PREDICTING THE VALUE OF FOOTBALL PLAYERS

A main Project Report submitted in the partial fulfillment of the

requirements for the award of the degree

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

Goli Bodhini (19471A05L6)

Under the esteemed guidance of

M.Satyam Reddy, M.Tech.

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

NARASARAOPETA ENGINEERING COLLEGE:

NARASARAOPET (AUTONOMOUS)

(Affiliated to J.N.T.U.Kakinada ,Approved by AICTE&Accredited by NBA)

2022-2023

NARASARAOPETA ENGINEERING COLLEGE: NARASARAOPETA

(AUTONOMOUS)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the main project entitled **"PREDICTING THE VALUE OF FOOTBALL PLAYERS"** is a bonafide work done by " G.Bodhini (19471A05L6) " in partial fulfilment of the requirements for the award of the degree of **BACHELOR OF TECHNOLOGY** in the department of **COMPUTER SCIENCE AND ENGINEERING during 2019-2020.**

PROJECT GUIDE
M.Satyam Reddy,MTech.

PROJECT CO-ORDINATOR
M.Sireesha,MTech.,(Ph.D)

HEAD OF THE DEPARTMENT
Dr.S.N.TirumalaRao,M.Tech.,Ph.D

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

We wish to express our thanks to carious personalities who are responsible for the completion of the project. We are extremely thankful to our beloved chairman sir **M.V. Koteswara Rao**, **B.sc** who took keen interest in us in every effort throughout this course. We owe out gratitude to our principal **Dr.M.Sreenivasa Kumar,** **M.Tech,Ph.D(UK),MISTE,FIE(1)** for his kind attention and valuable guidance throughout the course.

We express our deep felt gratitude to **Dr.S.N.Tirumala Rao,M.Tech, Ph.D H.O.D. CSE** department and our guide **M.Satyam Reddy,M.Tech.** of CSE department whose valuable guidance and unstinting encouragement enable us to accomplish our project successfully in time.

We extend our sincere thanks to **M.Sireesha,M.Tech,(Ph.D)** Associate Professor coordinator of the project for extending her encouragement. Their profound knowledge and willingness have been a constant source of inspiration for us throughout this project work.

We extend our sincere thanks to all other teaching and non-teaching staff to department for their cooperation and encouragement during our B.Tech degree. We have no words to acknowledge the warm affection, constant inspiration and encouragement that we receive from our parents. We affectionately acknowledge the encouragement received from our friends those who involved in giving valuable suggestions had clarifying out all doubts which had really helped us in successfully completing our project.

By

G.Bodhini (19471A05L6)

ABSTRACT

As we all know that football is a very popular and trending game across the globe, and the football players like Christiano Ronaldo,Lionel Messi, Mbappi are become very popular in recent games. We all know their names and Origin of these famous football players and many of us don’t know their net value.In this project we are going to predict every foot ball player using Machine Learning. In Machine learning we are going to use four algorithms or features namely linear regression, multiple linear regression, decision trees and random forests.

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

INSTITUTE VISION AND MISSION

INSTITUTION VISION

To emerge as a Centre of excellence in technical education with a blend of effective student centric teaching learning practices as well as research for the transformation of lives and community,

INSTITUTION MISSION

M1: Provide the best class infra-structure to explore the field of engineering and research

M2: Build a passionate and a determined team of faculty with student centric teaching,imbibing experiential, innovative skills

M3: Imbibe lifelong learning skills, entrepreneurial skills and ethical values in students for addressing societal problems

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VISION OF THE DEPARTMENT

To become a centre of excellence in nurturing the quality Computer Science & Engineering professionals embedded with software knowledge, aptitude for research and ethical values to cater to the needs of industry and society.

MISSION OF THE DEPARTMENT

The department of Computer Science and Engineering is committed to

M1: Mould the students to become Software Professionals, Researchers and Entrepreneurs by providing advanced laboratories.

M2: Impart high quality professional training to get expertize in modern software tools and technologies to cater to the real time requirements of the Industry.

M3: Inculcate team work and lifelong learning among students with a sense of societal and ethical responsibilities.

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Program Specific Outcomes (PSO's)

PSO1: Apply mathematical and scientific skills in numerous areas of Computer Science and Engineering to design and develop software-based systems.

PSO2: Acquaint module knowledge on emerging trends of the modern era in Computer Science and Engineering

PSO3: Promote novel applications that meet the needs of entrepreneur, environmental and social issues.

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Program Educational Objectives (PEO's)

The graduates of the programme are able to:

PEO1: Apply the knowledge of Mathematics, Science and Engineering fundamentals to identify and solve Computer Science and Engineering problems.

PEO2: Use various software tools and technologies to solve problems related to academia, industry and society.

PEO3: Work with ethical and moral values in the multi-disciplinary teams and can communicate effectively among team members with continuous learning.

PEO4: Pursue higher studies and develop their career in software industry.

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Program Outcomes

1. Engineering knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. Problem analysis: Identify, formulate, research literature, and analyzecomplex engineering problems reaching substantiated conclusions using firstprinciples of mathematics, natural sciences, and engineering sciences.

3. Design/development of solutions: Design solutions for complexengineering problems and design system components or processes that meetthe specified needs with appropriate consideration for the public health andsafety, and the cultural, societal, and environmental considerations.

4. Conduct investigations of complex problems: Use research-basedknowledge and research methods including design of experiments, analysis andinterpretation of data, and synthesis of the information to provide validconclusions.

5. **Modern tool usage:** Create, select, and apply appropriate techniques,resources, and modern engineering and IT tools including prediction andmodeling to complex engineering activities with an understanding of thelimitations.

5. **The engineer and society:** Apply reasoning informed by the contextualknowledge to assess societal, health, safety, legal and cultural issues and theconsequent responsibilities relevant to the professional engineering practice.

7. Environment and sustainability: Understand the impact of theprofessional engineering solutions in societal and environmental contexts,and demonstrate the knowledge of, and need for sustainable development.

8. Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. Individual and team work: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

10. Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

11. Project management and finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

12. Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Project Course Outcomes (CO'S):

CO425.1: Analyse the System of Examinations and identify the problem.

CO425.2: Identify and classify the requirements.

CO425.3: Review the Related Literature

CO425.4: Design and Modularize the project

CO425.5: Construct, Integrate, Test and Implement the Project.

CO425.6: Prepare the project Documentation and present the Report using appropriate method.

Course Outcomes – Program Outcomes mapping

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
C425.1		✓											✓		
C425.2	✓		✓		✓								✓		
C425.3				✓		✓	✓	✓					✓		
C425.4			✓			✓	✓	✓					✓	✓	
C425.5					✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
C425.6									✓	✓	✓		✓	✓	

Course Outcomes – Program Outcome correlation

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
C425.1	2	3											2		
C425.2			2		3								2		
C425.3				2		2	3	3					2		
C425.4			2			1	1	2					3	2	
C425.5					3	3	3	2	3	2	2	1	3	2	1
C425.6									3	2	1		2	3	

Note: The values in the above table represent the level of correlation between CO's and PO's:

1. **Low level**

2. **Medium level**

3. **High level**

Project mapping with various courses of Curriculum with Attained PO's:

Name of the course from which principles are applied in this project	**Description of the device**	**Attained PO**
C3.2.4, C3.2.5	Gathering the requirements and defining the problem, plan to develop a **smart bottle for health care using sensors.**	PO1, PO3
CC4.2.5	Each and every requirement is critically analyzed, the process model is identified and divided into **five modules**	PO2, PO3
CC4.2.5	Logical design is done by using the unified modelling language which involves individual team work	PO3, PO5, PO9
CC4.2.5	Each and every module is tested, integrated, and evaluated in our project	PO1, PO5
CC4.2.5	Documentation is done by all our four members in the form of a group	PO10
CC4.2.5	Each and every phase of the work in group is presented periodically	PO10, PO11
CC4.2.5	Implementation is done and the project will be handled by the **hospital management and in future updates in our project can be done based on air bubbles occurring in liquid in saline.**	PO4, PO7
CC4.2.8 CC4.2.	**The physical design includes hardware components like sensors, gsm module, software and Arduino.**	PO5, PO6

INDEX

S.NO	CONTENTS	PAGE NO
	3.7 Cross Validation	**22**
	3.8 Machine Learning Algorithms	22
	3.8.1 Decision Tree	22
	3.8.2 Random Forest	22
	3.8.3 Regression	23
	3.8.4 Linear Regression	23
	3.9 Implementing Code	24
	3.10 Result Analysis	37
4.	**Output Screens**	**40**
5.	**Conclusion**	**44**
6.	**Future Scope**	**45**
7.	**Bibliography**	**46**
8.	**Conference Paper**	**48**
9.	**Similarity Check Report**	**54**

List of figures

1. INTRODUCTION

1.1 Introduction

Football is a popular sport; however, it is a big business as well. As we all know that football is a very trending game across the globe, and the football players like Christiano Ronaldo,Lionel Messi, Mbappi are become very popular in recent games.Market values can be understood as estimates of transfer fee prices that could be paid for a player on the football market. Therefore, market values play an important role in transfer negotiations. The market has traditionally been estimated by football experts. However, expert judgments are inaccurate and not transparent.

