. 2021)

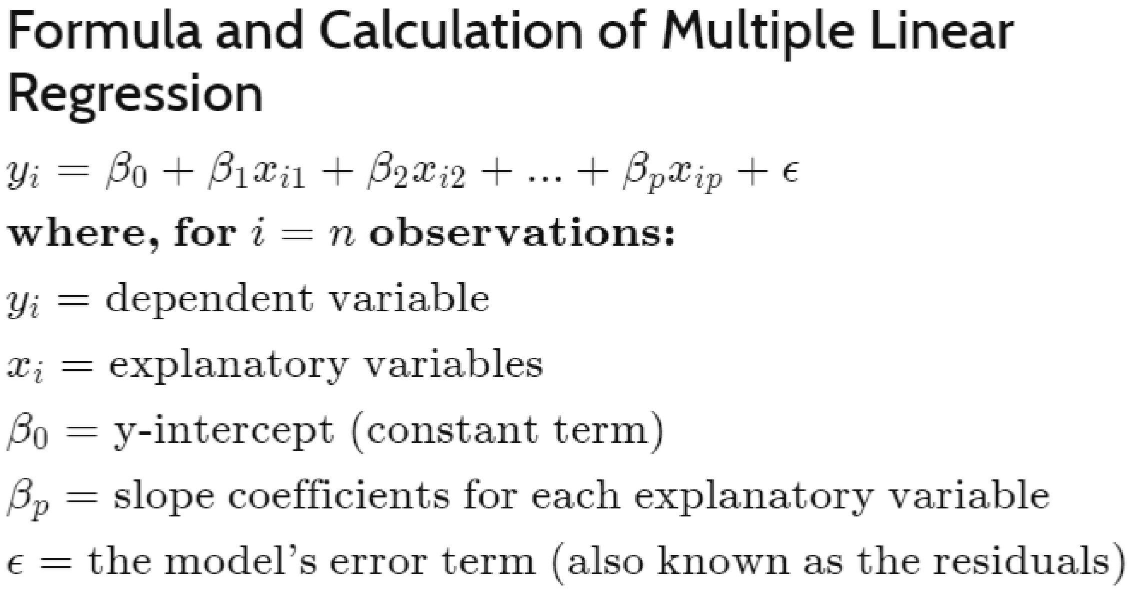
For our thesis and for the question in hand, we have to predict the player's predicted potential, where the algorithm learns to predict continuous values based on input features. The output labels in regression are continuous values

Types of Algorithms in Regression used in our research

Multiple Linear Regression

Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable, is a statistical method that forecasts the result of a response variable using a number of explanatory variables. Modelling the linear relationship between the explanatory (independent) factors and response (dependent) variables is the aim of multiple linear regression.

In a multiple regression, the impact of several explanatory variables on a particular outcome is taken into consideration. When all other model variables are kept constant, it assesses the relative impact of these independent or explanatory variables on the dependent variable. (Hayes, A. 2023)



Formula and Calculation of Multiple Linear Regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

x_i = explanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

Fig 1.2

K-Nearest Neighbours Algorithm

The k-nearest neighbours' algorithm, sometimes referred to as KNN or k-NN, is a supervised learning classifier that employs proximity to produce classifications or predictions about the grouping of a single data point. It can be applied to situations involving regression or classification. Similar to classification problems, regression issues use the notion, however in this case, the average of the k nearest neighbours is used to forecast a classification ('What is the k-nearest neighbors algorithm?', no date).

It is also known as a lazy learner algorithm since it saves the training dataset rather than learning from it immediately. Instead, it uses the dataset to perform an action when classifying data. The KNN method simply saves the information during the training phase, and when it receives new data, it categorizes it into a category that is quite similar to the new data

Flexibility: KNN can handle multi-dimensional data and doesn't require extensive preprocessing (K-Nearest Neighbor(KNN) Algorithm for Machine Learning – Javatpoint, 2021)

Support Vector Regression

Support vector regression (SVR) is a supervised machine learning algorithm that is used to predict continuous values. It is a type of regression analysis that finds the best fit hyperplane to a set of data points. The hyperplane is a line or curve that divides the data into two classes, with the target variable on one side and the predicted variable on the other. SVR is different from traditional linear regression in that it does not assume that the data is linearly separable. This means that the data points may not all lie on a straight line. Instead, SVR uses a kernel function to map the data into a higher dimensional space where it can be separated by a hyperplane. The goal of SVR is to find the hyperplane that has the maximum margin between the two classes of data points.

SVR is a non-parametric algorithm, which means that it does not make any assumptions about the distribution of the data. This makes it a robust algorithm that can be used with a variety of data types, SVR is a computationally efficient algorithm, which makes it suitable for large datasets, SVR is a versatile algorithm that can be used for both classification and regression tasks (Singh, A. 2022)

Random Forest Regression

A supervised learning approach called random forest regression use group learning techniques to build a more reliable and accurate model than a single decision tree. It operates by building a large number of decision trees during training, each of which is trained on a distinct bootstrap sample of the data. The final prediction is then created by averaging the projections from each individual tree. Powerful algorithms like random forest regression can be utilized for a range of regression applications like predicting housing prices, predicting sales, and streamlining manufacturing procedures. It excels in situations where the data is noisy or where there are missing values.

Gradient Boosting Regression

Gradient boosting regression is a supervised learning technique that employs ensemble methods to construct a more precise and resilient model compared to a single decision tree. It operates by progressively incorporating decision trees into the model, with each tree being trained to rectify the errors made by the preceding trees. The objective of gradient boosting regression is to minimize the loss function, which quantifies the disparity between the predicted values and the actual values.

The following algorithms were not selected for the regression task because they are not well-suited for this type of problem:

Naive Bayes is typically used for classification tasks, not regression tasks, Decision trees can be overfitting on regression tasks, but this issue is mitigated in ensemble methods like Random Forest and Gradient Boosting and Neural networks require a lot of data, which is not available in this case.

The following algorithms were selected for the regression task because they are well-suited for this type of problem:

Linear regression is a simple algorithm that is easy to understand and interpret, K-nearest neighbours is a non-parametric algorithm that can capture local patterns in the data, Support vector regression is a robust algorithm that is not easily overfitting, Random Forest and gradient boosting are ensemble methods that combine the strengths of multiple decision trees. These algorithms were selected to provide a comprehensive range of capabilities, including simplicity, robustness, ensemble capabilities, and high predictive accuracy. It is important to validate these algorithms using the specific dataset and explore different hyperparameters to fine-tune their performance

RESEARCH DESIGN METHODOLOGY.2

This section discusses about the methodologies that have been followed to carry out the research process. Planning, setting a roadmap and timeline is very essential in working toward the desired purpose of a research. The research has focussed on building a machine learning model for predicting player potential, the research also contains metrics and Key performance indicators which indicate the multiple implications to the world of footballing analytics. To implement the following actions on the dataset, many essential steps have been documented down below with its definition followed by its part in the research which would lead to the results expected.

Problem Definition and Data Collection:

Clearly define the problem you aim to solve with machine learning. Collect relevant data that will help address the problem effectively. Defining the problem involves understanding the task's scope, objectives, and potential impact on business or research. Data collection involves identifying data sources, acquiring necessary permissions, and ensuring data quality

Understanding every feature present in the dataset helps in deciding their relevance for the research. Features maybe independently chosen a part of research or can be an integral part of the dataset. Understanding the data helps to ensure its closeness to the business requirement and is essential for the planned research

Data has been gathered as per this research plan and each feature is explained as follows:

Known As: This is a common nickname or commonly known name of the player.

Full Name: The full legal name of the player.

Overall: This represents the overall rating or skill level of the player in the game.

Potential: The potential rating that the player can reach in the game.

Value (in Euro): The in-game value of the player in terms of virtual currency (Euros).

Positions Played: The different positions on the field that the player can play.

Best Position: The position at which the player performs best.

Nationality: The player's nationality.

Image Link: A link to an image of the player.

Age: The age of the player.

Height (in cm): The player's height in centimetres.

Weight (in kg): The player's weight in kilograms.

TotalStats: The total statistical value or sum of all the player's attributes.

BaseStats: The base attributes or fundamental skills of the player.

Club Name: The name of the club the player is associated with.

Wage (in Euro): The in-game wage of the player in virtual currency (Euros).

Release Clause: The amount of virtual currency (Euros) needed to release the player from their contract.

Club Position: The player's position in their current club.

Contract Until: The year until which the player's contract with the current club is valid.

Club Jersey Number: The jersey number worn by the player in their current club.

Joined On: The year the player joined the current club.

On Loan: Indicates if the player is currently on loan to another club.

Preferred Foot: The foot (left or right) that the player prefers to use.

Weak Foot Rating: A rating (1 to 5) indicating the player's proficiency with their weaker foot.

Skill Moves: The number of skill moves the player can perform (1 to 5).

International Reputation: The player's reputation on the international stage (1 to 5).

National Team Name: The name of the national team the player represents.

National Team Image Link: A link to an image of the player representing their national team.

National Team Position: The player's position in the national team.

National Team Jersey Number: The jersey number worn by the player in the national team.

Attacking Work Rate: The player's work rate when it comes to attacking (e.g., High, Medium, Low).

Defensive Work Rate: The player's work rate when it comes to defending (e.g., High, Medium, Low).

Pace Total: The player's overall pace attribute.

Shooting Total: The player's overall shooting attribute.

Passing Total: The player's overall passing attribute.

Dribbling Total: The player's overall dribbling attribute.

Defending Total: The player's overall defending attribute.

Physicality Total: The player's overall physicality attribute.

Crossing: The ability to deliver accurate and effective crosses from wide positions into the box during offensive plays.

Finishing: The skill to take accurate shots on goal, often in close proximity to the goalposts.

Heading Accuracy: The precision and effectiveness of a player's headers when attacking or defending aerial balls.

Short Passing: The accuracy and control of a player's passes over short distances.

Volleys: The proficiency in taking shots on goal from balls that are not on the ground, such as aerial crosses.

Dribbling: The capability to control the ball while maneuvering past defenders using close ball control skills.

Curve: The ability to put spin on the ball, resulting in curved trajectories during shots or passes.

Freekick Accuracy: The accuracy and precision of shots taken during free kicks, which require careful ball placement.

Long Passing: The accuracy and effectiveness of long-distance passes, often used for switching play or initiating attacks.

Ball Control: The skill to receive and maintain control of the ball upon receiving it, regardless of the ball's speed or direction

Acceleration: The speed at which a player can quickly reach their top speed.

Sprint Speed: The maximum speed a player can achieve during a sprint.

Agility: The ability to change direction quickly and with precision.

Reactions: How quickly a player responds to a situation, such as receiving a pass or reacting to an opponent's move.

Balance: The player's ability to stay stable and maintain control of the ball while changing direction or fending off opponents.

Shot Power: The strength and speed of a player's shots on goal.

Jumping: The height a player can reach when jumping, particularly useful for heading the ball.

Stamina: How well a player can maintain their energy and performance level throughout a match.

Strength: The physical power a player has to hold off opponents or win challenges.

Long Shots: The ability to take powerful and accurate shots from a distance.

Aggression: The player's tendency to engage in challenges and compete aggressively for the ball.

Interceptions: The capability to read the game and intercept passes or actions from opponents.

Positioning: How well a player positions themselves on the field to take advantage of tactical opportunities.

Vision: The ability to assess the field and make accurate passes or decisions based on the game situation.

Penalties: The proficiency in taking penalty shots during critical moments of a match.

Composure: The player's ability to maintain calm and make sound decisions under pressure

Marking: How well a player can stick to and track an opponent, often used in man-to-man defensive situations.

Standing Tackle: The skill of tackling an opponent while on their feet, without sliding.

Sliding Tackle: The skill of making a tackle by sliding on the ground to dispossess an opponent.

Goalkeeper Diving: The ability to make diving saves to reach the ball.

Goalkeeper Handling: How well a goalkeeper can control the ball after a save or during handling.

Goalkeeper Kicking: The accuracy and distance of a goalkeeper's kicks from their hands or the ground.

Goalkeeper Positioning: The ability to position oneself effectively in the goal to block shots.

Goalkeeper Reflexes: How quickly a goalkeeper can react to a shot or a situation, contributing to save success

ST Rating (Striker): This rating reflects a player's effectiveness in the central attacking role, often responsible for scoring goals and creating goal-scoring opportunities.

LW Rating (Left Winger): This rating represents how well a player performs as a left-sided forward, contributing to attacks from the left wing and delivering crosses.

LF Rating (Left Forward): This rating indicates a player's performance in a more advanced left-sided attacking role, focusing on goal-scoring opportunities.

CF Rating (Center Forward): This rating assesses a player's ability to play as the main forward, responsible for both scoring goals and assisting teammates.

RF Rating (Right Forward): This rating measures a player's effectiveness as a right-sided forward, contributing to attacks from the right wing and creating chances.

RW Rating (Right Winger): This rating evaluates how well a player performs as a right-sided forward, delivering crosses and participating in attacks from the right wing.

CAM Rating (Central Attacking Midfielder): This rating reflects a player's performance in a central, playmaking role, responsible for creating scoring opportunities and linking play between midfield and attack.

LM Rating (Left Midfielder): This rating assesses a player's ability to contribute from the left midfield position, supporting attacks and providing defensive cover.

CM Rating (Central Midfielder): This rating reflects a player's effectiveness as a central midfielder, involved in both offensive and defensive aspects of the game.

RM Rating (Right Midfielder): This rating evaluates a player's performance from the right midfield position, contributing to both attacking and defensive phases of play.

LWB Rating (Left Wing Back): This rating measures a player's ability to play as a left wing-back, combining defensive duties with overlapping runs and crossing from deep positions.

CDM Rating (Central Defensive Midfielder): This rating reflects a player's proficiency as a central defensive midfielder, focusing on breaking up opponent attacks and distributing the ball.

RWB Rating (Right Wing Back): This rating assesses a player's effectiveness as a right wing-back, combining defensive responsibilities with supporting attacks from wide areas.

LB Rating (Left Back): This rating evaluates a player's performance as a left full-back, contributing to both defensive efforts and providing width in attacks.

CB Rating (Center Back): This rating reflects a player's abilities as a central defender, focusing on defensive positioning, tackling, and aerial duels.

RB Rating (Right Back): This rating measures a player's effectiveness as a right full-back, contributing to both defensive stability and attacking support.

GK Rating (Goalkeeper): This rating is specific to goalkeepers and represents their overall performance in guarding the goal, making saves, and organizing the defensive line.

Data Preparation:

Clean and preprocess the collected data. Handle missing values, outliers, and inconsistencies. Convert the data into a format suitable for analysis.

Data preparation is a crucial step to ensure the quality and reliability of the model's training data. This step might involve normalization, transformation, and encoding categorical variables.

Feature Engineering:

Select and create relevant features from the data that will aid the machine learning algorithm's learning process.

Additional Info: Feature engineering involves extracting meaningful insights from raw data. It could include techniques like creating new features, dimensionality reduction, or selecting the most informative attributes.

Model Selection:

Choose an appropriate machine learning algorithm based on the problem type, available data, and desired outcomes.

The models selected are Linear Regression, K Nearest Neighbours, Support Vector Regression, Random Forest Regression and Gradient Boosting Regression

Model Training and Tuning:

Train the chosen model on the prepared data. Fine-tune its parameters to achieve optimal performance.

Additional Info: Model training involves adjusting the internal parameters of the algorithm using the training data. Hyperparameter tuning, a subset of this step, focuses on optimizing settings that aren't learned by the model itself.

Model Evaluation: Evaluate the model's performance using validation data. Select appropriate evaluation metrics to measure how well the model is performing.

Additional Info: Model evaluation helps understand how well the trained model is likely to generalize to new, unseen data. The metrics used in the research are

MAE (Mean Absolute Error): MAE measures the average absolute difference between predicted and actual values. It gives an overall sense of prediction accuracy, without considering the direction of errors.