We all know their names and Origin of these famous football players and many of us don't know their net value.In this project we are going to predict every foot ball player using Machine Learning.Data analytics may thus provide a sound alternative or a complementary approach to experts-based estimations of market value.The method is based on the application of machine learning algorithms to the performance data of football players. The data used in the experiment are FIFA 20 video game data, collected from sofifa.com.

We estimate players' market values using four regression models that were tested on the full set of features linear regression, multiple linear regression, decision factors affecting the determination of the market value. In the experimental results, random forest performed better than other algorithms for predicting the players' market values.

The results show that our methods are capable to address this task efficiently, surpassing the performance reported in previous works. Finally, we believe our results can play an important role in the negotiations that take place between football clubs and a player's agents.

1.2 Existing System

Nowadays, decisions are made by agents based on their experience and knowledge.This practice may lead to errors, consume a lot of time and excessive costs which affects the value of football players.

Disadvantages:

1. Doesn’t generate accurate and efficient results.
2. Computation time is very high.
3. Lacking of accuracy may result in lack of value of football players.

1.3 Proposed System

By using these algorithms we can reduce the complexity in predicting the value the football players, reduce errors, enhance accurate results. It is easier to predict the value and it will also help the club agents to makequick decisions.

Advantages:

1. Generates accurate and efficient results.
2. Computation time is greatly reduced.
3. Reduces manual work.
4. Efficient and accurate results.

1.4 System Requirements

1.4.1 Hardware Requirements:

- Processor : Intel®core™ i7-7500UCPU@2.70gh
- RAM : 8 GB
- System Type : 64-bit operating system, x64-based processor

1.4.2 Software Requirements:

- Operating system : Windows 10
- Coding language : Python
- Platform : Google COLAB
- Browser : Any Latest Browser like Chrome

2.LITERATURE SURVEY

2.1 Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

The most common indicators for assessing the market value fall into three categories:player characteristics, player performance and player popularity.

2.1.1 Player Characteristics:

Player characteristics are described as both physical and demographic attributes.Age is a important indicator of market value, as it reflects both experience and ability.Most studies used the age factor to estimate market value, bearing in mind that players' values usually increase until their mid-20s and decrease thereafter. Besides, it has been found that player height leads to a significant increase in salary returns. because it indicates good header ability that may increase the likelihood of scoring or preventing a goal. Another characteristic that has been studied in player-valuation research is footedness.

The researchers also studied whether the players nationalities affected their market values. For example, in their study of the Spanish professional football league, they found that non Spanish European players were systematically overrated, while non- European players were systematically underrated. Finally, the player position--goalkeeper, defender, midfielder or forward player---is important in estimating market value.

2.1.2 Player Performance:

Several player performance metrics can be used to estimate market values. Goals, including field goals, headers and penalties, refer to players' ability to score and so are a largely unambiguous measure of performance. Apart from the above mentioned metric, many researchers used other performance metrics that helped explain the value and the fees. Passing are used frequently duelling (or tackles) in the form of clearances; dribbles committed fouls and yellow and red cards.

2.1.3 Player Popularity:

In football, not only is the talent of the player crucial in determining the market value. The ***popularity*** also can explain the demand for football players. In other words, the market value of football players also depends on their crowdpulling power, independent of what they show on the pitch. The image of a player outside the football pitch influences the number of jerseys sold and money earned from portrait rights. Accordingly, studies of the football transfer market have investigated popularity-related factors. Popular athletes have commercial value, which is important for the club. Even though players like Messi, Ronaldo or even Ibrahimovic are close to retirement, their brand value is still very high as they have gained international stature during their careers. Everyone knows their face, and this gives them extra ammunition when negotiating sponsor deals with popular brands.

Finally,by using the machine learning algorithms like Linear Regression,Multiple Regression, Decision Tree and Random Forest, we will find the market value of football players.

2.2 Some machine learning methods:

Machine learning algorithms are often categorized as supervised and unsupervised.

Supervised machine learning algorithms can apply what has been learned in thepast to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

Unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behaviour within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best. This is known as the reinforcement signal.

2.3 Applications of machine learning

1. Virtual Personal Assistants
2. Predictions while Commuting
3. Videos Surveillance
4. Social Media Services
5. Email Spam and Malware Filtering
6. Online Customer Support
7. Search Engine Result Refining
8. Product Recommendations
9. Online Fraud Detection

3. SYSTEM ANALYSIS

3.1 Implementation of machine learning using Python

Python is a popular programming language. It was created in 1991 by Guido van Rossum.

It is used for:

1.web development (server-side),

2.software development,

3.mathematics,

4.system scripting.

The most recent major version of Python is Python 3. However, Python 2, although not being updated with anything other than security updates, is still quite popular.It is possible to write Python in an Integrated Development Environment, such as Thonny, Pycharm, Netbeans or Eclipse, Anaconda which are particularly useful when managing larger collections of Python files.

Python was designed for its readability. Python uses new lines to complete a command, as opposed to other programming languages which often use semicolons or parentheses.Python relies on indentation, using whitespace, to define scope; such as the scope of loops, functions and classes. Other programming languages often use curly-brackets for this purpose.

In the older days, people used to perform Machine Learning tasks manually by coding all the algorithms and mathematical and statistical formula. This made the process time consuming, tedious and inefficient. But in the modern days, it is become very much easy and efficient compared to the olden days by various python libraries, frameworks,

and modules. Today, Python is one of the most popular programming languages for this task and it has replaced many languages in the industry, one of the reason is its vast collection of libraries. Python libraries that used in Machine Learning are:

1.Numpy

2.Scipy

3.Scikit-learn

4.Pandas

5.Matplotlib

NumPy is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is very useful for fundamental scientific computations in Machine Learning. It is particularly useful for linear algebra, Fourier transform, and random number capabilities. High-end libraries like TensorFlow uses NumPy internally for manipulation of Tensors.

SciPy is a very popular library among Machine Learning enthusiasts as it contains different modules for optimization, linear algebra, integration and statistics. There is a difference between the SciPy library and the SciPy stack. The SciPy is one of the core packages that make up the SciPy stack. SciPy is also very useful for image manipulation.

Skit-learn is one of the most popular Machine Learning libraries for classical Machine Learning algorithms. It is built on top of two basic Python libraries, NumPy and SciPy. Scikit-learn supports most of the supervised and unsupervised learning algorithms. Scikit learn can also be used for data-mining and data-analysis, which makes it a great tool who is starting out with Machine Learning.

Pandas is a popular Python library for data analysis. It is not directly related to Machine Learning. As we know that the dataset must be prepared before training. In this case, Pandas comes handy as it was developed specifically for data extraction and preparation. It provides high-level data structures and wide variety tools for data analysis. It provides many inbuilt methods for groping, combining and filtering data.

Matpoltlib is a very popular Python library for data visualization. Like Pandas, it is not directly related to Machine Learning. It particularly comes in handy when a programmer wants to visualize the patterns in the data. It is a 2D plotting library used for creating 2D graphs and plots. A module named pyplot makes it easy for programmers for plotting as it provides features to control line styles, font properties, formatting axes, etc. It provides various kinds of graphs and plots for data visualization, histogram, error charts, bar chats, etc.

3.2 Scope of the project

The scope of this system is to maintain player details in datasets, train the model using the large quantity of data present in datasets and predict the value of the football players on new data during testing.

3.3 Analysis

The dataset contains 11 attributes which are used to predict the value of football players such as:

1. International reputation
2. Overall
3. Potential
4. Mentality_composure
5. Age
6. Height_cm
7. Weight_kg
8. Shooting
9. Passing
10. Dribbling
11. Value_eur

International_reputation:

- The popularity also can explain the demand for football player.
- In other words, the market value of football players also depends on their crowdpulling power, independent of what they show on the pitch.
- The normal range is between 1.00-5.00

Overall:

- The Overall Rating of a player will indicate the overall performance of the player in all matches.
- The Overall Rating is in between 48 to 94.

Potential:

- The potential of a player will represent the capability of the player.
- The Potential is in between 49 to 95.

Mentality_Composure:

- It indicates the player's mentality in terms of confident and being calm and control.
- The interval is in range of 12 to 96.

Age:

- It is the one of the main characteristic of a player that is taken into consideration while predicting the value of a football player.
- The age should be in the range of 16 to 42 years.

Height_cm:

- Height plays one of the most important role.If the height of a player is more,he has a

chance of catching more goals.

- The height of a player probably in the range of 156 to 205.

Weight_Kg:

- Weight is the most significant characteristic in predicting the value of a football player.
- The weight of the player should be in the range of 50 to 110kgs.

Shooting:

- It represents the hitting of a ball in an attempt to score the goal.
- It ranges between 15 to 93.

Passing:

- It represents the passing of a ball intentionally from one player to another player in the same team.
- It ranges between 24 to 92.

Dribbling:

- It represents the passing of a ball in a given direction and avoiding the defender's attempts to intercept the ball.
- It is in between 23 to 96.

Value_eur:

- Based on the above characteristics, we will determine the value of a football players.
- The values are exponential.

3.4 DataSet:

age	height_cm	weight_kg	overall	potential	internation	mentality_	shooting	passing	dribbling	value_eur
32	170	72	94	94	5	96	92	92	96	95500000
34	187	83	93	93	5	95	93	82	89	58500000
27	175	68	92	92	5	94	85	87	95	105500000
26	188	87	91	93	3	68	83	82	83	77500000
28	175	74	91	91	4	91	83	86	94	90000000
28	181	70	91	91	4	91	86	92	86	90000000
27	187	85	90	93	3	70	92	84	67	67500000
27	193	92	90	91	3	89	60	70	71	78000000
33	172	66	90	90	4	92	76	89	89	45000000
27	175	71	90	90	3	91	86	81	89	80500000
20	178	73	89	95	3	84	84	78	90	93500000
28	187	89	89	91	3	82	28	54	67	67500000
25	188	89	89	91	3	91	91	79	81	83000000
26	191	91	89	91	3	65	84	70	81	58000000
28	192	82	89	90	4	68	46	86	83	56000000
28	168	72	89	90	3	85	65	92	81	66000000
34	187	85	89	89	4	84	46	58	60	24500000
31	173	70	89	89	4	90	90	77	88	60000000
33	184	82	89	89	4	84	68	75	73	31500000
32	182	86	89	89	5	85	89	80	84	53000000
30	184	80	89	89	4	86	87	74	85	64500000
30	189	76	89	89	4	93	62	80	80	55000000
28	176	73	89	89	4	89	86	84	89	69000000

Fig.1 Dataset

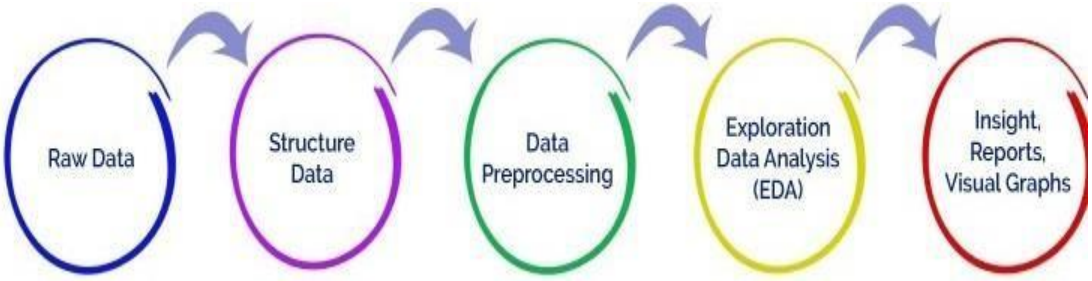
Fig.1 is the data set of football players contains attributes International reputation,Overall, Potential,Mentality_composure, Age, Height_cm, Weight_kg,Shooting, Passing, Dribbling, Value_eur.

3.5 Data Pre-processing

Before feeding data to an algorithm we have to apply transformations to our data which is referred as pre-processing. By performing pre-processing the raw data which is not feasible for analysis is converted into clean data. In-order to achieve better results using a model in Machine Learning, data format has to be in a proper manner. The data should be in a particular format for different algorithms. For example, if we consider Random Forest algorithm it does not support null values. So that those null values have to be managed using raw data.

Data Pre-processing:

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Pre-processing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.



Need of Data Preprocessing:

For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format. For example, Random Forest algorithm does not support null values, therefore to execute random forest algorithm null values have to be managed from the original raw data set. Another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in one data set, and best out of them is chosen.

3.6 Missing values

Filling missing values is one of the pre-processing techniques. The missing values in the dataset is represented as '?' but it a non-standard missing value and it has to be converted into a standard missing value NaN. So that pandas can detect the missing values. We have filled that missing values with 0.

3.6.1 Correlation coefficient method

We can find dependency between two attributes p and q using Correlation coefficient method using the formula.

$$r_{p,q} = \sum(p_i - p)(q_i - q)/n\sigma_p\sigma_q$$

$$= \sum(p_i q_i) - np\,q / n\sigma_p\sigma_q$$

n is the total number of patterns, pi and qi are respective values of p and q attributes in patterns i, p and q are respective mean values of p and q attributes, σ_p , σ_q are respective standard deviations values of p and q attributes. Generally, $-1 \leq r_{p,q} \leq +1$. If $r_{p,q} < 0$, then p and q are negatively correlated. If $r_{p,q} = 0$, then p and q are independent attributes and there is no correlation between them. If $r_{p,q} > 0$, then p and q are positively correlated. Wecan drop the attributes that are having correlation coefficient value as 0 as it indicates that the variables are independent with respect to the prediction attribute. Fig:3.8.2 is the correlation heat map. After applying correlation the attributes are International reputation,Overall, Potential,Mentality_composure, Age, Height_cm, Weight_kg,Shooting, Passing, Dribbling, Value_eur.