MSE (Mean Squared Error): MSE calculates the average of squared differences between predicted and actual values. It emphasizes larger errors and penalizes outliers more than MAE.

R-squared (Coefficient of Determination): R-squared quantifies how well the regression model fits the data. It represents the proportion of the variance in the dependent variable explained by the independent variables. Higher values indicate a better fit.

Model Deployment and Monitoring:

Deploy the trained model to make predictions on new, unseen data. Monitor the model's performance in real-world applications and retrain it as needed.

Additional Info: Deployment requires integrating the model into the operational environment. Monitoring ensures that the model's performance remains accurate and reliable over time, and updates are made when necessary ('Frameworks for Approaching the Machine Learning Process – Kdnuggets', 2018)

The programming language used to conduct the research is Python Programming language

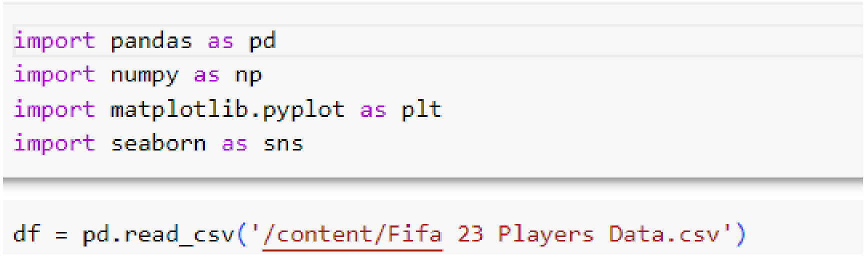
The models and the metrics have been deployed on Google Collab

IMPLEMENTATION & ANALYSIS

Leveraging Metrics and Data Analysis for Data-Driven KPIs

For our research, along with the building of the prediction model, there is also a metric and KPI section which deals with analysing the data in hand, leveraging the data headers multiple new metrics have been created to indicated the patterns and performance of the data, using these metrics and data relationships insights Is generated to indicate the key performance indicators and the correlation between the data points

The primary step is to import the required libraries into Google colab and then load the dataset into the given dataframe



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df = pd.read_csv('/content/Fifa 23 Players Data.csv')
```

Fig 1.3

Matplotlib, Seaborn, and NumPy are powerful Python libraries for data visualization (Matplotlib and Seaborn) and numerical computations (NumPy), enhancing data analysis and presentation

There are five questions, each dealing with a specific aspect of the dataset

Each question has info regarding the question under each sub question, this has been represented in the Colab file, so hence for the report only the screengrab and basic info regarding the question is mentioned

Question 1:

How does the Body Mass Index (BMI) of football players correlate with their physical attributes, and how does it vary across different countries? This investigation entails calculating and visualizing the average BMI of players for each country, identifying the ten countries with both the highest and lowest average BMI scores. Furthermore, it seeks to pinpoint the player with the lowest and the player with the highest BMI in the dataset. The analysis aims to reveal any discernible patterns or relationships between player BMI and their physical characteristics within the realm of football

BMI (Body Mass Index):

"Body Mass Index (BMI) is a numerical value representing a person's body fat based on their weight and height. Formula: BMI = weight (kg) / (height (m)/100) ^2

The main financial benefits of BMI are

Performance optimization:

Maintaining the right BMI can improve a player's physical performance, leading to better results on the court. Improved performance can attract more fans, increase ticket sales, merchandise revenue, and potentially secure higher league rankings, all of which contribute to increased revenue.

Sponsorship and endorsement:

Successful players with good physical fitness can become brand ambassadors, attracting lucrative sponsorship deals and player and club endorsements. A team of athletes with marketing capabilities can secure valuable partnerships, increasing revenue streams.

Player transfers:

Clubs whose players maintain an optimal BMI are more likely to attract interest from other clubs for player transfers. This can lead to negotiating favourable transfer terms and fees, generating additional revenue.

Long-term investment:

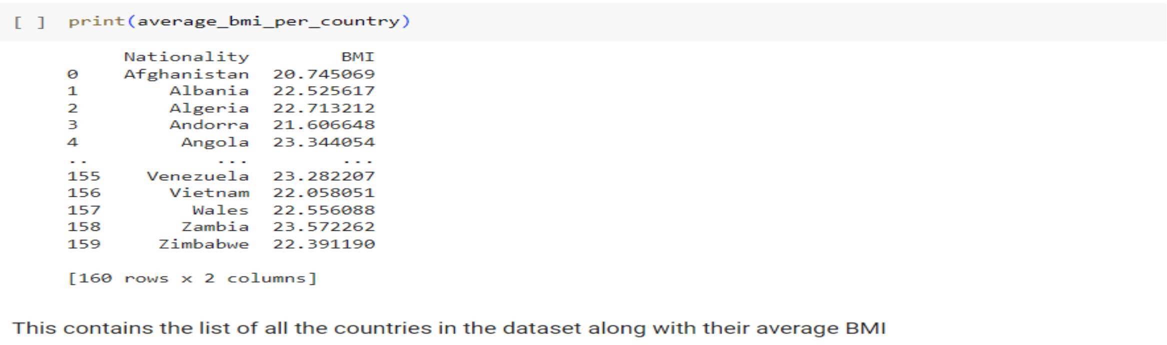
Developing a culture of fitness and health through an optimal BMI contributes to the club's long-term success. This sustained success can attract loyal fans, sponsors and investors, providing stable financial support. Image and Reputation:

Insurance and Contracts:

Clubs can negotiate better premiums and contract terms as players maintain a healthier BMI. This can reduce the financial risk associated with injury and health problems.

The sub questions are as follows with the screengrab from colab file

a) **Calculate the average BMI of each country**

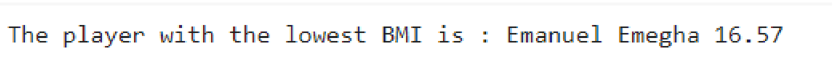


```
[ ]  print(average_bmi_per_country)

         Nationality        BMI
    0    Afghanistan  20.745069
    1        Albania  22.525617
    2        Algeria  22.713212
    3        Andorra  21.606648
    4         Angola  23.344054
    ..           ...        ...
    155    Venezuela  23.282207
    156      Vietnam  22.058051
    157        Wales  22.556088
    158       Zambia  23.572262
    159     Zimbabwe  22.391190

    [160 rows x 2 columns]
```

This contains the list of all the countries in the dataset along with their average BMI

Fig 1.4

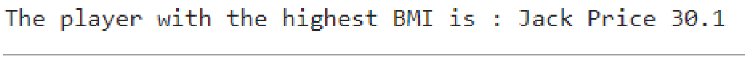
b) **Player with the lowest BMI**



```
The player with the lowest BMI is : Emanuel Emegha 16.57
```

Fig 1.5

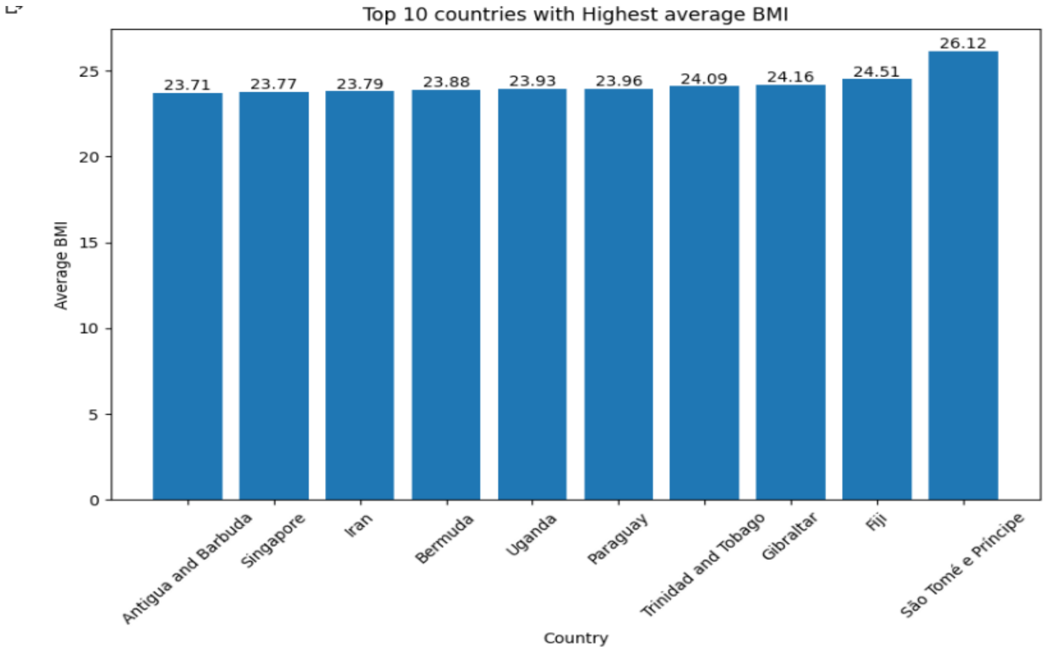
c) **Player with the highest BMI**



```
The player with the highest BMI is : Jack Price 30.1
```

Fig 1.6

d) **Calculate and visualize 10 countries with highest average BMI**



Advantages of selecting players from these countries:

Physical Strength: Players from high average BMI countries might possess greater natural strength

Physical Presence: These players can create a formidable presence on the field, influencing opponents' tactics and creating space for teammates

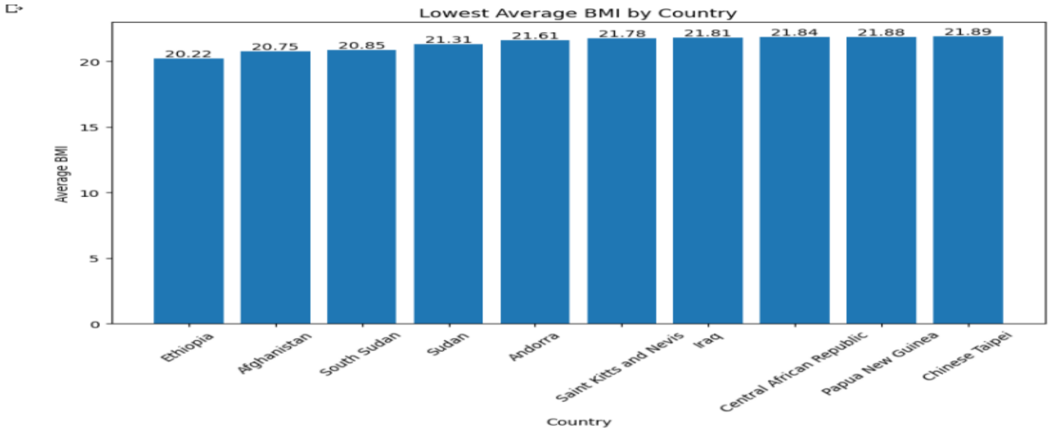
Disadvantages of selecting players from these countries

Fitness Requirements : Maintaining fitness levels and weight management could be more challenging for players with naturally higher BMI

Fig 1.7

c) **Calculate and visualize 10 countries with lowest average BMI**

Fig 1.8



Advantages of selecting players from these countries:

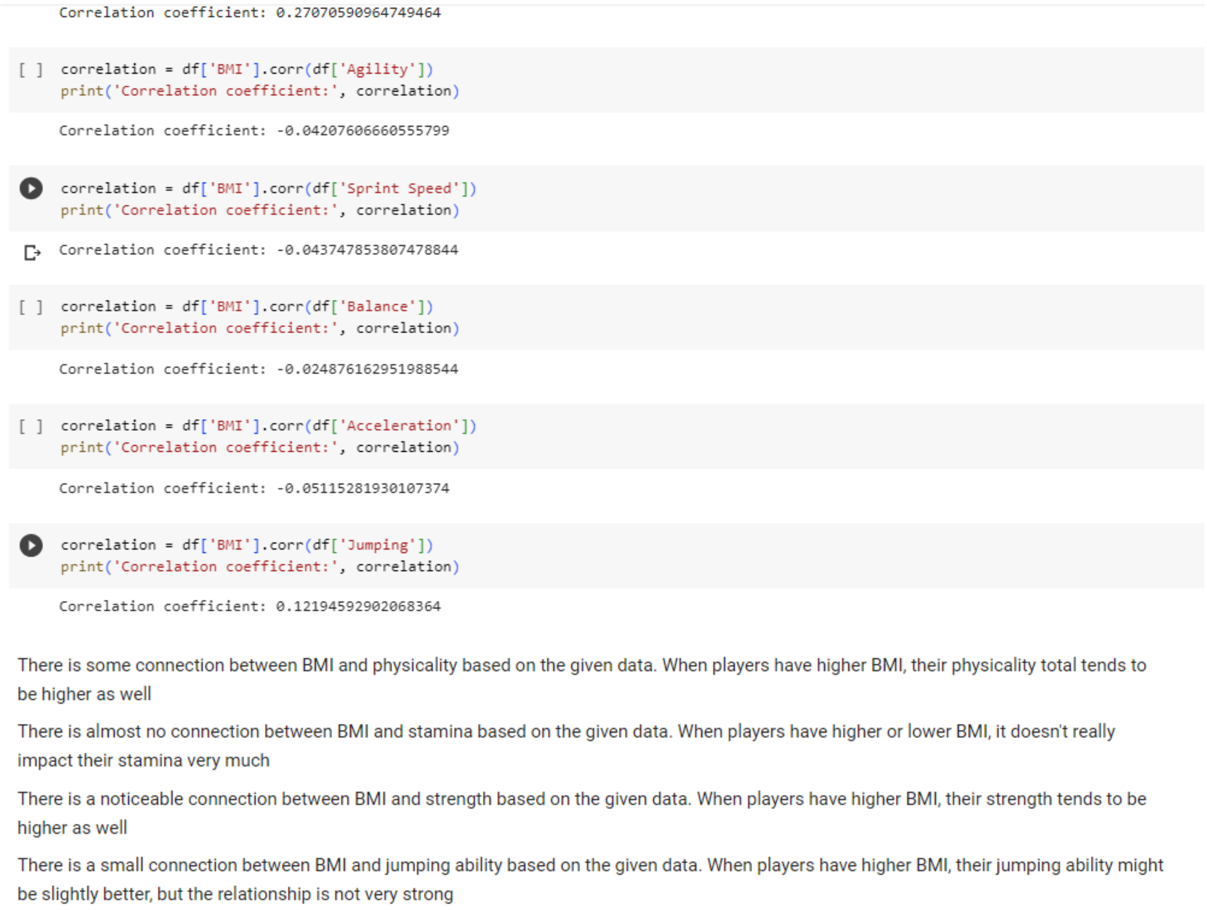
Technical Skills: Lower body mass can contribute to better ball control, dribbling, and intricate footwork

Pressing Game: Lighter players may be better suited for high-pressing tactics, quickly closing down opponents

Disadvantages of selecting players from these countries:

Physical Presence: Players might lack physical strength and presence, affecting their performance in challenges and physical contests

f) **Correlation between BMI and features with respect to the physical attributes of a player and their relationship between them**



```
Correlation coefficient: 0.27070590964749464
```

```
[ ] correlation = df['BMI'].corr(df['Agility'])
    print('Correlation coefficient:', correlation)
```

```
Correlation coefficient: -0.04207606660555799
```

```
correlation = df['BMI'].corr(df['Sprint Speed'])
print('Correlation coefficient:', correlation)
```

```
Correlation coefficient: -0.043747853807478844
```

```
[ ] correlation = df['BMI'].corr(df['Balance'])
    print('Correlation coefficient:', correlation)
```

```
Correlation coefficient: -0.024876162951988544
```

```
[ ] correlation = df['BMI'].corr(df['Acceleration'])
    print('Correlation coefficient:', correlation)
```

```
Correlation coefficient: -0.05115281930107374
```

```
correlation = df['BMI'].corr(df['Jumping'])
print('Correlation coefficient:', correlation)
```