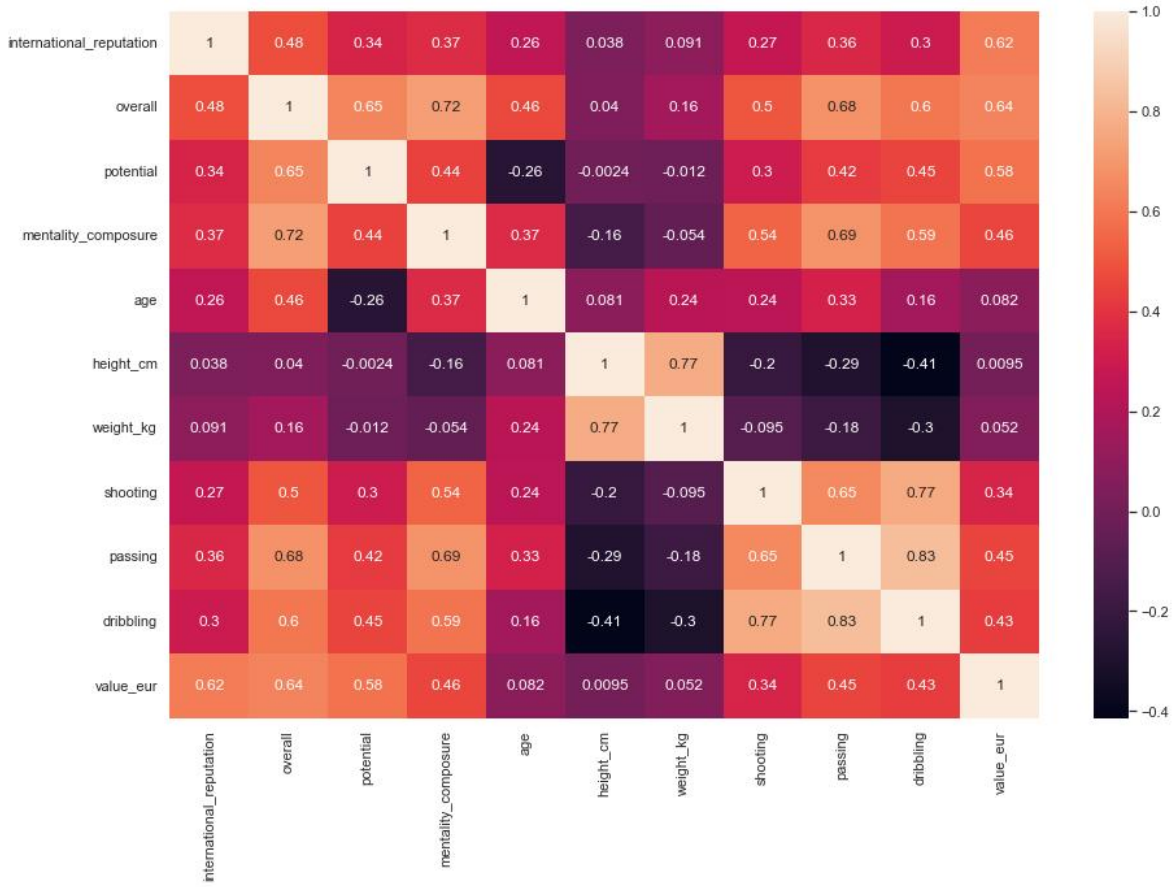


Fig:2 Correlation

Let us see the distribution among all the players.