```
Correlation coefficient: 0.12194592902068364
```

There is some connection between BMI and physicality based on the given data. When players have higher BMI, their physicality total tends to be higher as well

There is almost no connection between BMI and stamina based on the given data. When players have higher or lower BMI, it doesn't really impact their stamina very much

There is a noticeable connection between BMI and strength based on the given data. When players have higher BMI, their strength tends to be higher as well

There is a small connection between BMI and jumping ability based on the given data. When players have higher BMI, their jumping ability might be slightly better, but the relationship is not very strong

Fig 1.9

Question 2:

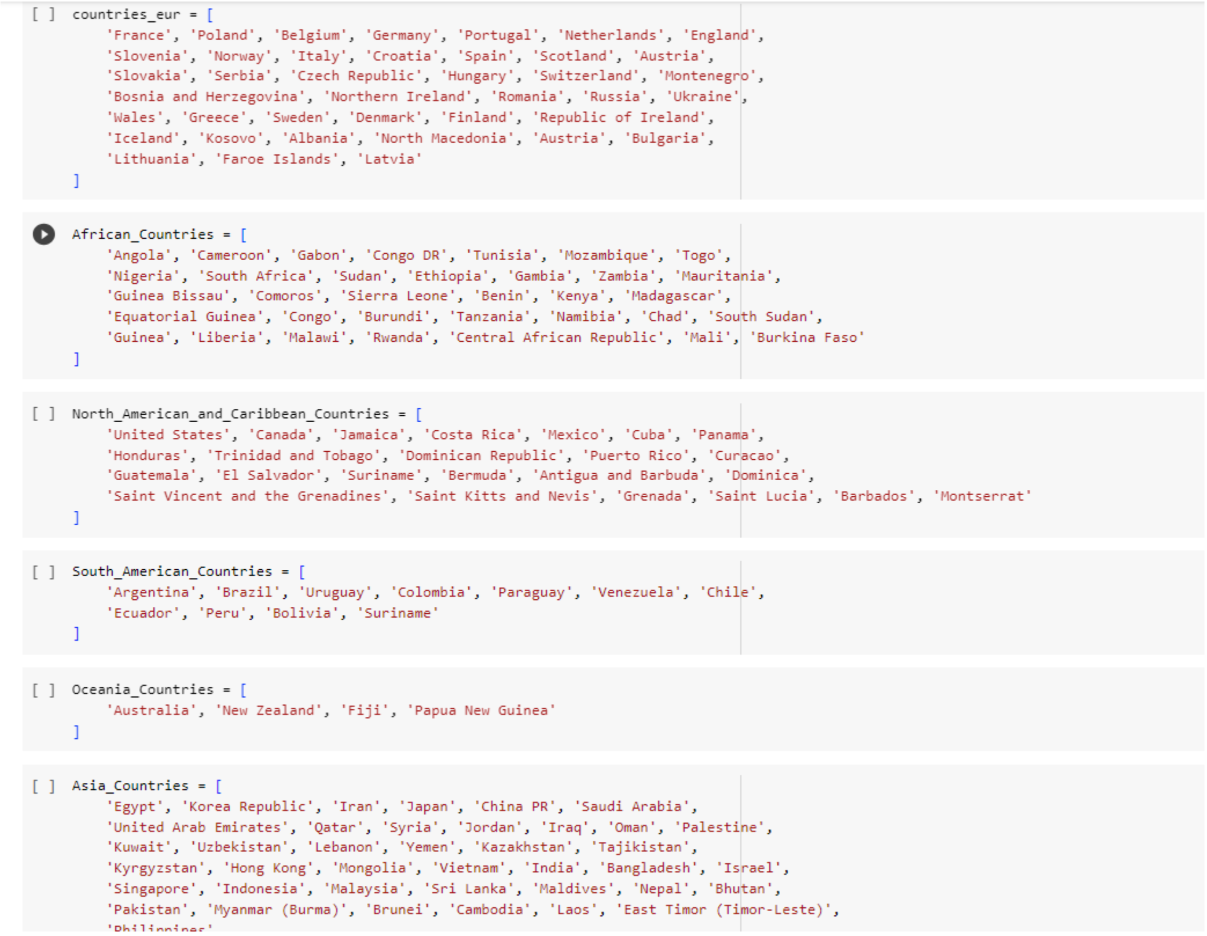
How do physical attributes such as physicality, stamina, strength, and acceleration vary across different continents in football, and can we determine which continent exhibits superiority in these aspects? This analysis seeks to quantify and compare the performance metrics of players from various continents, shedding light on the disparities in physical attributes and identifying potential strengths among continents in the realm of football

Since we have a list of countries, we have to segregate the countries as per their continents, I have divided them into six continents

Europe, Africa, North America and Caribbean, South American, Oceanic and Asia

Then for each continent we take into consideration all the different factors and calculate the average value for each feature for each and every continent

With the results of all continents, we can come to multiple conclusions and inferences on the basis of factors like Team Selection, Transfer Market, Scouting and Injuries



```
countries_eur = [
    'France', 'Poland', 'Belgium', 'Germany', 'Portugal', 'Netherlands', 'England',
    'Slovenia', 'Norway', 'Italy', 'Croatia', 'Spain', 'Scotland', 'Austria',
    'Slovakia', 'Serbia', 'Czech Republic', 'Hungary', 'Switzerland', 'Montenegro',
    'Bosnia and Herzegovina', 'Northern Ireland', 'Romania', 'Russia', 'Ukraine',
    'Wales', 'Greece', 'Sweden', 'Denmark', 'Finland', 'Republic of Ireland',
    'Iceland', 'Kosovo', 'Albania', 'North Macedonia', 'Austria', 'Bulgaria',
    'Lithuania', 'Faroe Islands', 'Latvia'
]
```

```
African_Countries = [
    'Angola', 'Cameroon', 'Gabon', 'Congo DR', 'Tunisia', 'Mozambique', 'Togo',
    'Nigeria', 'South Africa', 'Sudan', 'Ethiopia', 'Gambia', 'Zambia', 'Mauritania',
    'Guinea Bissau', 'Comoros', 'Sierra Leone', 'Benin', 'Kenya', 'Madagascar',
    'Equatorial Guinea', 'Congo', 'Burundi', 'Tanzania', 'Namibia', 'Chad', 'South Sudan',
    'Guinea', 'Liberia', 'Malawi', 'Rwanda', 'Central African Republic', 'Mali', 'Burkina Faso'
]
```

```
North_American_and_Caribbean_Countries = [
    'United States', 'Canada', 'Jamaica', 'Costa Rica', 'Mexico', 'Cuba', 'Panama',
    'Honduras', 'Trinidad and Tobago', 'Dominican Republic', 'Puerto Rico', 'Curacao',
    'Guatemala', 'El Salvador', 'Suriname', 'Bermuda', 'Antigua and Barbuda', 'Dominica',
    'Saint Vincent and the Grenadines', 'Saint Kitts and Nevis', 'Grenada', 'Saint Lucia', 'Barbados', 'Montserrat'
]
```

```
South_American_Countries = [
    'Argentina', 'Brazil', 'Uruguay', 'Colombia', 'Paraguay', 'Venezuela', 'Chile',
    'Ecuador', 'Peru', 'Bolivia', 'Suriname'
]
```

```
Oceania_Countries = [
    'Australia', 'New Zealand', 'Fiji', 'Papua New Guinea'
]
```

```
Asia_Countries = [
    'Egypt', 'Korea Republic', 'Iran', 'Japan', 'China PR', 'Saudi Arabia',
    'United Arab Emirates', 'Qatar', 'Syria', 'Jordan', 'Iraq', 'Oman', 'Palestine',
    'Kuwait', 'Uzbekistan', 'Lebanon', 'Yemen', 'Kazakhstan', 'Tajikistan',
    'Kyrgyzstan', 'Hong Kong', 'Mongolia', 'Vietnam', 'India', 'Bangladesh', 'Israel',
    'Singapore', 'Indonesia', 'Malaysia', 'Sri Lanka', 'Maldives', 'Nepal', 'Bhutan',
    'Pakistan', 'Myanmar (Burma)', 'Brunei', 'Cambodia', 'Laos', 'East Timor (Timor-Leste)',
    'Philippines'
```

Fig 1.10

Potential applications of the average scores and their interpretive utility in connection with these attributes.

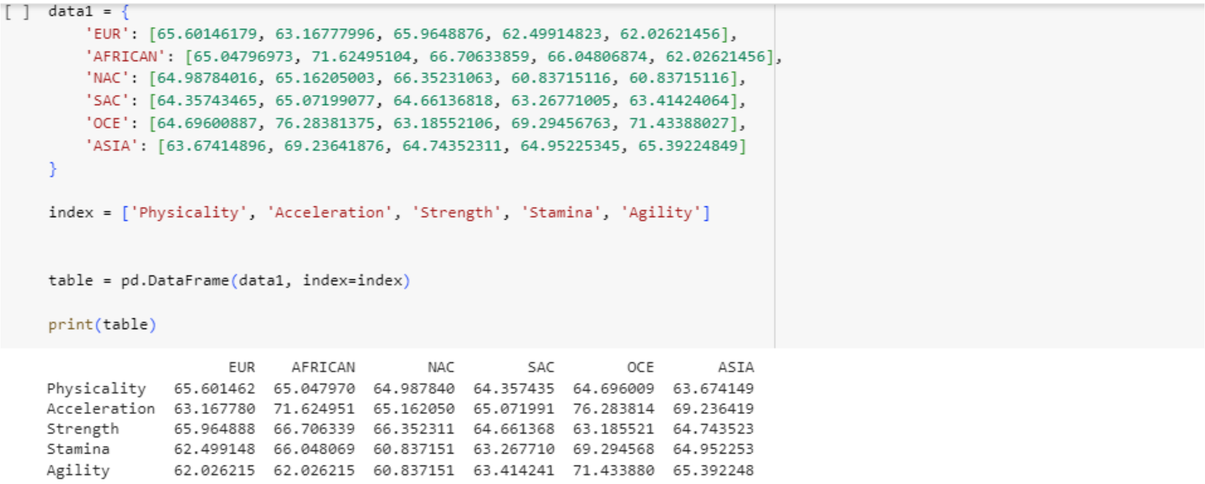
Team selection: Coaches and managers might consider these attributes when selecting players for their teams

Transfer market: Clubs often scout and sign players from all around the world

Scouting: When scouting for new talents, clubs might pay attention to regions that have higher average scores in certain attributes

Injuries: Certain attributes, such as strength and stamina, can influence a player's susceptibility to injuries

The table below consists of average values of all features for all the continents



```
[ ]  data1 = {
         'EUR': [65.60146179, 63.16777996, 65.9648876, 62.49914823, 62.02621456],
         'AFRICAN': [65.04796973, 71.62495104, 66.70633859, 66.04806874, 62.02621456],
         'NAC': [64.98784016, 65.16205003, 66.35231063, 60.83715116, 60.83715116],
         'SAC': [64.35743465, 65.07199077, 64.66136818, 63.26771005, 63.41424064],
         'OCE': [64.69600887, 76.28381375, 63.18552106, 69.29456763, 71.43388027],
         'ASIA': [63.67414896, 69.23641876, 64.74352311, 64.95225345, 65.39224849]
     }

     index = ['Physicality', 'Acceleration', 'Strength', 'Stamina', 'Agility']


     table = pd.DataFrame(data1, index=index)

     print(table)
```

```
                    EUR    AFRICAN        NAC        SAC        OCE       ASIA
Physicality   65.601462  65.047970  64.987840  64.357435  64.696009  63.674149
Acceleration  63.167780  71.624951  65.162050  65.071991  76.283814  69.236419
Strength      65.964888  66.706339  66.352311  64.661368  63.185521  64.743523
Stamina       62.499148  66.048069  60.837151  63.267710  69.294568  64.952253
Agility       62.026215  62.026215  60.837151  63.414241  71.433880  65.392248
```

Physicality:

African players have the highest average physicality score (65.047970), followed closely by European players (65.601462). This could be attributed to the playing styles in these continents, where physicality is often emphasized. African players, in particular, are known for their strength and athleticism.

Acceleration:

Asian players have the highest average acceleration score (76.283814), followed by African players (71.624951). This might be due to the emphasis on speed and agility in Asian football. Many Asian countries focus on developing quick and agile players.

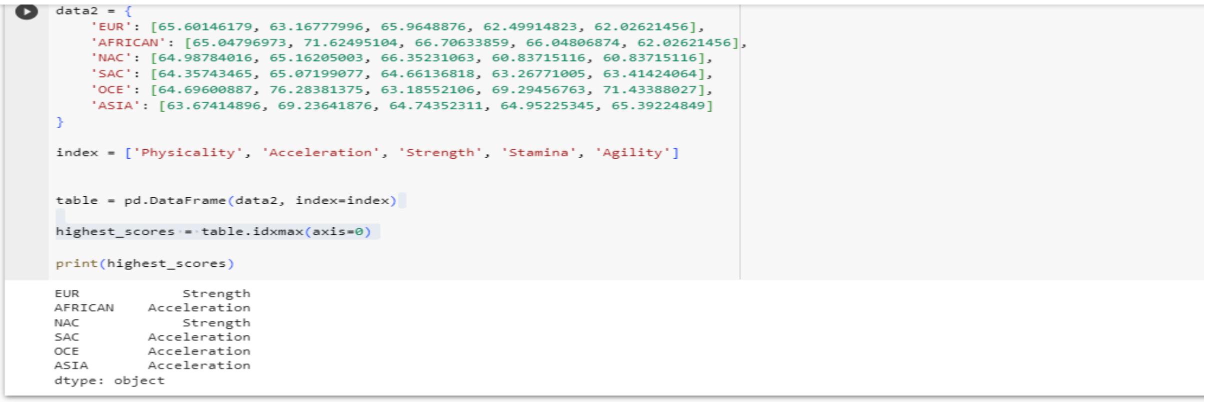
Strength:

African players again stand out with the highest average strength score (66.706339), followed by European players (65.964888). The physical nature of African football and the style of play often requiring strength could contribute to this trend.

Stamina:

African players have the highest average stamina score (66.048069), with North American players (NAC) having the lowest (60.837151). The difference could be due to varying playing conditions, training approaches, and game strategies in these continents.

Fig 1.11