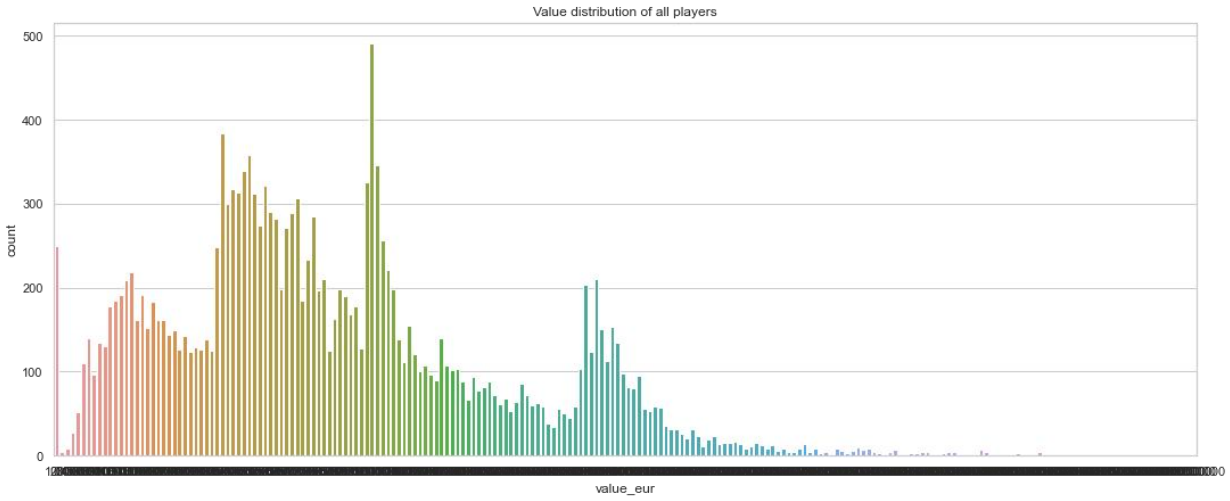


Fig: 3 Value distribution of all players

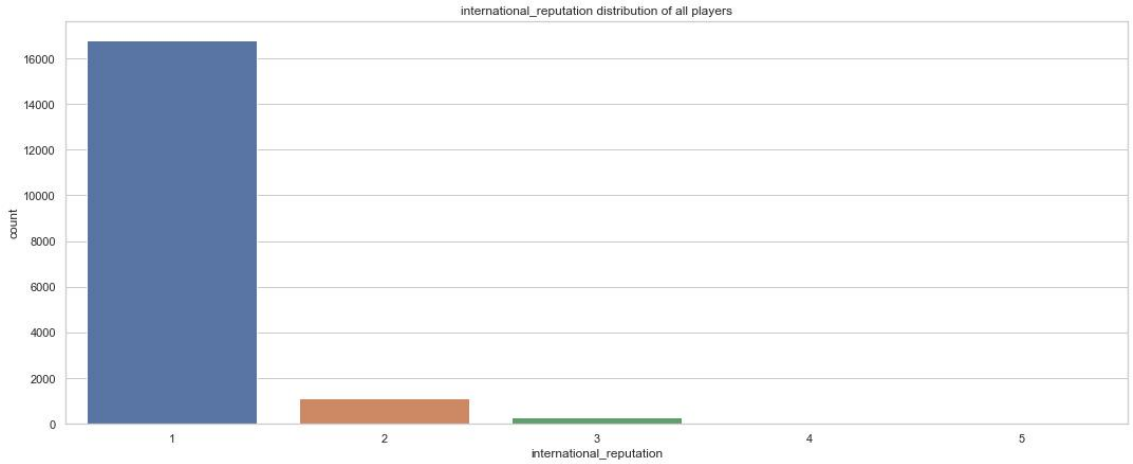


Fig:4 International distribution of all players

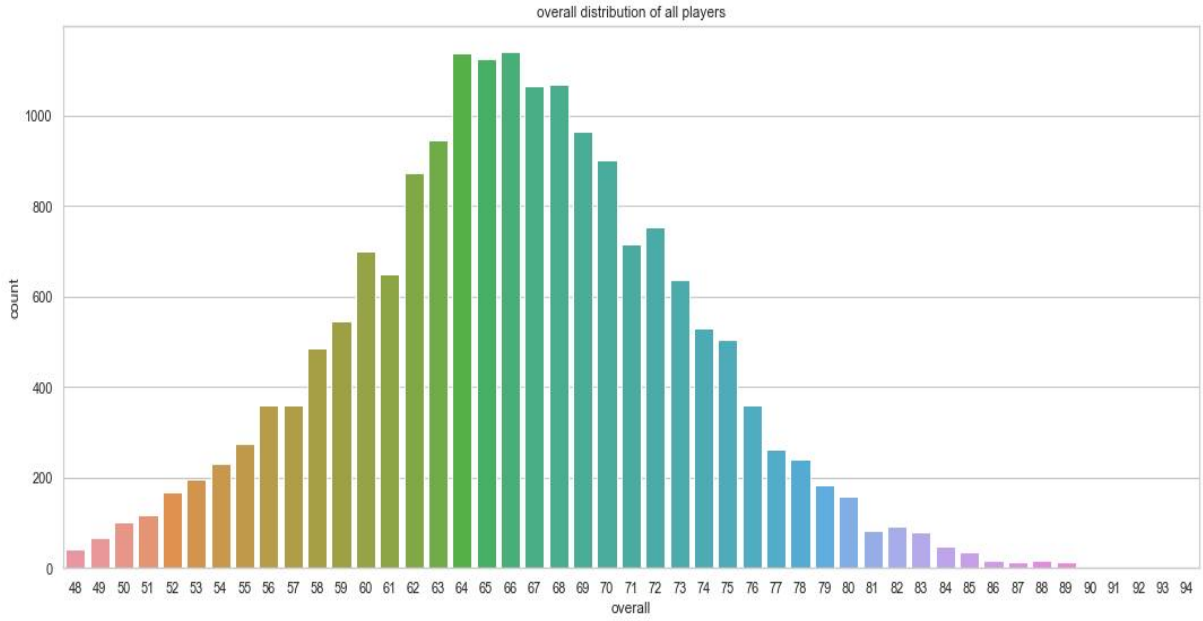


Fig:5 Overall Rating distribution of all players

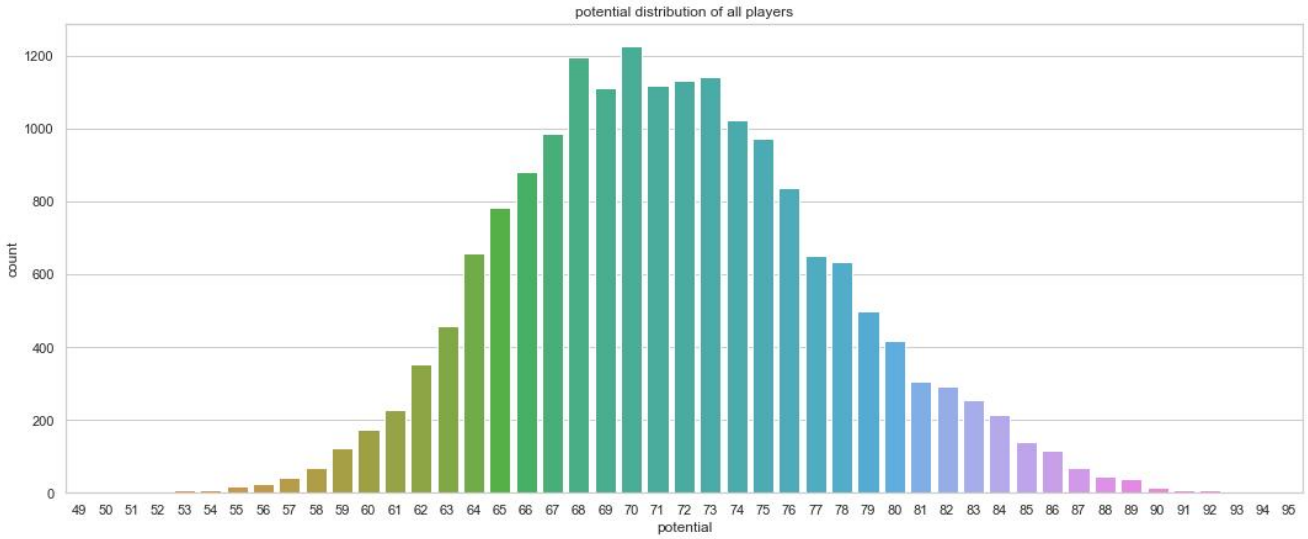


Fig:6 Potential distribution of all players

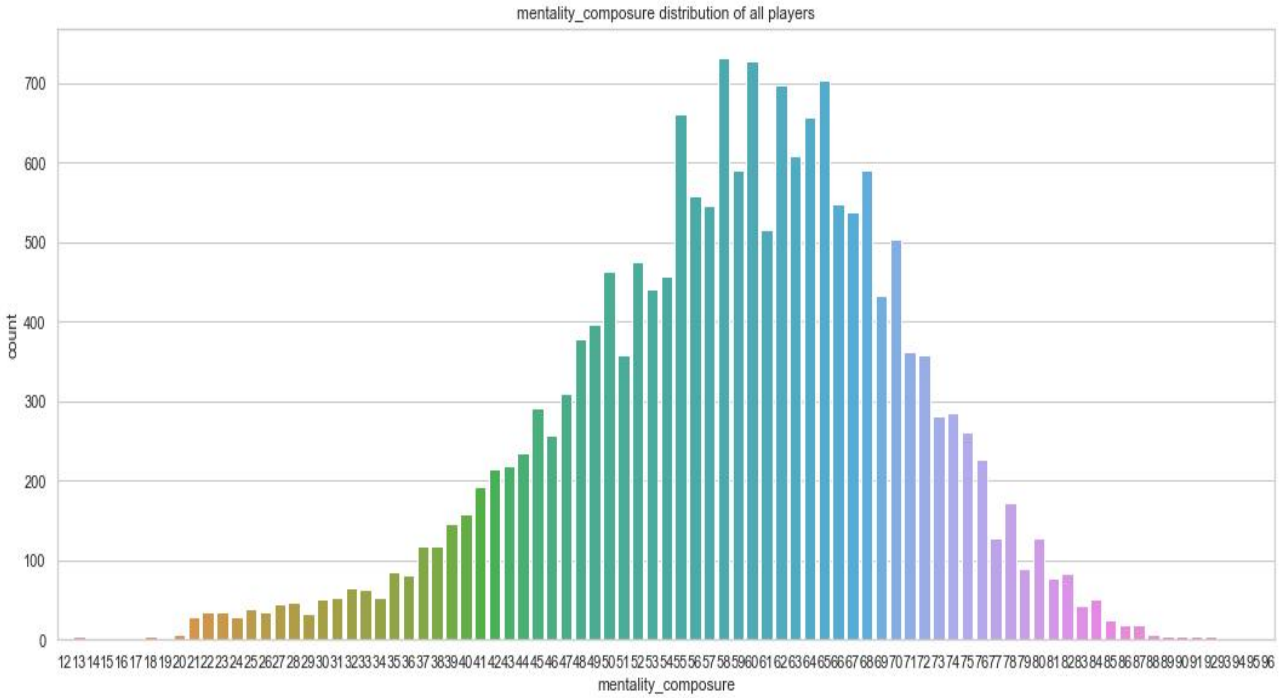


Fig:7 mentality_composure distribution of all players

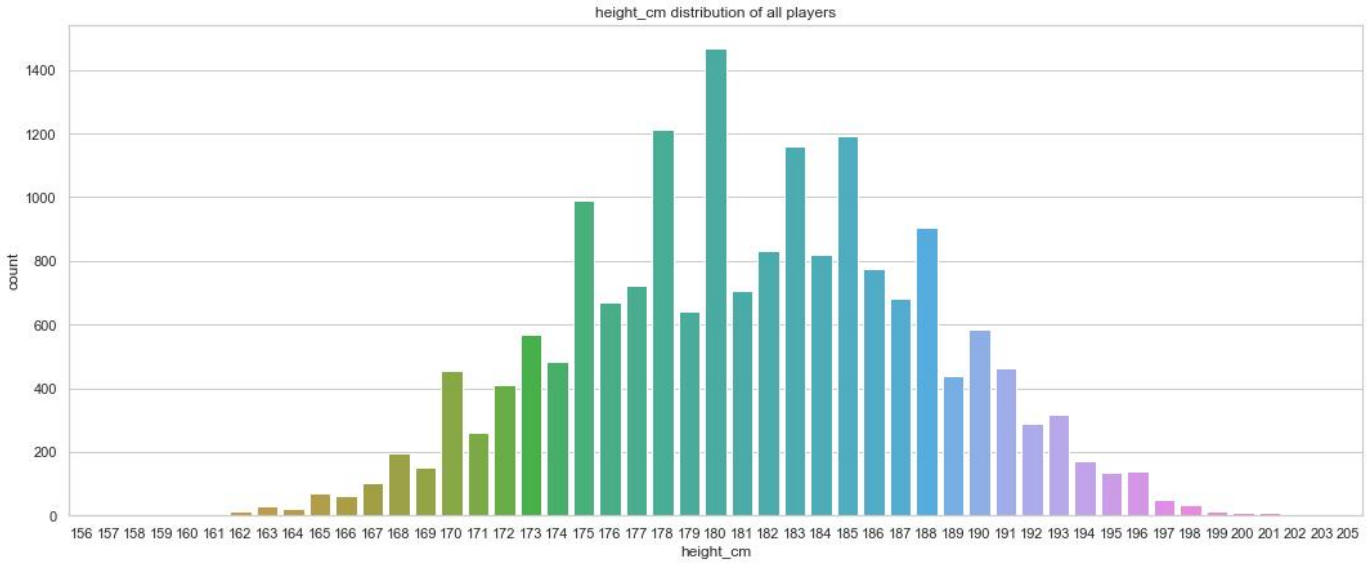


Fig:8 height_cm distribution of all players

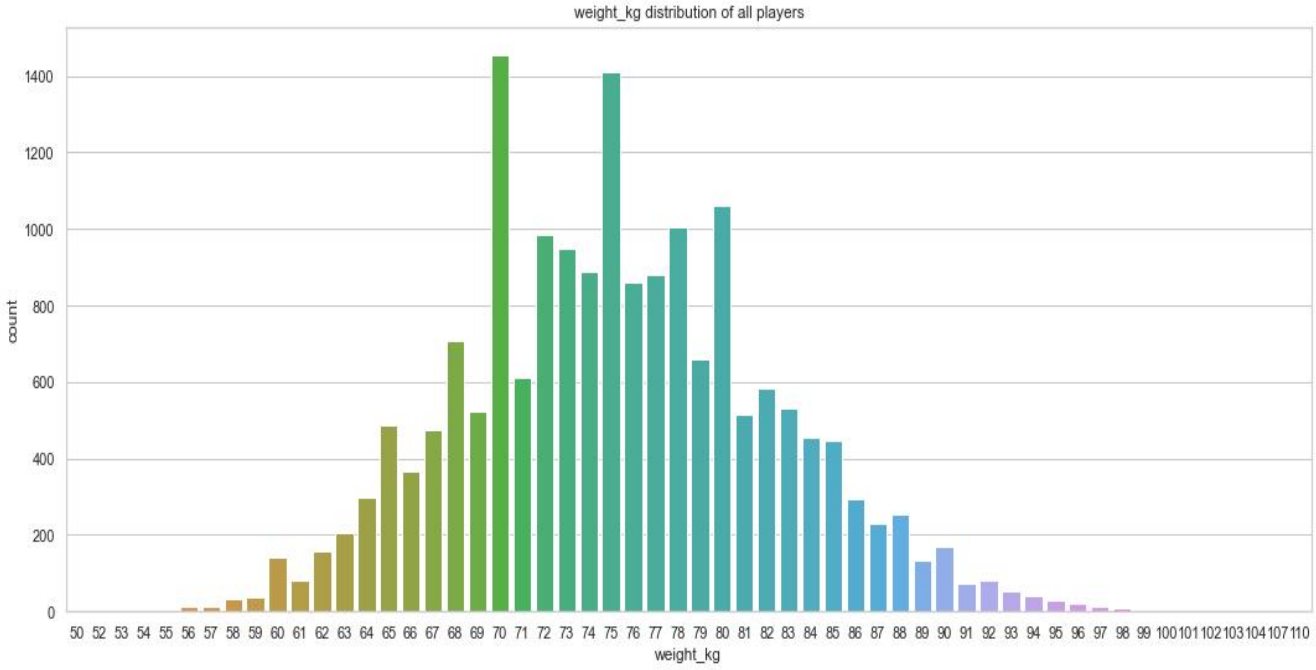


Fig:9 Weight_kg distribution of all players

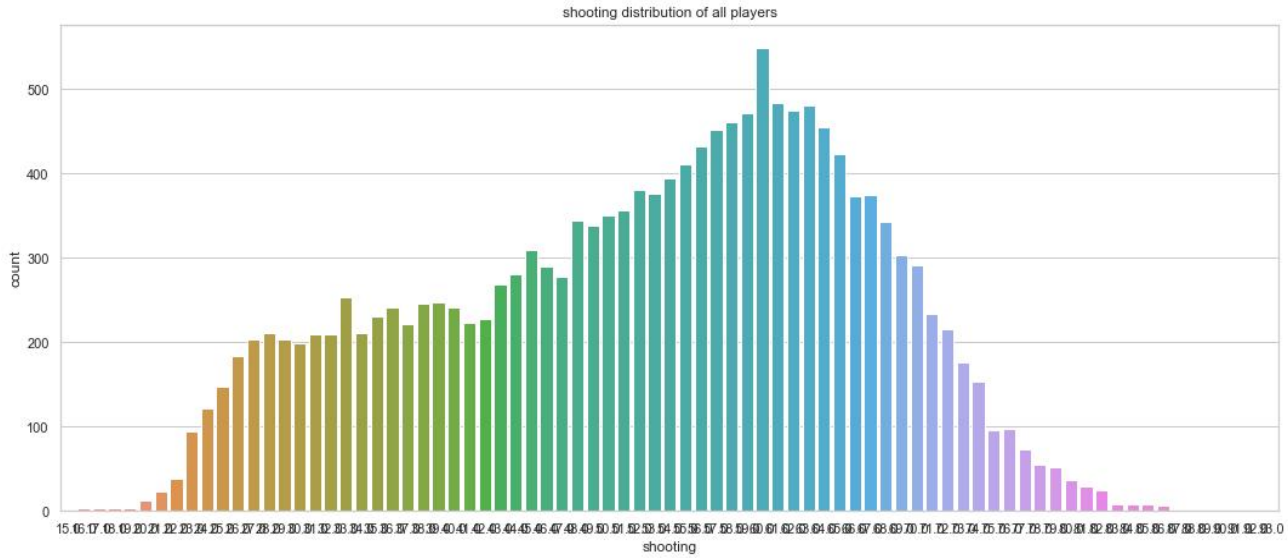


Fig:10 Shooting distribution of all players

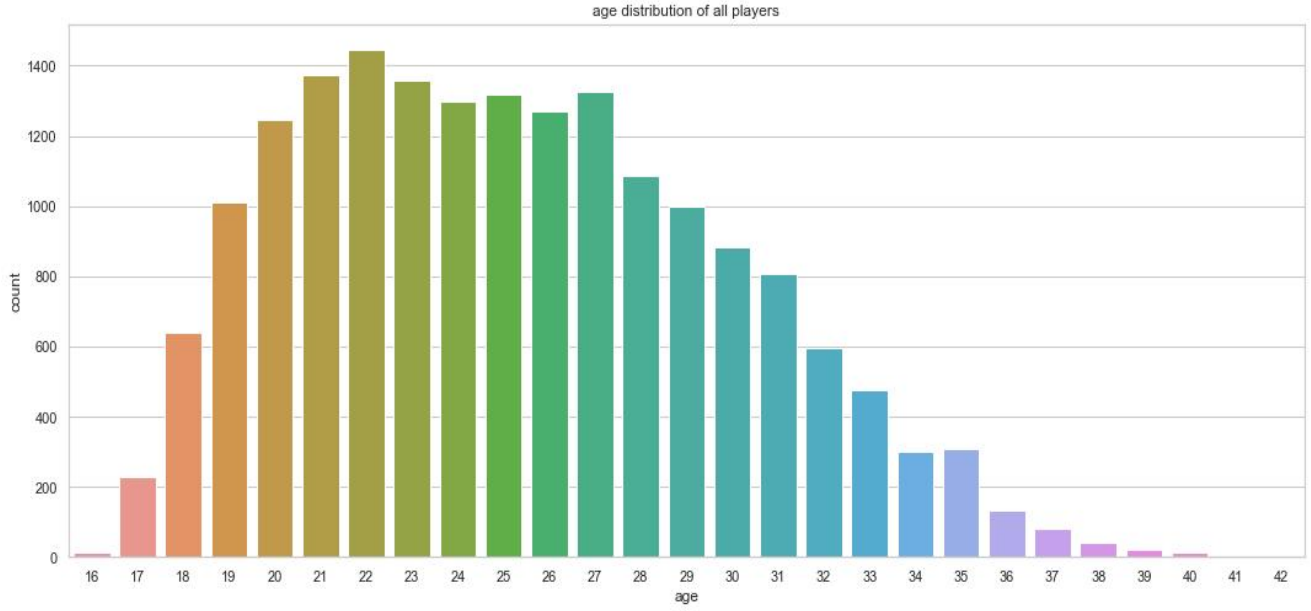


Fig:11 age distribution of all players

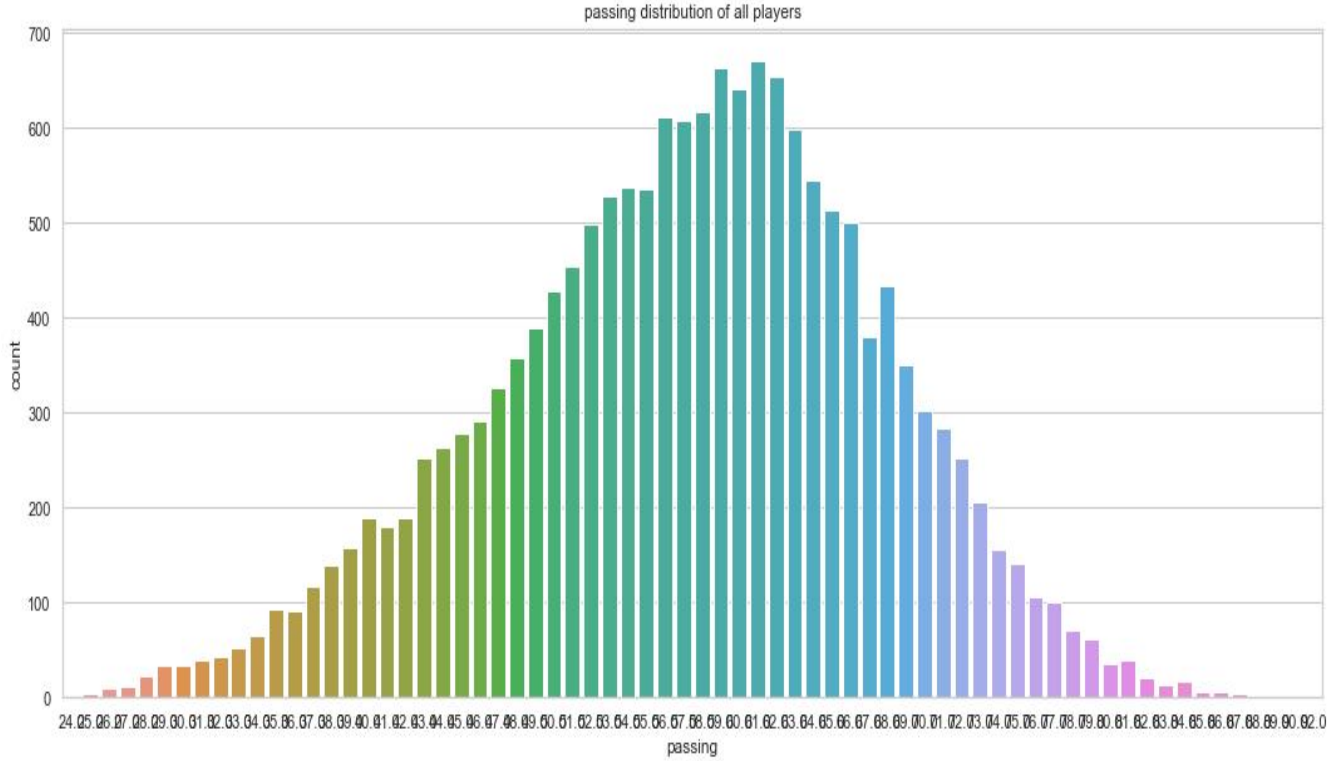


Fig:12 Passing distribution of all players

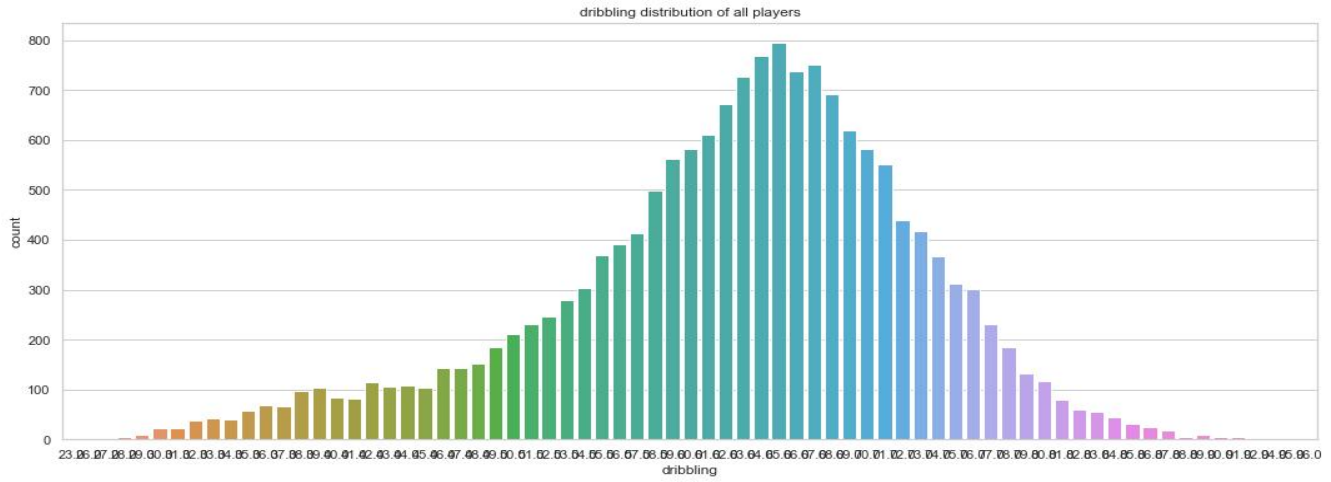


Fig: 13 dribbling distribution of all players

3.7 Cross Validation:

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set. The three steps involved in cross-validation are as follows :

– Reserve some portion of sample data-set.

– Using the rest data-set train the model.

– Test the model using the reserve portion of the data-set.

3.8 Machine learning algorithms

Research on data mining has led to the formulation of several data mining algorithms. These algorithms can be directly used on a dataset for creating some models or to draw vital conclusions and inferences from that dataset. Some popular data mining algorithms are Random forest, Decision tree, Linear Regression..

3.8.1 Decision Tree:

Decision Tree Analysis is a general, predictive modelling tool that has applications spanning a number of different areas. In general, decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The decision rules are generally in form of if-then-else statements. The deeper the tree, the more complex the rules and fitter the model.

3.8.2 Random Forest:

Random Forest is a popular machine learning algorithm that belongs to the supervised

learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

3.8.3 Regression:

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as **temperature, age, salary, price,** etc.

3.8.4 Linear Regression:

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

3.9 Implementation code:

Importing Libraries

```
import pandas as pd
import numpy as np

# Data visualization
import matplotlib.pyplot as plt
import seaborn as sb
from pandas.plotting import scatter_matrix

# Machine Learning Algorithms
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor


# Model Selection and Evaluation
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV

# Performance
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
from sklearn.metrics import mean_absolute_error

# For Missing Values
from sklearn.impute import SimpleImputer
```

Loading the data set

```
fifa_raw_dataset = pd.read_csv("C:/Users/Hello/Documents/players_20.csv")
```

```
fifa_raw_dataset .head()
```

Exploring the dataset

```
fifa_raw_dataset.info()

fifa_raw_dataset.shape

fifa_dataset.shape

fifa_dataset.describe()
```

Visualizing the data

```
 import matplotlib.pyplot as plt
import seaborn as sb
plt.figure(1, figsize=(18, 7))
sb.set(style="whitegrid")
sb.countplot( x= 'value_eur', data=fifa_dataset)
plt.title('Value distribution of all players')
plt.show()

import matplotlib.pyplot as plt
import seaborn as sb
plt.figure(1, figsize=(18, 7))
sb.set(style="whitegrid")
sb.countplot( x= 'international_reputation', data=fifa_dataset)
plt.title('international_reputation distribution of all players')
plt.show()

import matplotlib.pyplot as plt
import seaborn as sb
plt.figure(1, figsize=(18, 7))
sb.set(style="whitegrid")