```
data2 = {
    'EUR': [65.60146179, 63.16777996, 65.9648876, 62.49914823, 62.02621456],
    'AFRICAN': [65.04796973, 71.62495104, 66.70633859, 66.04806874, 62.02621456],
    'NAC': [64.98784016, 65.16205003, 66.35231063, 60.83715116, 60.83715116],
    'SAC': [64.35743465, 65.07199077, 64.66136818, 63.26771005, 63.41424064],
    'OCE': [64.69600887, 76.28381375, 63.18552106, 69.29456763, 71.43388027],
    'ASIA': [63.67414896, 69.23641876, 64.74352311, 64.95225345, 65.39224849]
}

index = ['Physicality', 'Acceleration', 'Strength', 'Stamina', 'Agility']


table = pd.DataFrame(data2, index=index)

highest_scores = table.idxmax(axis=0)

print(highest_scores)
```

```
EUR            Strength
AFRICAN    Acceleration
NAC            Strength
SAC        Acceleration
OCE        Acceleration
ASIA       Acceleration
dtype: object
```

African Players: African players tend to excel in physicality, strength, and stamina.African players' attributes suggest suitability for a physical and high-energy style of play. This might involve strong defensive capabilities, powerful attacking plays, and an ability to maintain intensity throughout matches This aligns with the perception of African football as being physically demanding and competitive.

Asian Players: Asian players show strengths in acceleration and agility, which reflects a focus on quick and agile gameplay, Asian players' strengths in acceleration and agility could lead to a fast-paced and dynamic playing style. This might involve quick transitions, intricate passing sequences, and emphasis on rapid attacks

European Players: European players generally have well-rounded scores across all attributes, with relatively high scores in physicality, acceleration, and strength,European players' balanced attributes make them adaptable to a variety of playing styles. European football often features a mix of technical skills, tactical discipline, and physicality. Teams might focus on possession-based play, pressing, or counter-attacks, depending on the situation

North American Players (NAC): North American players have comparatively lower scores in acceleration, agility, and stamina. This could be due to differing training methods and playing styles in North America,North American players' attributes could lead to a style that emphasizes strategic gameplay, set pieces, and organized defense. While still capable of offensive plays, this style might prioritize maintaining structure

Fig 1.12

From the list of all values, the below table represents the highest average attribute of each continent for example for strength Europe has the highest average value

Each continent has its own set of physical traits and it can be suited to certain sort of playing styles

There are many playing styles in world football for which these insights can be used for like

Total Football, Tiki Taka, Catenaccio and Gegenpressing

QUESTION 3:

What is the connection between a player's international reputation and key attributes such as wage, market value, player count, and age? This investigation aims to delve into the intricate relationship between a player's standing on the international stage and these fundamental attributes, seeking to analyse and quantify the interplay between reputation and player characteristics within the realm of football

Equating international reputation to revenue terms in football involves estimating how a player's reputation on the global stage can impact the financial aspects of the sport, the following features can be equated with international reputation of a player

- Increased Sponsorship and Endorsement Deals
- Ticket Sales and Merchandise
- TV and Broadcasting Rights
- Increased Commercial Opportunities

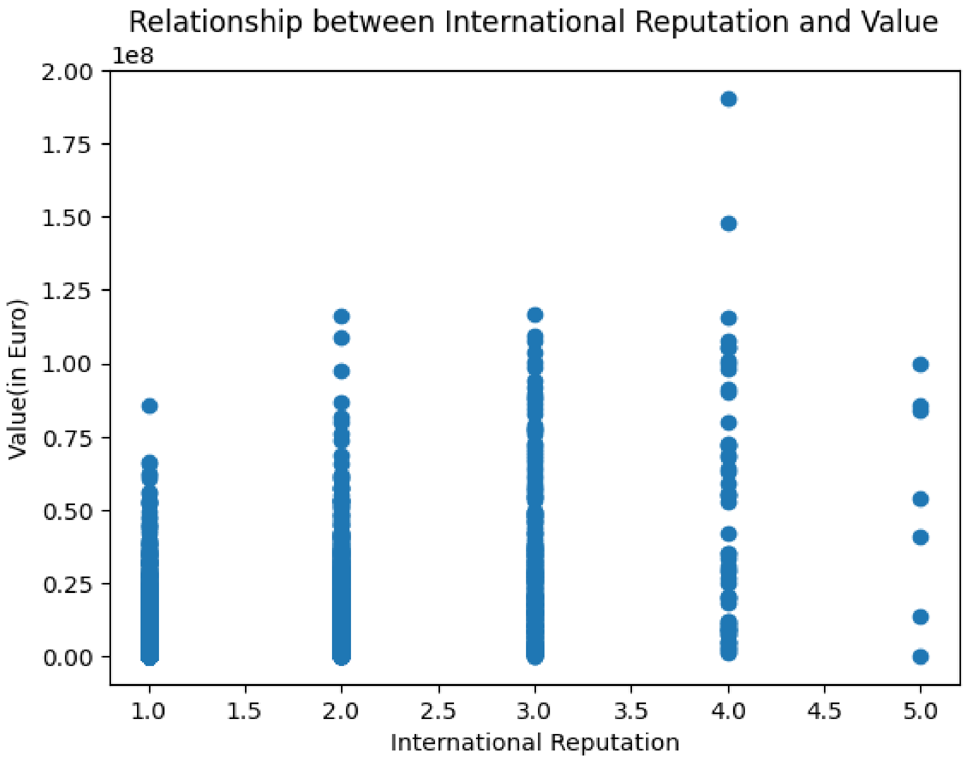
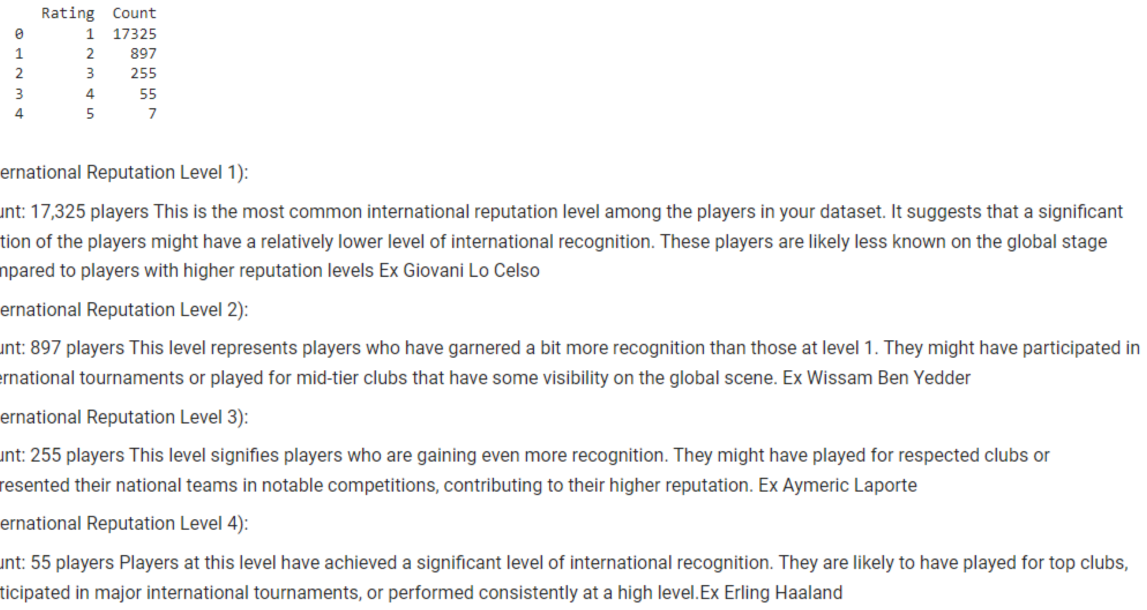


Fig 1.13 The above graph represents the International Reputation vs the value of each player



```
   Rating  Count
0       1  17325
1       2    897
2       3    255
3       4     55
4       5      7
```

(International Reputation Level 1):

Count: 17,325 players This is the most common international reputation level among the players in your dataset. It suggests that a significant portion of the players might have a relatively lower level of international recognition. These players are likely less known on the global stage compared to players with higher reputation levels Ex Giovani Lo Celso

(International Reputation Level 2):

Count: 897 players This level represents players who have garnered a bit more recognition than those at level 1. They might have participated in international tournaments or played for mid-tier clubs that have some visibility on the global scene. Ex Wissam Ben Yedder

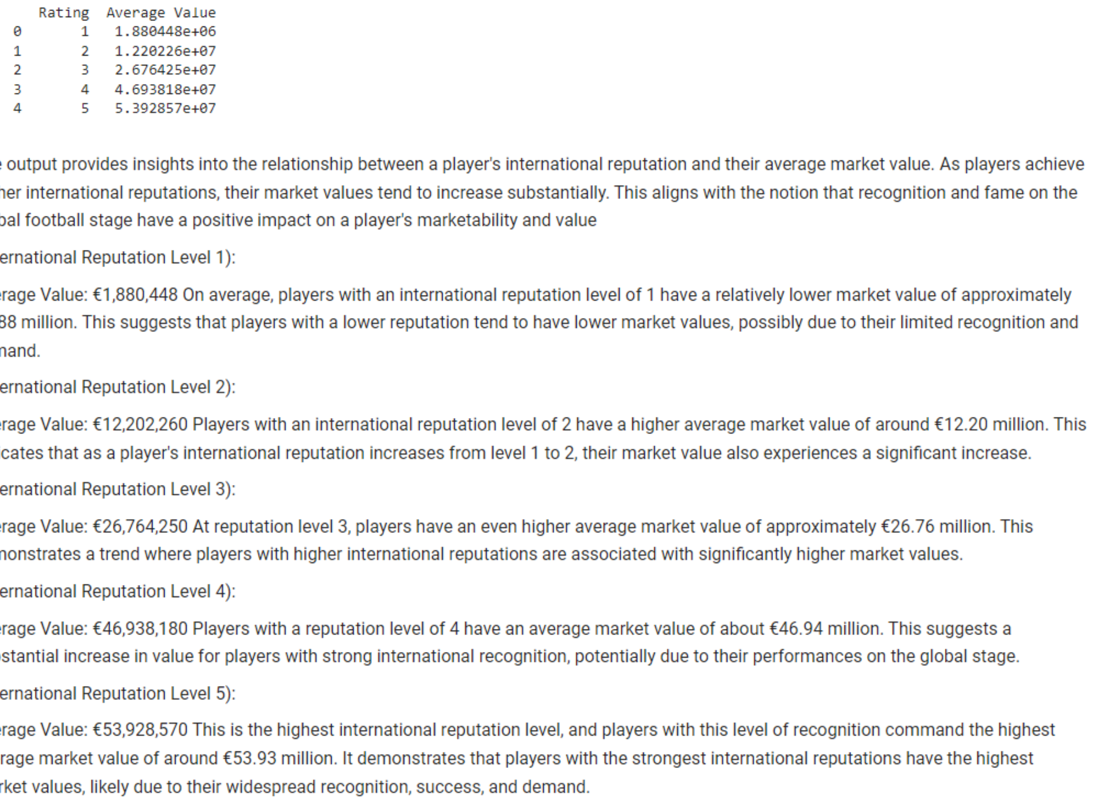
(International Reputation Level 3):

Count: 255 players This level signifies players who are gaining even more recognition. They might have played for respected clubs or represented their national teams in notable competitions, contributing to their higher reputation. Ex Aymeric Laporte

(International Reputation Level 4):

Count: 55 players Players at this level have achieved a significant level of international recognition. They are likely to have played for top clubs, participated in major international tournaments, or performed consistently at a high level.Ex Erling Haaland

Fig 1.14 The above screengrab represents the count of players as per the international reputation ratings with the insights of each rating level



```
   Rating  Average Value
0       1   1.880448e+06
1       2   1.220226e+07
2       3   2.676425e+07
3       4   4.693818e+07
4       5   5.392857e+07
```

The output provides insights into the relationship between a player's international reputation and their average market value. As players achieve higher international reputations, their market values tend to increase substantially. This aligns with the notion that recognition and fame on the global football stage have a positive impact on a player's marketability and value

(International Reputation Level 1):

Average Value: €1,880,448 On average, players with an international reputation level of 1 have a relatively lower market value of approximately €1.88 million. This suggests that players with a lower reputation tend to have lower market values, possibly due to their limited recognition and demand.

(International Reputation Level 2):

Average Value: €12,202,260 Players with an international reputation level of 2 have a higher average market value of around €12.20 million. This indicates that as a player's international reputation increases from level 1 to 2, their market value also experiences a significant increase.

(International Reputation Level 3):

Average Value: €26,764,250 At reputation level 3, players have an even higher average market value of approximately €26.76 million. This demonstrates a trend where players with higher international reputations are associated with significantly higher market values.

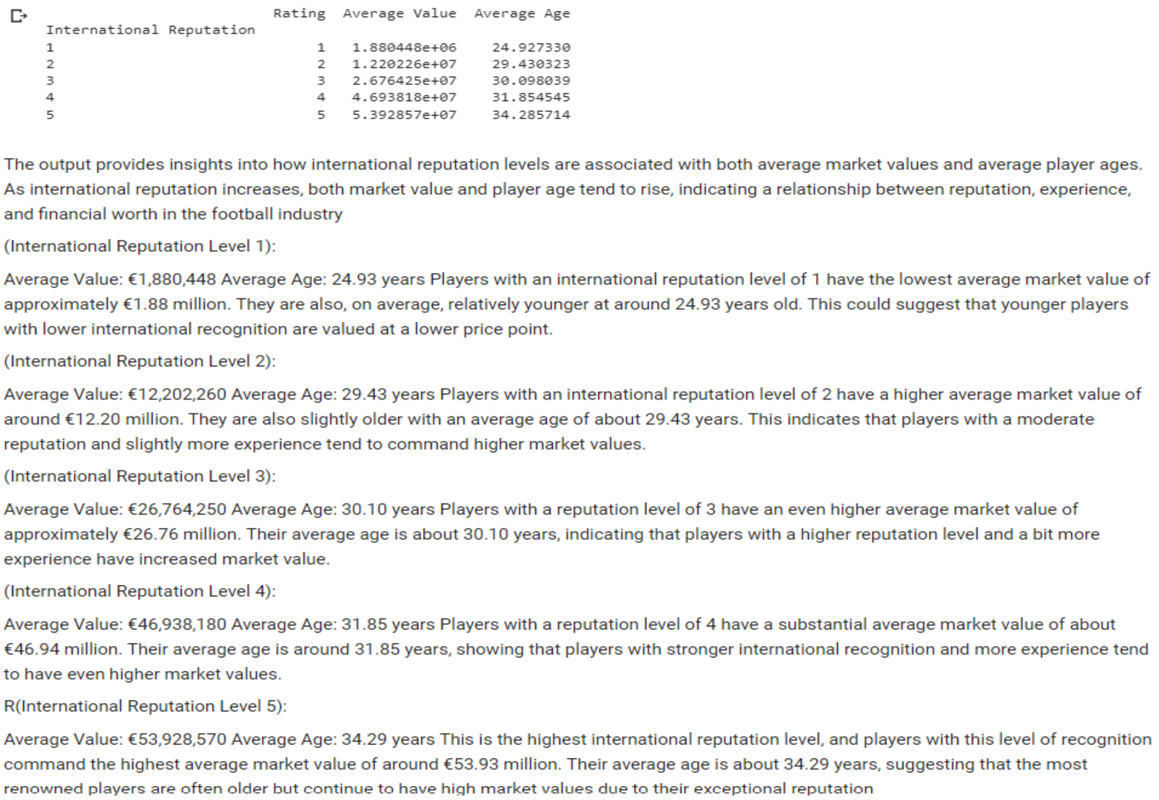
(International Reputation Level 4):

Average Value: €46,938,180 Players with a reputation level of 4 have an average market value of about €46.94 million. This suggests a substantial increase in value for players with strong international recognition, potentially due to their performances on the global stage.

(International Reputation Level 5):

Average Value: €53,928,570 This is the highest international reputation level, and players with this level of recognition command the highest average market value of around €53.93 million. It demonstrates that players with the strongest international reputations have the highest market values, likely due to their widespread recognition, success, and demand.

Fig 1.15 The above screengrab represents the average market value of players from different categories of international reputation with insights



```
                              Rating  Average Value  Average Age
International Reputation
1                                  1   1.880448e+06    24.927330
2                                  2   1.220226e+07    29.430323
3                                  3   2.676425e+07    30.098039
4                                  4   4.693818e+07    31.854545
5                                  5   5.392857e+07    34.285714
```

The output provides insights into how international reputation levels are associated with both average market values and average player ages. As international reputation increases, both market value and player age tend to rise, indicating a relationship between reputation, experience, and financial worth in the football industry

(International Reputation Level 1):

Average Value: €1,880,448 Average Age: 24.93 years Players with an international reputation level of 1 have the lowest average market value of approximately €1.88 million. They are also, on average, relatively younger at around 24.93 years old. This could suggest that younger players with lower international recognition are valued at a lower price point.

(International Reputation Level 2):

Average Value: €12,202,260 Average Age: 29.43 years Players with an international reputation level of 2 have a higher average market value of around €12.20 million. They are also slightly older with an average age of about 29.43 years. This indicates that players with a moderate reputation and slightly more experience tend to command higher market values.

(International Reputation Level 3):

Average Value: €26,764,250 Average Age: 30.10 years Players with a reputation level of 3 have an even higher average market value of approximately €26.76 million. Their average age is about 30.10 years, indicating that players with a higher reputation level and a bit more experience have increased market value.

(International Reputation Level 4):

Average Value: €46,938,180 Average Age: 31.85 years Players with a reputation level of 4 have a substantial average market value of about €46.94 million. Their average age is around 31.85 years, showing that players with stronger international recognition and more experience tend to have even higher market values.

R(International Reputation Level 5):

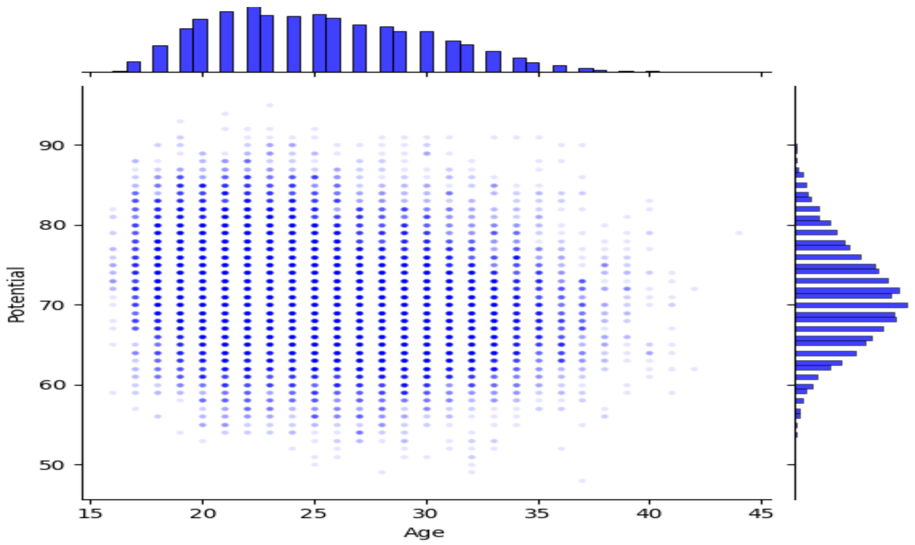
Average Value: €53,928,570 Average Age: 34.29 years This is the highest international reputation level, and players with this level of recognition command the highest average market value of around €53.93 million. Their average age is about 34.29 years, suggesting that the most renowned players are often older but continue to have high market values due to their exceptional reputation

Fig 1.16 The above screengrab represents the average market value and age of players from different categories of international reputation with insights

Question 4:

What is the correlation between the age of football players and crucial attributes such as wage, potential, market value, and sprint speed? This inquiry aims to investigate and quantify the associations between player age and these key characteristics

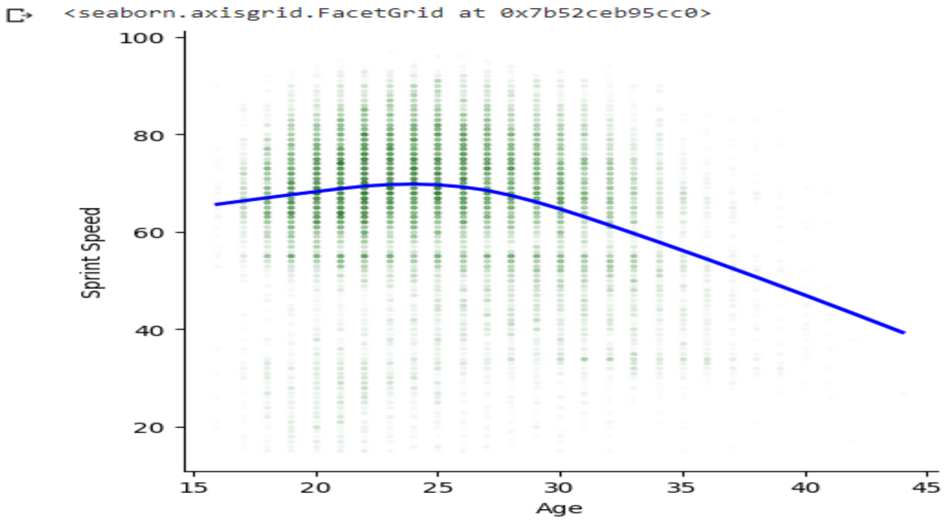
Age vs Potential



Potential tends to fall as you grow old in terms of age

Fig 1.17

Age vs Sprint Speed

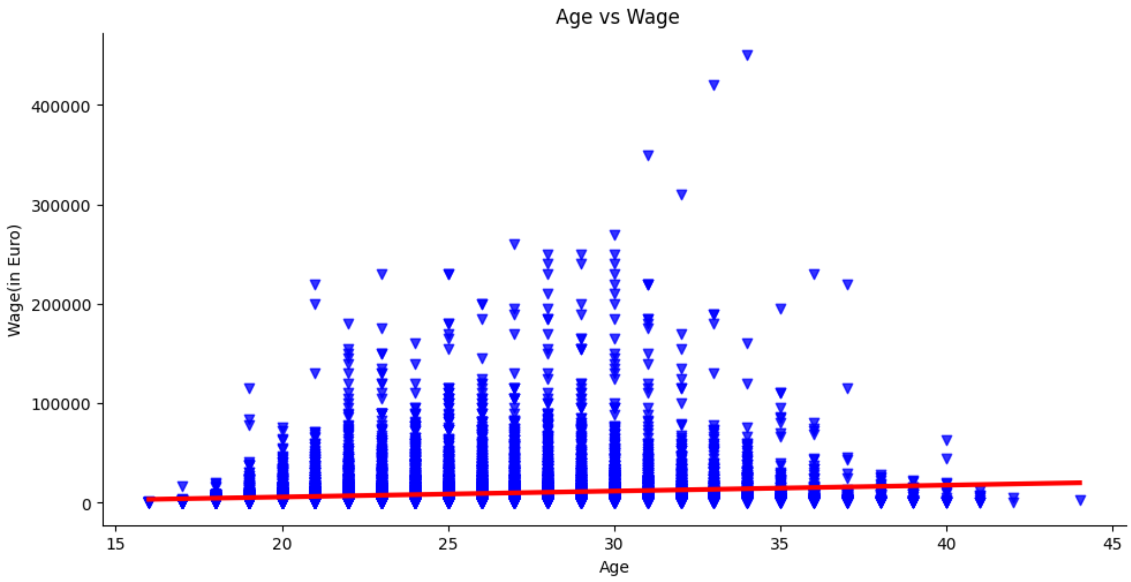


```
<seaborn.axisgrid.FacetGrid at 0x7b52ceb95cc0>
```

As the age increases the sprint speed decreases drastically

Fig 1.18

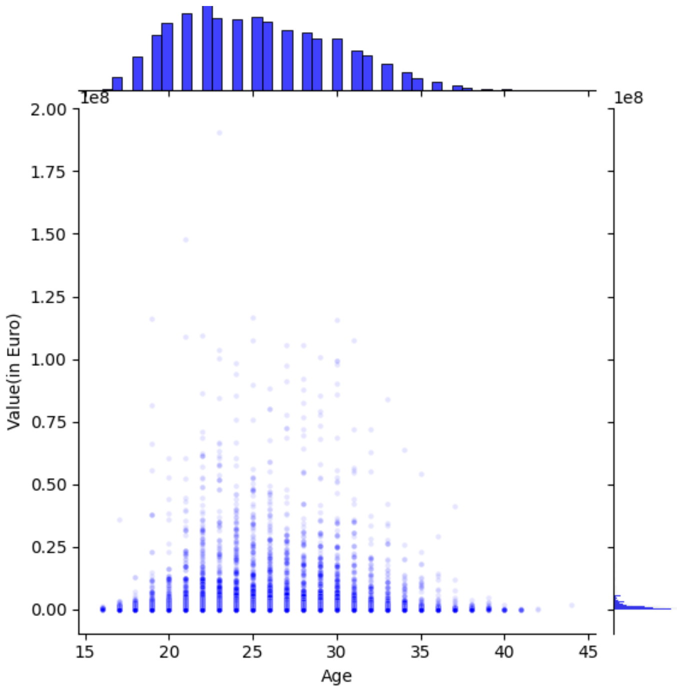
Age vs Wage



Wage Decrease as Age Increases, wage gradually decreases as the age approaches 30 to 40 years except for a few exceptions

Fig 1.19

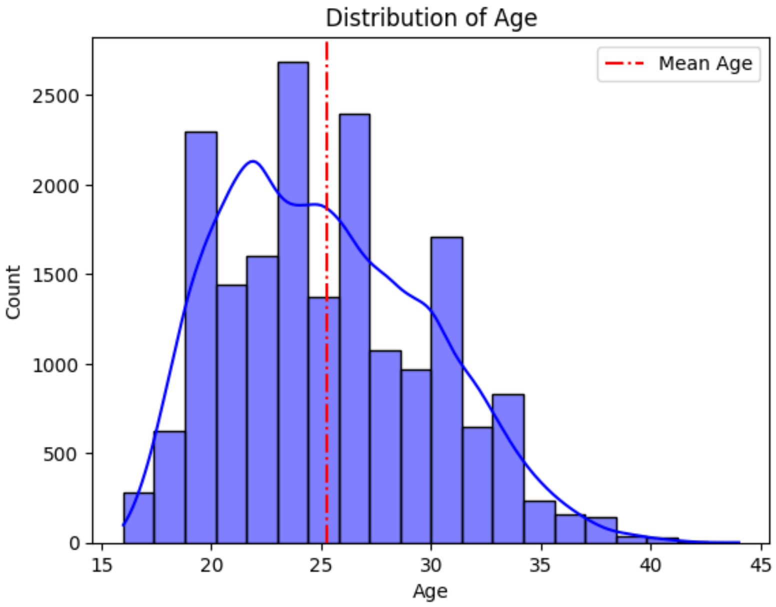
Age vs Value



Player value experiences a significant decline beyond the age of 23, while 23 marks the peak age for players, characterized by their highest value

Fig 1.20

Distribution of Age



The depicted graph illustrates the age distribution of players in the dataset, with the dataset's mean age being 25.24 years.

Fig 1.21

QUESTION 5:

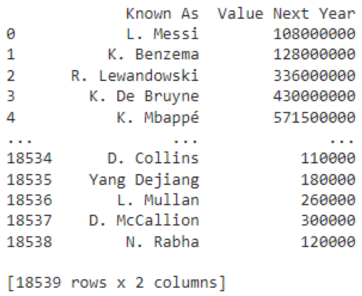
a) Create a metric called as Value next year, that give's the player's expected value next year

b) Visualize the top 10 players with highest expected values for next year

c)Countries with the greatest number of players

d)Clubs with the greatest number of players as present in the dataset

a)



```
              Known As  Value Next Year
0             L. Messi        108000000
1           K. Benzema        128000000
2       R. Lewandowski        336000000
3         K. De Bruyne        430000000
4            K. Mbappé        571500000
...                ...              ...
18534       D. Collins           110000
18535      Yang Dejiang           180000
18536        L. Mullan           260000
18537      D. McCallion           300000
18538         N. Rabha           120000