sb.countplot( x= 'overall', data=fifa_dataset)
plt.title('overall distribution of all players')
plt.show()
```

```
import matplotlib.pyplot as plt
import seaborn as sb
plt.figure(1, figsize=(18, 7))
sb.set(style="whitegrid")
sb.countplot( x= 'potential', data=fifa_dataset)
plt.title('potential distribution of all players')
plt.show()

import matplotlib.pyplot as plt
import seaborn as sb
plt.figure(1, figsize=(18, 7))
sb.set(style="whitegrid")
sb.countplot( x= 'mentality_composure', data=fifa_dataset)
plt.title('mentality_composure distribution of all players')
plt.show()

import matplotlib.pyplot as plt
import seaborn as sb
plt.figure(1, figsize=(18, 7))
sb.set(style="whitegrid")
sb.countplot( x= 'height_cm', data=fifa_dataset)
plt.title('height_cm distribution of all players')
plt.show()

import matplotlib.pyplot as plt
import seaborn as sb
plt.figure(1, figsize=(18, 7))
sb.set(style="whitegrid")
sb.countplot( x= 'weight_kg', data=fifa_dataset)
plt.title('weight_kg distribution of all players')
plt.show()
```

```
import matplotlib.pyplot as plt
import seaborn as sb
plt.figure(1, figsize=(18, 7))
sb.set(style="whitegrid")
sb.countplot( x= 'shooting', data=fifa_dataset)
plt.title('shooting distribution of all players')
plt.show()

import matplotlib.pyplot as plt
import seaborn as sb
plt.figure(1, figsize=(18, 7))
sb.set(style="whitegrid")
sb.countplot( x= 'age', data=fifa_dataset)
plt.title('age distribution of all players')
plt.show()

import matplotlib.pyplot as plt
import seaborn as sb
plt.figure(1, figsize=(18, 7))
sb.set(style="whitegrid")
sb.countplot( x= 'passing', data=fifa_dataset)
plt.title('passing distribution of all players')
plt.show()

import matplotlib.pyplot as plt
import seaborn as sb
plt.figure(1, figsize=(18, 7))
sb.set(style="whitegrid")
sb.countplot( x= 'dribbling', data=fifa_dataset)
plt.title('dribbling distribution of all players')
plt.show()
```