[18539 rows x 2 columns]
```

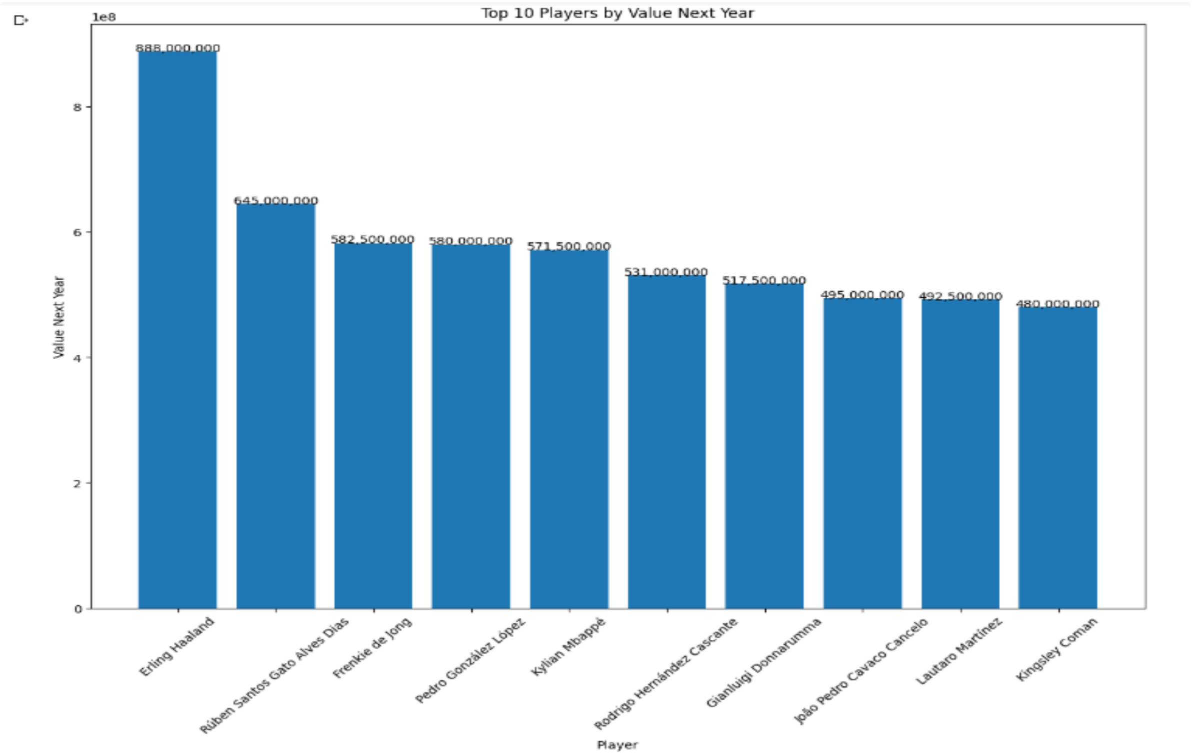
This metric calculates and displays an estimate of how players' values might change in the next year based on their current values and the number of years remaining in their contracts. This estimation considers the influence of contract length on a player's perceived market value

For each player, it subtracts the current year (2022) from the year their contract ends, yielding the number of years left in their contract

This 'Value Next Year' estimation considers the impact of contract length on a player's perceived market value

Fig 1.22

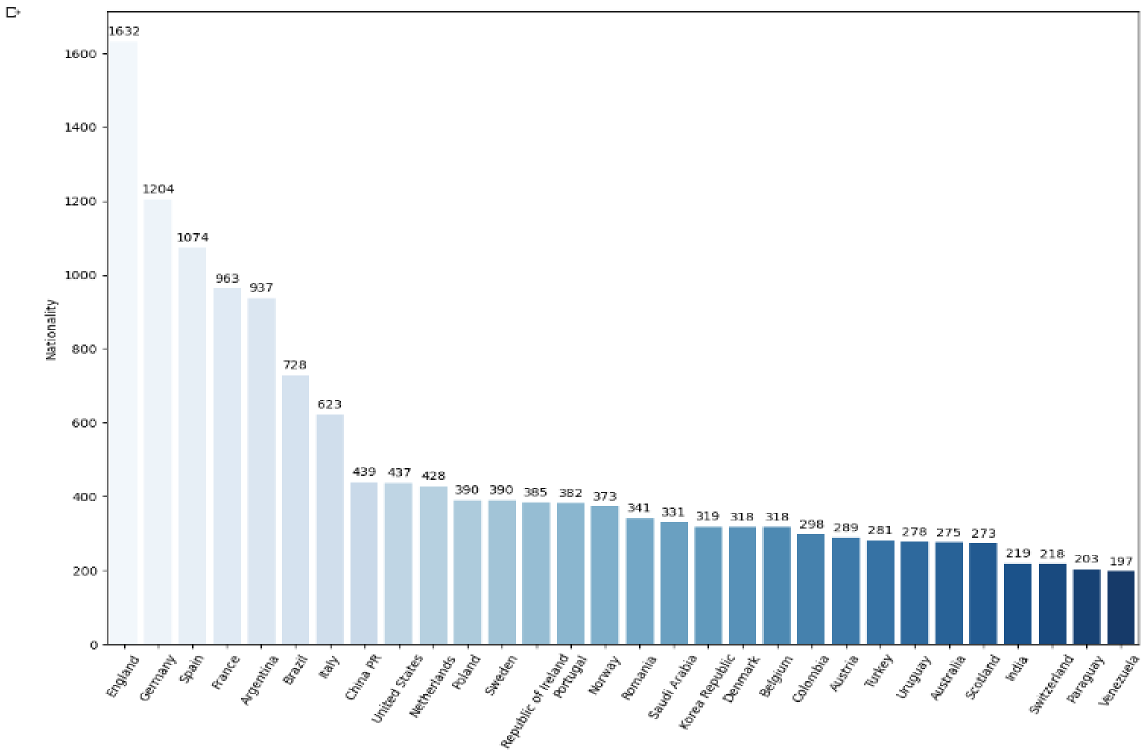
b)



The depicted graph showcases the roster of the top 10 players anticipated to have the greatest value in the upcoming year. Erling Haaland claims the top spot as the most valuable player, closely followed by his Manchester City colleague Ruben Dias. The compilation features four Manchester City players, underscoring the presence of youthful and exceptionally valuable talents within the team.

Fig 1.23

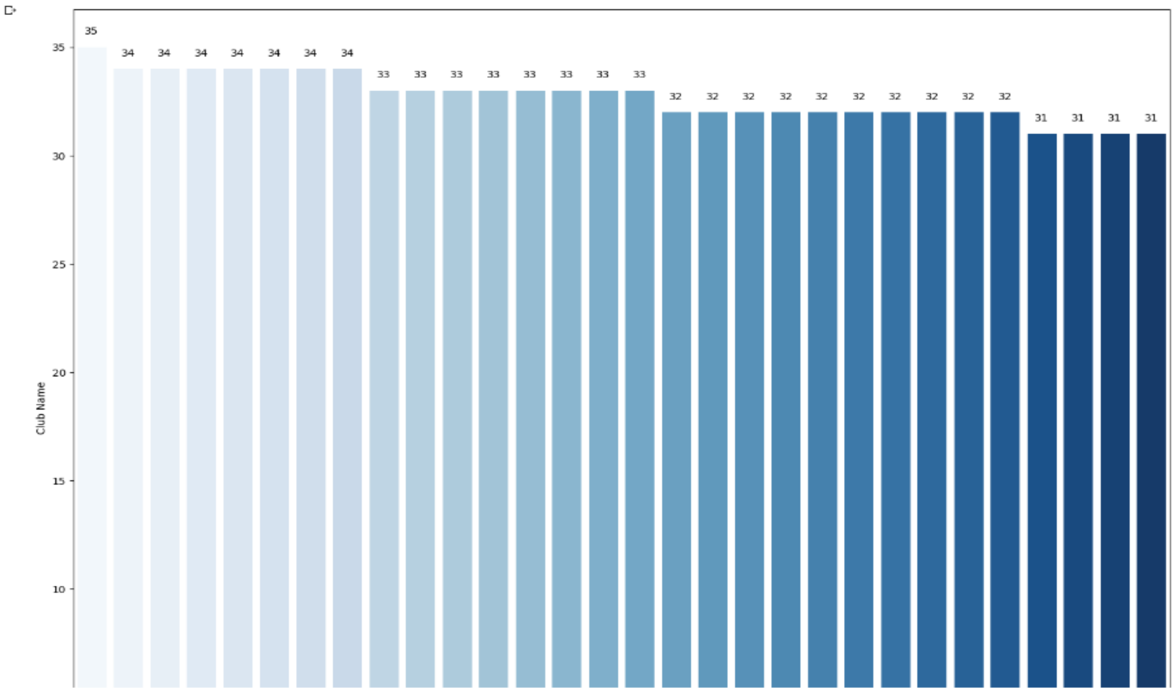
c)



The bar plot illustrates how the top 30 nationalities are distributed among the players in the dataset. Each bar corresponds to a specific nationality and its height reflects the number of players belonging to that nationality.England has the highest number of players in the dataset.

Fig 1.24

d)



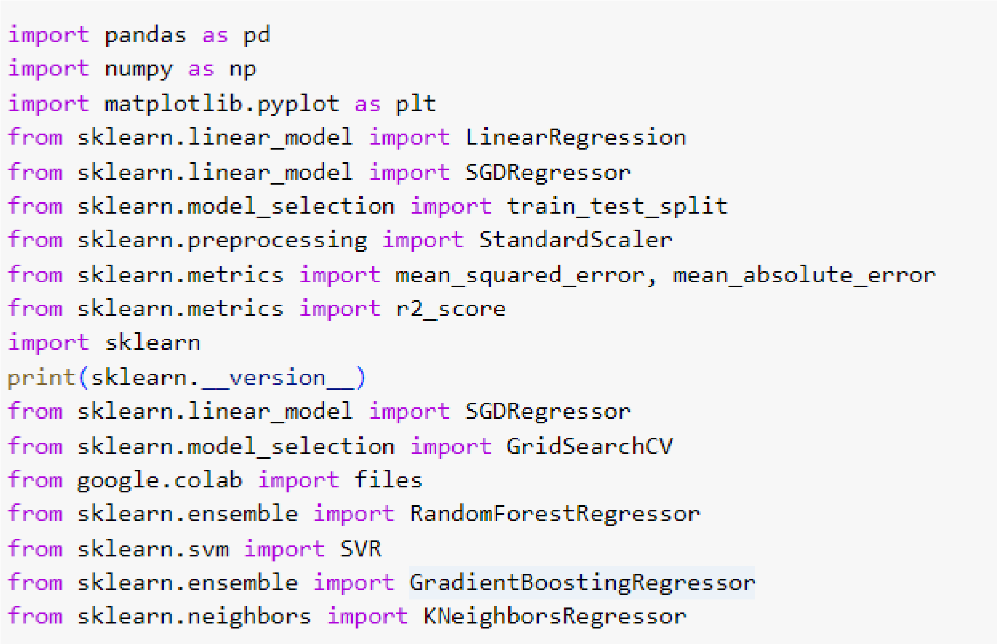
The bar plot provides a visual depiction of how the dataset's top 30 football clubs are distributed among the players. Each bar is associated with a distinct club, and its vertical extent signifies the count of players affiliated with that club. Notably, Olympique de Marseille stands out as the club with the largest representation of players among all individual club

Fig 1.25

MODEL BUILDING AND IMPLEMENTATION OF RESULTS

There are multiple steps with regards to building a prediction model as mentioned earlier in the research

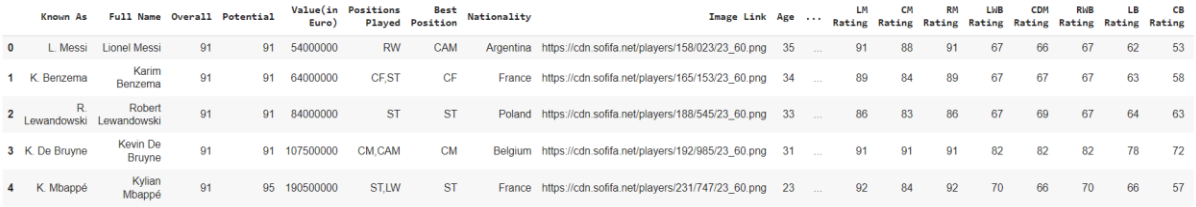
Step 1: Load all the libraries required to build the prediction model



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import SGDRegressor
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error, mean_absolute_error
from sklearn.metrics import r2_score
import sklearn
print(sklearn.__version__)
from sklearn.linear_model import SGDRegressor
from sklearn.model_selection import GridSearchCV
from google.colab import files
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.neighbors import KNeighborsRegressor
```

Fig 1.26

Step 2: Load the required the dataset into the python environment



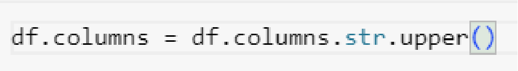
	Known As	Full Name	Overall	Potential	Value(in Euro)	Positions Played	Best Position	Nationality	Image Link	Age	...	LM Rating	CM Rating	RM Rating	LWB Rating	CDM Rating	RWB Rating	LB Rating	CB Rating
0	L. Messi	Lionel Messi	91	91	54000000	RW	CAM	Argentina	https://cdn.sofifa.net/players/158/023/23_60.png	35	...	91	88	91	67	66	67	62	53
1	K. Benzema	Karim Benzema	91	91	64000000	CF,ST	CF	France	https://cdn.sofifa.net/players/165/153/23_60.png	34	...	89	84	89	67	67	67	63	58
2	R. Lewandowski	Robert Lewandowski	91	91	84000000	ST	ST	Poland	https://cdn.sofifa.net/players/188/545/23_60.png	33	...	86	83	86	67	69	67	64	63
3	K. De Bruyne	Kevin De Bruyne	91	91	107500000	CM,CAM	CM	Belgium	https://cdn.sofifa.net/players/192/985/23_60.png	31	...	91	91	91	82	82	82	78	72
4	K. Mbappé	Kylian Mbappé	91	95	190500000	ST,LW	ST	France	https://cdn.sofifa.net/players/231/747/23_60.png	23	...	92	84	92	70	66	70	66	57

Fig 1.27

Step 3: Data Cleaning

For prediction models to produce accurate, dependable results, data cleaning is essential because it gets rid of mistakes, outliers, and inconsistencies that could skew model training and predictions.

The multiple steps in our data cleaning process are as follows



```
df.columns = df.columns.str.upper()
```

Fig 1.28 The above step converts the column names from lower to upper case as it offers advantages like improved readability and consistent formatting

To remove features which are irrelevant, less interpretable and of no use while building the prediction models, so we need to drop columns from the dataset.

The prediction model we are building is one wherein the potential Is predicted for the future, so keeping that in mind all the features have to selected factually as so it can contribute towards the efficiency of the model

The following columns have been removed from the dataset:

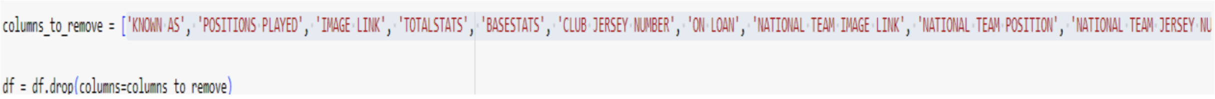
'KNOWN AS', 'POSITIONS PLAYED', 'IMAGE LINK', 'TOTALSTATS', 'BASESTATS', 'CLUB JERSEY NUMBER', 'ON LOAN', 'NATIONAL TEAM IMAGE LINK', 'NATIONAL TEAM POSITION', 'NATIONAL TEAM JERSEY NUMBER', 'INTERNATIONAL REPUTATION', 'HEIGHT (IN CM)', 'WEIGHT (IN KG)','NATIONAL TEAM NAME.

There are a multitude of reasons so as to why these columns have been removed from the dataset, some of them are as follows

Similar data points (Known as and Full name)-Both of these headers have similar data points, so it makes sense to opt for of them

Data headers like 'NATIONAL TEAM IMAGE LINK', 'NATIONAL TEAM JERSEY NUMBER', 'NATIONAL TEAM POSITION' and 'CLUB JERSEY NUMBER' have no relation at all with what we are trying to predict from our prediction models

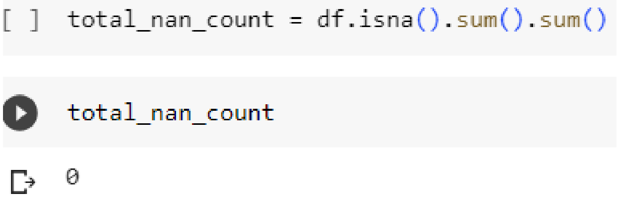
Data headers like 'HEIGHT (IN CM)', 'WEIGHT (IN KG)', 'TOTALSTATS' and 'BASESTATS' although containing numerical values have no bearings with regards to predicting the potential, hence we have to drop them.



```
columns_to_remove = ['KNOWN AS', 'POSITIONS PLAYED', 'IMAGE LINK', 'TOTALSTATS', 'BASESTATS', 'CLUB JERSEY NUMBER', 'ON LOAN', 'NATIONAL TEAM IMAGE LINK', 'NATIONAL TEAM POSITION', 'NATIONAL TEAM JERSEY NU

df = df.drop(columns=columns to remove)
```

Fig 1.29

We have to check for any null values in the dataframe and then they need to be removed from the dataframe as null values often cause alarming issues such as Data Integrity issues, Algorithm compatibility issues and mainly model performance issues.



```
[ ]  total_nan_count = df.isna().sum().sum()

     total_nan_count

     0
```

Fig 1.30 Total number of null values in our dataframe

CORRELATION HEATMAP

Building a correlation heatmap is an essential step in prediction model building as it is the mainstay in identifying the main entities for features that we want to use for predicting the target variable POTENTIAL

The features which have the highest value in the heatmap with regards to POTENTIAL is selected, the following features have been selected to predict the target variable (POTENTIAL)

OVERALL (0.66)

WAGE IN EURO (0.50)

VALUE IN EURO (0.45)

RELEASE CLAUSE (0.51)

DRIBBLING (0.50)

PASSING (0.43)

REACTION (0.54)

The range of values from 0.43-0.66

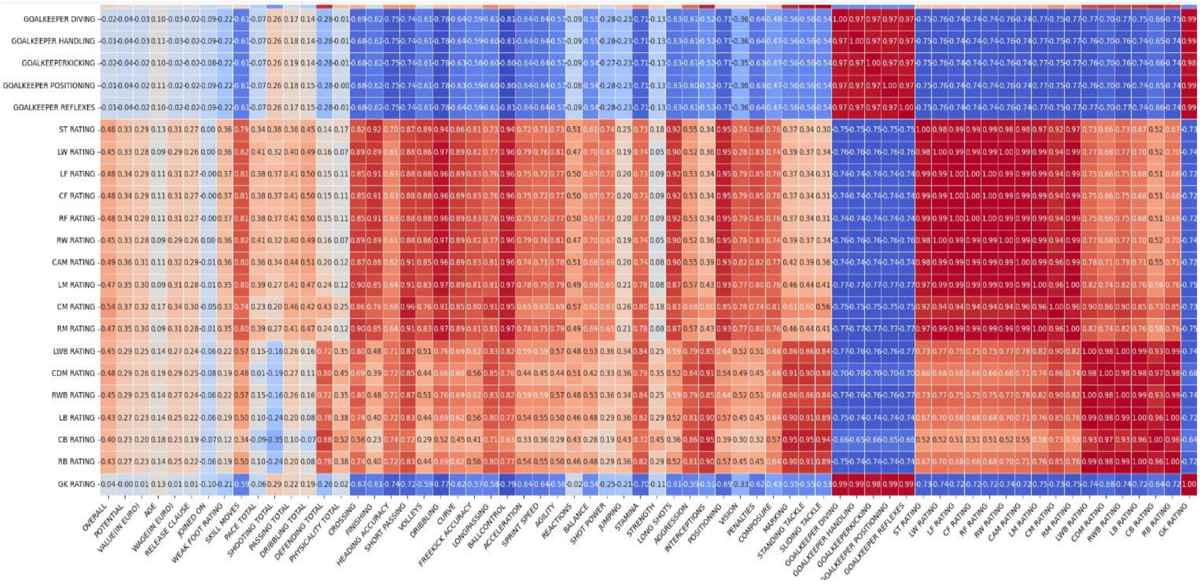


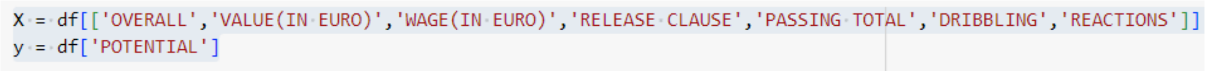
Fig 1.31

MODEL BUILDING

The five models used in my research are

LINEAR REGRESSION:

For our Linear Regression model the feature and the target variables are



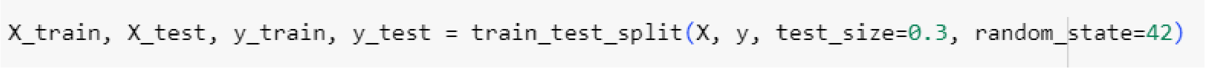
```
X = df[['OVERALL','VALUE(IN EURO)','WAGE(IN EURO)','RELEASE CLAUSE','PASSING TOTAL','DRIBBLING','REACTIONS']]
y = df['POTENTIAL']
```

Fig 1.32

The feature variables are selected on the basis of the highest correlation scores as based on the heatmap while the target variable is the one that we need to predict

X is the independent variable while Y is dependent variable

After this we need to split the dataset into training and testing sets using the train_test_split function



```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Fig 1.33

X_train: This will contain a subset of our predictor variables X that we'll use for our Linear Regression model, it contains 70% (in this case) of our data because we've specified test_size=0.3, meaning we want to reserve 30% of our data for testing

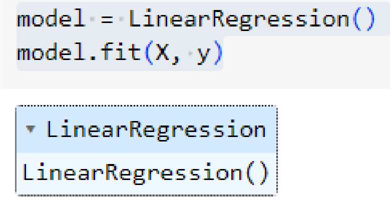
X_test: This will contain the remaining 30% of our predictor variables. We will use this set to evaluate our model's performance on data it has never seen before. This is essential to check how well our model reacts to new and unseen data.

y_train: This contains a subset of our target variable Y (corresponding to X_train). It's used for training our model.

y_test: This contains the remaining portion of our target variable corresponding to X_test. We will use this set to evaluate our model's predictions against the actual target values

The random_state parameter is set to 42 to ensure that the data split is reproducible when running the code multiple times.

Our machine learning model can be trained on X_train and y_train following this data split, and its performance can then be evaluated on X_test and y_test to determine how well it reacts to new data.

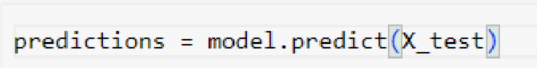


```
model = LinearRegression()
model.fit(X, y)
```

```
▾ LinearRegression
LinearRegression()
```

Fig 1.34

Using our predictor variables X and Y, this code builds a linear regression model and fits it, enabling the model to discover the link between them.



```
predictions = model.predict(X_test)
```

Fig 1.35

This makes predictions on the testing dataset X_test using the trained linear regression model, saving the results in the predictions variable.

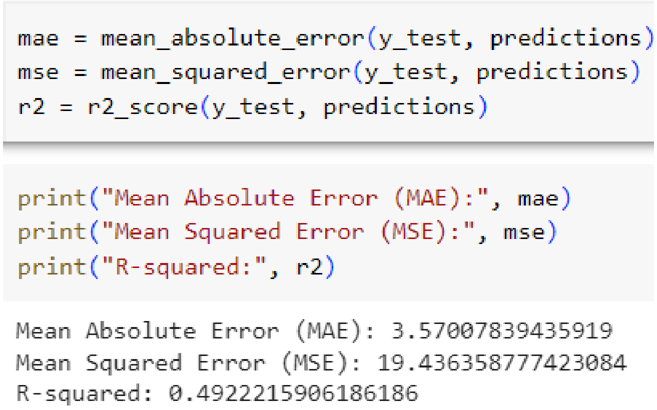
Following these steps, we have to now calculate the metrics which assess the performance and the accuracy of the linear regression model on the test dataset

The three metrics under consideration are MAE, MSE and R SQUARED

MAE-Mean Absolute Error represents the average absolute difference between your model's predictions and the actual values

MSE- Mean Square Error represents the average squared difference between your model's predictions and the actual values

The coefficient of determination (R-squared) is a measure of how well the model fits the data



```
mae = mean_absolute_error(y_test, predictions)
mse = mean_squared_error(y_test, predictions)
r2 = r2_score(y_test, predictions)
```

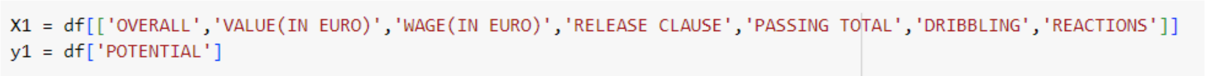
```
print("Mean Absolute Error (MAE):", mae)
print("Mean Squared Error (MSE):", mse)
print("R-squared:", r2)
```

```
Mean Absolute Error (MAE): 3.57007839435919
Mean Squared Error (MSE): 19.436358777423084
R-squared: 0.4922215906186186
```

Fig 1.36

K NEAREST NEIGHBORS REGRESSION

For our K Nearest Neighbors Regression model the feature and the target variables are



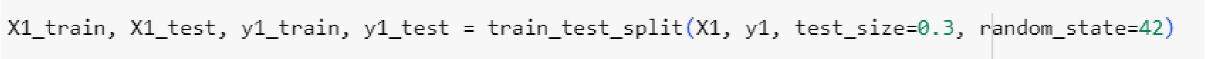
```
X1 = df[['OVERALL','VALUE(IN EURO)','WAGE(IN EURO)','RELEASE CLAUSE','PASSING TOTAL','DRIBBLING','REACTIONS']]
y1 = df['POTENTIAL']
```

Fig 1.37

The feature variables are selected on the basis of the highest correlation scores as based on the heatmap while the target variable is the one that we need to predict

X is the independent variable while Y is dependent variable

After this we need to split the dataset into training and testing sets using the train_test_split function



```
X1_train, X1_test, y1_train, y1_test = train_test_split(X1, y1, test_size=0.3, random_state=42)
```

Fig 1.38

X_train: This will contain a subset of our predictor variables X that we'll use for our K Nearest Neighbors Regression model, it contains 70% (in this case) of our data because we've specified test_size=0.3, meaning we want to reserve 30% of our data for testing

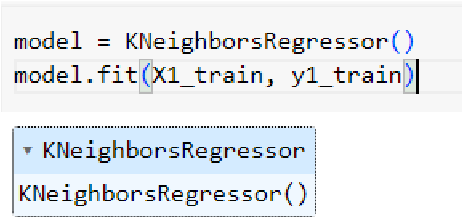
X_test: This will contain the remaining 30% of our predictor variables. We will use this set to evaluate our model's performance on data it has never seen before. This is essential to check how well our model reacts to new and unseen data.

y_train: This contains a subset of our target variable Y (corresponding to X_train). It's used for training our model.

y_test: This contains the remaining portion of our target variable corresponding to X_test. We will use this set to evaluate our model's predictions against the actual target values

The random_state parameter is set to 42 to ensure that the data split is reproducible when running the code multiple times.

Our machine learning model can be trained on X_train and y_train following this data split, and its performance can then be evaluated on X_test and y_test to determine how well it reacts to new data

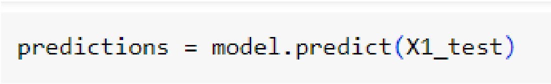


```
model = KNeighborsRegressor()
model.fit(X1_train, y1_train)
```

```
▾ KNeighborsRegressor
KNeighborsRegressor()
```

Fig 1.39

Using our predictor variables X and Y, this code builds a K Nearest Neighbors and fits it, enabling the model to discover the link between them



```
predictions = model.predict(X1_test)
```

Fig 1.40

This makes predictions on the testing dataset X_test using the trained K Nearest Neighbors Regression, saving the results in the predictions variable.

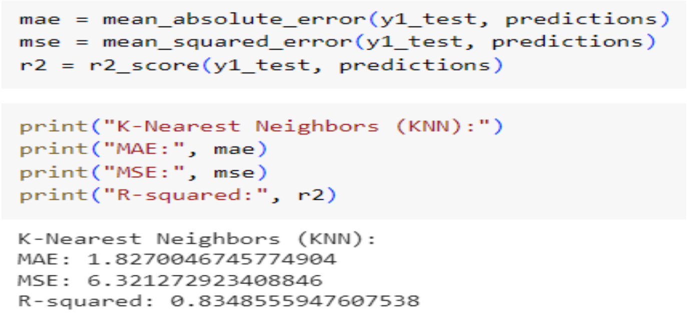
Following these steps, we have to now calculate the metrics which assess the performance and the accuracy of the K Nearest Neighbors Regression on the test dataset

The three metrics under consideration are MAE, MSE and R SQUARED

MAE-Mean Absolute Error represents the average absolute difference between your model's predictions and the actual values

MSE- Mean Square Error represents the average squared difference between your model's predictions and the actual values

The coefficient of determination (R-squared) is a measure of how well the model fits the data



```
mae = mean_absolute_error(y1_test, predictions)
mse = mean_squared_error(y1_test, predictions)
r2 = r2_score(y1_test, predictions)
```

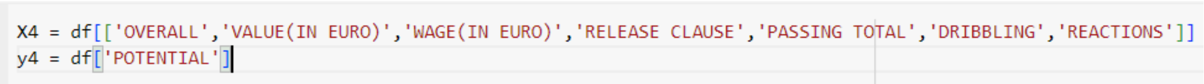
```
print("K-Nearest Neighbors (KNN):")
print("MAE:", mae)
print("MSE:", mse)
print("R-squared:", r2)
```

```
K-Nearest Neighbors (KNN):
MAE: 1.8270046745774904
MSE: 6.321272923408846
R-squared: 0.8348555947607538
```

Fig 1.41 Final result of KNN model

SUPPORT VECTOR REGRESSION

For our Support Vector Regression model the feature and the target variables are



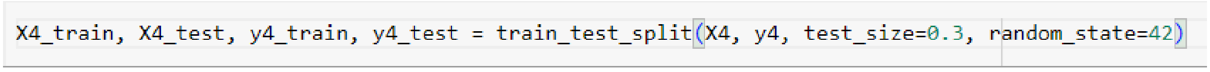
```
X4 = df[['OVERALL','VALUE(IN EURO)','WAGE(IN EURO)','RELEASE CLAUSE','PASSING TOTAL','DRIBBLING','REACTIONS']]
y4 = df['POTENTIAL']
```

Fig 1.42

The feature variables are selected on the basis of the highest correlation scores as based on the heatmap while the target variable is the one that we need to predict

X is the independent variable while Y is dependent variable

After this we need to split the dataset into training and testing sets using the train_test_split function



```
X4_train, X4_test, y4_train, y4_test = train_test_split(X4, y4, test_size=0.3, random_state=42)
```

Fig 1.43

X_train: This will contain a subset of our predictor variables X that we'll use for our Support Vector Regression model, it contains 70% (in this case) of our data because we've specified test_size=0.3, meaning we want to reserve 30% of our data for testing

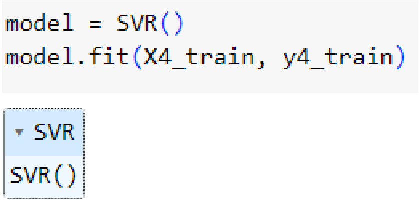
X_test: This will contain the remaining 30% of our predictor variables. We will use this set to evaluate our model's performance on data it has never seen before. This is essential to check how well our model reacts to new and unseen data.

y_train: This contains a subset of our target variable Y (corresponding to X_train). It's used for training our model.

y_test: This contains the remaining portion of our target variable corresponding to X_test. We will use this set to evaluate our model's predictions against the actual target values

The random_state parameter is set to 42 to ensure that the data split is reproducible when running the code multiple times.

Our machine learning model can be trained on X_train and y_train following this data split, and its performance can then be evaluated on X_test and y_test to determine how well it reacts to new data

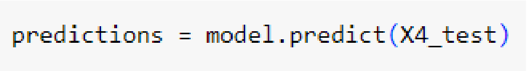


```
model = SVR()
model.fit(X4_train, y4_train)
```

```
▾ SVR
SVR()
```

Fig 1.44

Using our predictor variables X and Y, this code builds a Support Vector Regression model and fits it, enabling the model to discover the link between them



```
predictions = model.predict(X4_test)
```

Fig 1.45

This makes predictions on the testing dataset X_test using the trained Support Vector Regression, saving the results in the predictions variable.

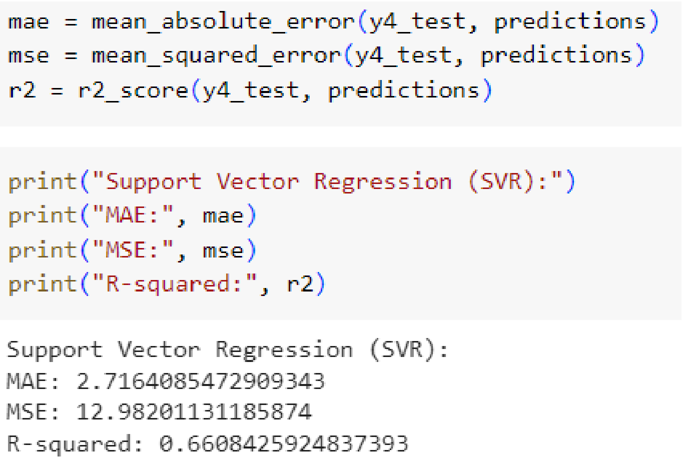
Following these steps, we have to now calculate the metrics which assess the performance and the accuracy of the Support Vector Regression on the test dataset

The three metrics under consideration are MAE, MSE and R SQUARED

MAE-Mean Absolute Error represents the average absolute difference between your model's predictions and the actual values

MSE- Mean Square Error represents the average squared difference between your model's predictions and the actual values

The coefficient of determination (R-squared) is a measure of how well the model fits the data



```
mae = mean_absolute_error(y4_test, predictions)
mse = mean_squared_error(y4_test, predictions)
r2 = r2_score(y4_test, predictions)
```

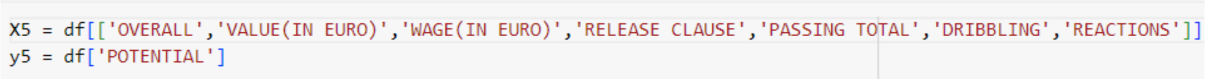
```
print("Support Vector Regression (SVR):")
print("MAE:", mae)
print("MSE:", mse)
print("R-squared:", r2)
```

```
Support Vector Regression (SVR):
MAE: 2.7164085472909343
MSE: 12.98201131185874
R-squared: 0.6608425924837393
```

Fig 1.46 Final result of SVR model

GRADIENT BOOSTING REGRESSION

For our Gradient Boosting Regression model the feature and the target variables are



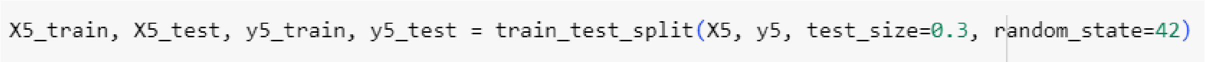
```
X5 = df[['OVERALL','VALUE(IN EURO)','WAGE(IN EURO)','RELEASE CLAUSE','PASSING TOTAL','DRIBBLING','REACTIONS']]
y5 = df['POTENTIAL']
```

Fig 1.47

The feature variables are selected on the basis of the highest correlation scores as based on the heatmap while the target variable is the one that we need to predict

X is the independent variable while Y is dependent variable

After this we need to split the dataset into training and testing sets using the train_test_split function



```
X5_train, X5_test, y5_train, y5_test = train_test_split(X5, y5, test_size=0.3, random_state=42)
```

Fig 1.48

X_train: This will contain a subset of our predictor variables X that we'll use for our Gradient Boosting Regression model, it contains 70% (in this case) of our data because we've specified test_size=0.3, meaning we want to reserve 30% of our data for testing

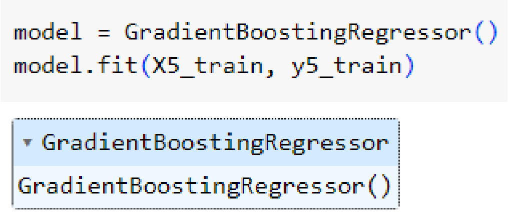
X_test: This will contain the remaining 30% of our predictor variables. We will use this set to evaluate our model's performance on data it has never seen before. This is essential to check how well our model reacts to new and unseen data.

y_train: This contains a subset of our target variable Y (corresponding to X_train). It's used for training our model.

y_test: This contains the remaining portion of our target variable corresponding to X_test. We will use this set to evaluate our model's predictions against the actual target values

The random_state parameter is set to 42 to ensure that the data split is reproducible when running the code multiple times.

Our machine learning model can be trained on X_train and y_train following this data split, and its performance can then be evaluated on X_test and y_test to determine how well it reacts to new data

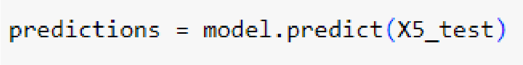


```
model = GradientBoostingRegressor()
model.fit(X5_train, y5_train)
```

```
▾ GradientBoostingRegressor
GradientBoostingRegressor()
```

Fig 1.49

Using our predictor variables X and Y, this code builds a Gradient Boosting Regression model and fits it, enabling the model to discover the link between them



```
predictions = model.predict(X5_test)
```

Fig 1.50

This makes predictions on the testing dataset X_test using the trained Gradient Boosting Regression model, saving the results in the predictions variable.

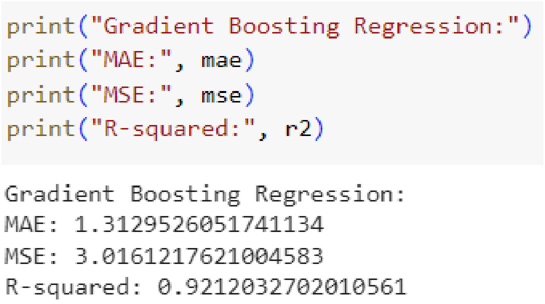
Following these steps, we have to now calculate the metrics which assess the performance and the accuracy of the Gradient Boosting Regression on the test dataset

The three metrics under consideration are MAE, MSE and R SQUARED

MAE-Mean Absolute Error represents the average absolute difference between your model's predictions and the actual values

MSE- Mean Square Error represents the average squared difference between your model's predictions and the actual values

The coefficient of determination (R-squared) is a measure of how well the model fits the data



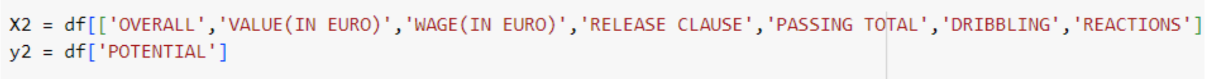
```
print("Gradient Boosting Regression:")
print("MAE:", mae)
print("MSE:", mse)
print("R-squared:", r2)
```

```
Gradient Boosting Regression:
MAE: 1.3129526051741134
MSE: 3.0161217621004583
R-squared: 0.9212032702010561
```

Fig 1.51 Final result of GBR model

RANDOM FOREST REGRESSION

For our Random Forest Regression model the feature and the target variables are



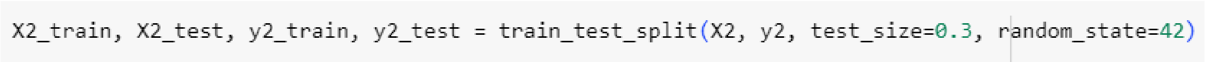
```
X2 = df[['OVERALL','VALUE(IN EURO)','WAGE(IN EURO)','RELEASE CLAUSE','PASSING TOTAL','DRIBBLING','REACTIONS']]
y2 = df['POTENTIAL']
```

Fig 1.52

The feature variables are selected on the basis of the highest correlation scores as based on the heatmap while the target variable is the one that we need to predict

X is the independent variable while Y is dependent variable

After this we need to split the dataset into training and testing sets using the train_test_split function



```
X2_train, X2_test, y2_train, y2_test = train_test_split(X2, y2, test_size=0.3, random_state=42)
```

Fig 1.53

X_train: This will contain a subset of our predictor variables X that we'll use for our Gradient Boosting Regression model, it contains 70% (in this case) of our data because we've specified test_size=0.3, meaning we want to reserve 30% of our data for testing

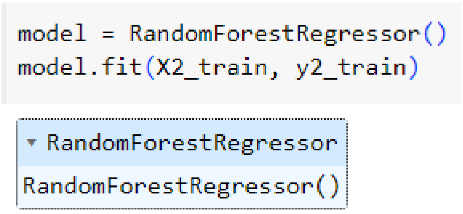
X_test: This will contain the remaining 30% of our predictor variables. We will use this set to evaluate our model's performance on data it has never seen before. This is essential to check how well our model reacts to new and unseen data.

y_train: This contains a subset of our target variable Y (corresponding to X_train). It's used for training our model.

y_test: This contains the remaining portion of our target variable corresponding to X_test. We will use this set to evaluate our model's predictions against the actual target values

The random_state parameter is set to 42 to ensure that the data split is reproducible when running the code multiple times.

Our machine learning model can be trained on X_train and y_train following this data split, and its performance can then be evaluated on X_test and y_test to determine how well it reacts to new data

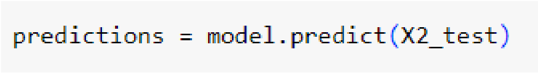


```
model = RandomForestRegressor()
model.fit(X2_train, y2_train)
```

```
▾ RandomForestRegressor
RandomForestRegressor()
```

Fig 1.54

Using our predictor variables X and Y, this code builds a Random Forest Regression model and fits it, enabling the model to discover the link between them



```
predictions = model.predict(X2_test)
```

Fig 1.55

This makes predictions on the testing dataset X_test using the trained Random Forest Regression model, saving the results in the predictions variable.

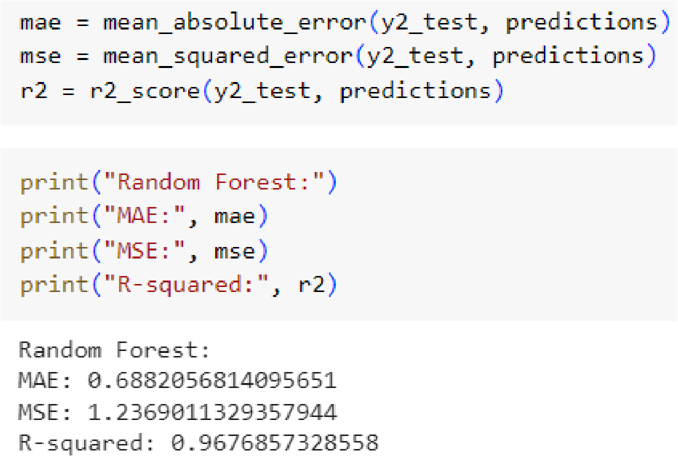
Following these steps, we have to now calculate the metrics which assess the performance and the accuracy of the Random Forest Regression on the test dataset

The three metrics under consideration are MAE, MSE and R SQUARED

MAE-Mean Absolute Error represents the average absolute difference between your model's predictions and the actual values

MSE- Mean Square Error represents the average squared difference between your model's predictions and the actual values

The coefficient of determination (R-squared) is a measure of how well the model fits the data



```
mae = mean_absolute_error(y2_test, predictions)
mse = mean_squared_error(y2_test, predictions)
r2 = r2_score(y2_test, predictions)
```

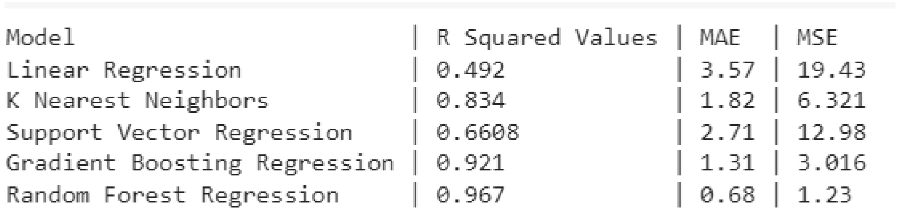
```
print("Random Forest:")
print("MAE:", mae)
print("MSE:", mse)
print("R-squared:", r2)
```

```
Random Forest:
MAE: 0.6882056814095651
MSE: 1.2369011329357944
R-squared: 0.9676857328558
```

Fig 1.56 Final result of RF model

RESULT AND CONCLUSION

Now since we have all the results of all the five algorithms



```
Model                         | R Squared Values | MAE  | MSE
Linear Regression             | 0.492            | 3.57 | 19.43
K Nearest Neighbors           | 0.834            | 1.82 | 6.321
Support Vector Regression     | 0.6608           | 2.71 | 12.98
Gradient Boosting Regression  | 0.921            | 1.31 | 3.016
Random Forest Regression      | 0.967            | 0.68 | 1.23
```

Fig 1.57 Results of all the algorithms

Conditions to judge the results are as follows

A lower MAE indicates that the model is closer to the actual values

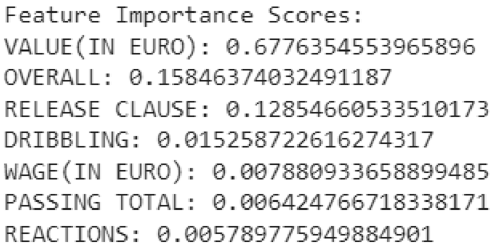
A lower MSE indicates that the model is closer to the actual values

A higher R-squared indicates that the model fits the data better

The Random Forest Regression model creates a number of decisions trees and averages their predictions. The R-squared value for the Random Forest Regression model is 0.967, which indicates that the model fits the data very well and it also has the lowest MAE and MSE, which indicates that it is the most accurate model

Random Forest Regression model is performing exceptionally well, and it's likely capturing meaningful patterns in our data

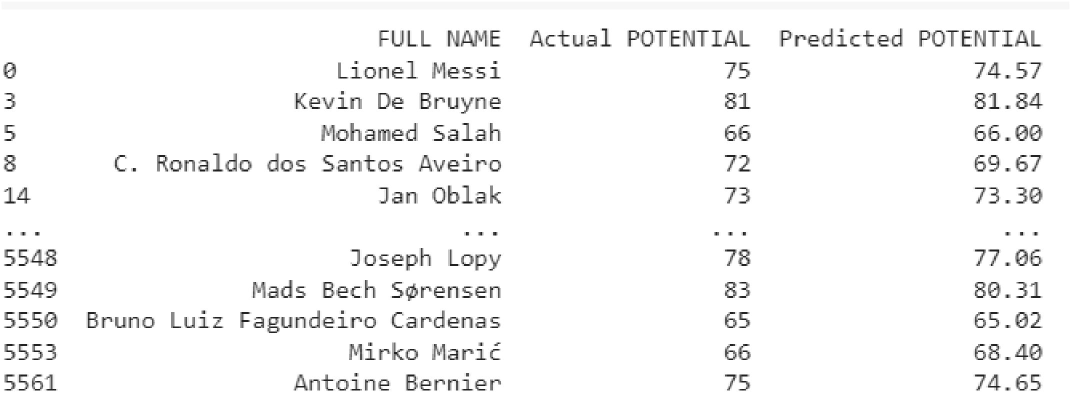
Now since we know Random Forest is the best performing model, we need to know which features contribute the most towards achieving the highest accuracy score



```
Feature Importance Scores:
VALUE(IN EURO): 0.6776354553965896
OVERALL: 0.15846374032491187
RELEASE CLAUSE: 0.12854660533510173
DRIBBLING: 0.015258722616274317
WAGE(IN EURO): 0.007880933658899485
PASSING TOTAL: 0.006424766718338171
REACTIONS: 0.005789775949884901
```

Fig 1.58 List of highest importance scores

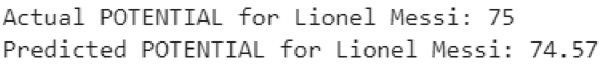
Now, we need to acquire a list that showcases the player's name, along with their actual potential and predicted potential, to assess how well it performs in terms of ratings.



```
                          FULL NAME  Actual POTENTIAL  Predicted POTENTIAL
0                      Lionel Messi                75                74.57
3                   Kevin De Bruyne                81                81.84
5                     Mohamed Salah                66                66.00
8       C. Ronaldo dos Santos Aveiro                72                69.67
14                        Jan Oblak                73                73.30
...                             ...               ...                  ...
5548                   Joseph Lopy                78                77.06
5549              Mads Bech Sørensen                83                80.31
5550  Bruno Luiz Fagundeiro Cardenas                65                65.02
5553                   Mirko Marić                66                68.40
5561               Antoine Bernier                75                74.65
```

Fig 1.59 Actual vs Predicted Values

If our objective is to choose a specific player and conduct an analysis by contrasting their real and forecasted values, we have the option to do that as well



```
Actual POTENTIAL for Lionel Messi: 75
Predicted POTENTIAL for Lionel Messi: 74.57
```

Fig 1.60 Potential comparison for Lionel Messi

The Random Forest prediction model effectively addresses the challenges in the FIFA dataset, providing a data-driven solution for assessing football player potential. With an impressive R-squared score of 0.967, it demonstrates a remarkable ability to capture complex relationships between player attributes and future potential. Its low Mean Absolute Error (MAE) of 0.68 and Mean Squared Error (MSE) of 1.23 underscore its precision in predicting player ratings

This innovation eliminates prior uncertainties and inconsistencies in player ranking, paving the way for data-driven decision-making in sports and gaming. Organizations can confidently rely on this model year after year, ensuring consistent and accurate player ratings. By utilizing up-to-date statistics, this model enables organizations to easily project expected player ratings for the upcoming year, facilitating informed choices in player recruitment and team composition. The Random Forest model's exceptional performance makes it the preferred choice for anticipating football player potential based on FIFA dataset attributes.

FUTURE DEVELOPMENTS

Examine how expected player ratings affect fan engagement, such as whether there are more discussions and community involvement. Examine the impact of virtual gaming on in-person fan interactions.

Study how predictive models can be included into the e-sports environment. Examine how these models may impact player choice, team tactics, and the dynamics of competition in football video game competitions.

Sports organizations can help you build and validate your model in practical settings. For example, work with the football groups. collaborating with these organizations can result in insightful feedback

Player Development Insights: Analyse how the projections of the model can influence player development plans, examine how clubs and academies might use this knowledge to properly develop emerging talents.

REFERENCES

- www.kaggle.com. (n.d.). Fifa 23 Players Dataset. [online] Available at: https://www.kaggle.com/datasets/sanjeetsinghnaik/fifa-23-players-dataset.

- AL-ASADI, M.A. and Tasdemir, S. (2022). Predict the Value of Football Players Using FIFA video game data and Machine Learning Techniques. IEEE Access, pp.1–1. doi:https://doi.org/10.1109/access.2022.3154767.

- Alvin, T.P. (2022). Predicting the Market Value of FIFA Soccer Players with Regression. [online] Medium. Available at: https://towardsdatascience.com/predictingmarket-value-of-fifa-soccer-players-withregression-5d79aed207d9.

- Sibanda, E. (2019). Using a multivariable linear regression model to predict the sprint speed of players in FIFA 19. [online] HackerNoon.com. Available at: https://medium.com/hackernoon/using-a-multivariable-linear-regression-model-to-predict-the-sprint-speed-of-players-in-fifa-19-530618986e1c.

- Advani, V. (2020). What is Machine Learning? How Machine Learning Works and future of it? [online] GreatLearning. Available at: https://www.mygreatlearning.com/blog/what-is-machine-learning/.

- Gupta, S. (2021). Regression vs. Classification in Machine Learning: What's the Difference? [online] Springboard Blog. Available at: https://www.springboard.com/blog/data-science/regression-vs-classification/.

- Hayes, A. (2023). How Multiple Linear Regression Works. [online] Investopedia. Available at: https://www.investopedia.com/terms/m/mlr.asp.

- IBM (n.d.). What is the k-nearest neighbors algorithm? | IBM. [online] www.ibm.com. Available at: https://www.ibm.com/topics/knn.

- JavaTpoint (2021). K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint. [online] www.javatpoint.com. Available at: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning.

- https://www.facebook.com/kdnuggets (2018). Frameworks for Approaching the Machine Learning Process - KDnuggets. [online] KDnuggets. Available at: https://www.kdnuggets.com/2018/05/general-approaches-machine-learning-process.html.