Finding coorelations

```
corr_matrix = fifa_dataset.corr()
corr_matrix.shape
corr_matrix["value_eur"].sort_values(ascending=False)
plt.figure(figsize=(15,10))
sb.heatmap(fifa_dataset.corr(), annot=True, cbar=True)
from pandas.plotting import scatter_matrix
attributes = ["value_eur", "international_reputation", "overall",
            "potential",
"mentality_composure","age","height_cm","weight_kg","shooting","passing","dribbling"]
scatter_matrix(fifa_dataset[attributes], figsize=(12, 8))
plt.show()
```

Data Cleaning:

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
#l = list(train_set['value_eur'] == 0)
#print('Zeros in output label: ',len([v for v in l if v==True] ))
print('\nNaN values in following features:')
fifa_dataset.isnull().any()


import numpy as np
fifa_dataset = fifa_dataset.replace(0,np.nan)
fifa_dataset.head()
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(strategy="median")
imputer.fit(fifa_dataset)

imputer.statistics_

tf = imputer.transform(fifa_dataset)

fifa_dataset_tf = pd.DataFrame(tf, columns=fifa_dataset.columns)
```

```
fifa_dataset_tf.head()

fifa_dataset_tf.isnull().any()
```

Creating the test set

```
import numpy as np
np.random.seed(42)
from sklearn.model_selection import train_test_split
train_set, test_set = train_test_split(fifa_dataset_tf, test_size=0.2, random_state=42)
print('Train',' ','Test')
print(len(train_set),'+',len(test_set),'=',len(train_set)+len(test_set))
```

Separate the Features and Label

```
fifa_dataset_features = fifa_dataset_tf.drop("value_eur", axis=1)
 fifa_dataset_labels = fifa_dataset_tf["value_eur"].copy()

X_test=test_set.drop("value_eur", axis=1)
Y_test=test_set["value_eur"].copy()
```

Machine Learning Algorithms:

Select and train the model:

Linear Regression

```
from sklearn.linear_model import LinearRegression
lin_reg = LinearRegression()
lin_reg.fit(fifa_dataset_features, fifa_dataset_labels)


from sklearn.metrics import mean_squared_error
fifa_dataset_predictions = lin_reg.predict(X_test) #x_test
lin_mse = mean_squared_error(Y_test, fifa_dataset_predictions) #y_test
lin_rmse = np.sqrt(lin_mse)
#lin_rmse
```

```
print(f"MSE for Linear Regression is {lin_mse} and RMSE is {lin_rmse}")


from sklearn.metrics import r2_score
score = r2_score(Y_test, fifa_dataset_predictions
```

Decision Trees

```
from sklearn.tree import DecisionTreeRegressor
tree_reg = DecisionTreeRegressor(random_state=42)
tree_reg.fit(fifa_dataset_features, fifa_dataset_labels)
fifa_dataset_predictions = tree_reg.predict(X_test)#X_TEST
tree_mse = mean_squared_error(Y_test, fifa_dataset_predictions)#Y_TEST
tree_rmse = np.sqrt(tree_mse)
print(f'MSE for Decision tree is {tree_mse} & RMSE is {tree_rmse}')
score = r2_score(Y_test, fifa_dataset_predictions)
print('Accuracy:',format(score*100,'.2f'),'%')
```

Random Forest

```
from sklearn.ensemble import RandomForestRegressor
forest_reg = RandomForestRegressor(n_estimators=100, random_state=42)
forest_reg.fit(fifa_dataset_features, fifa_dataset_labels)
fifa_dataset_predictions = forest_reg.predict(X_test)
forest_mse = mean_squared_error(Y_test, fifa_dataset_predictions)
forest_rmse = np.sqrt(forest_mse)
print(f'MSE for Random Forest is {forest_mse} & RMSE is {forest_rmse}')
score = r2_score(Y_test, fifa_dataset_predictions)
print('Accuracy:',format(score*100,'.2f'),'%')
```

Evaluation using Cross-Validation

```
from sklearn.model_selection import cross_val_score
scores = cross_val_score(tree_reg, fifa_dataset_features, fifa_dataset_labels,
```

```
                    scoring="neg_mean_squared_error", cv=10)
tree_rmse_scores = np.sqrt(-scores)
def display_scores(scores):
   print("Scores:", scores)
   print("Mean:", scores.mean())
   print("Standard deviation:", scores.std())
display_scores(tree_rmse_scores)
lin_scores = cross_val_score(lin_reg, fifa_dataset_features, fifa_dataset_labels,
                         scoring="neg_mean_squared_error", cv=10)
lin_rmse_scores = np.sqrt(-lin_scores)
display_scores(lin_rmse_scores)
forest_scores = cross_val_score(forest_reg, fifa_dataset_features, fifa_dataset_labels,
                           scoring="neg_mean_squared_error", cv=10)
forest_rmse_scores = np.sqrt(-forest_scores)
display_scores(forest_rmse_scores)
```

Fine-Tune the Model

```
from sklearn.model_selection import GridSearchCV
param_grid = [
   {'n_estimators': [3, 10, 30], 'max_features': [2, 3, 4]},
   {'bootstrap': [False], 'n_estimators': [3, 10], 'max_features': [2, 3, 4]},
 ]
forest_reg = RandomForestRegressor(random_state=42)
grid_search = GridSearchCV(forest_reg, param_grid, cv=5,
                        scoring='neg_mean_squared_error',
                        return_train_score=True)
grid_search.fit(fifa_dataset_features, fifa_dataset_labels)
grid_search.best_params_
grid_search.best_estimator_
cvres = grid_search.cv_results_
for mean_score, params in zip(cvres["mean_test_score"], cvres["params"]):
```

```
print(np.sqrt(-mean_score), params)
```

Evaluate the model on the Test set:

```
test_set = test_set.replace(0, np.nan)
tf = imputer.transform(test_set)
fifa_dataset_tf = pd.DataFrame(tf, columns=fifa_dataset.columns)
fifa_dataset_features = fifa_dataset_tf.drop("value_eur", axis=1)
fifa_dataset_labels = fifa_dataset_tf["value_eur"].copy()
final_model = grid_search.best_estimator_
final_predictions = final_model.predict(fifa_dataset_features)
final_mse = mean_squared_error(fifa_dataset_labels, final_predictions)
final_rmse = np.sqrt(final_mse)
final_rmse
final_model_score = r2_score(fifa_dataset_labels, final_predictions)
print('Accuracy:',format(final_model_score*100,'.2f'),'%')
```

Model.py:

```
import pandas as pd
import numpy as np
# Data visualization
import matplotlib.pyplot as plt
import seaborn as sb
from pandas.plotting import scatter_matrix
# Machine Learning Algorithms
from sklearn.ensemble import RandomForestRegressor
# For Missing Values
from sklearn.impute import SimpleImputer
#For Pickle
import pickle

fifa_raw_dataset = pd.read_csv("C:/Users/Hello/Documents/players_20.csv")
features = ['international_reputation', 'overall', 'potential', 'mentality_composure', 'age',
'height_cm', 'weight_kg','shooting','passing','dribbling','value_eur']
fifa_dataset = fifa_raw_dataset[[*features]]

fifa_dataset = fifa_dataset.replace(0,np.nan)

imputer = SimpleImputer(strategy="median")
imputer.fit(fifa_dataset)
tf = imputer.transform(fifa_dataset)
fifa_dataset_tf = pd.DataFrame(tf, columns=fifa_dataset.columns)
```

App.py:

```
import numpy as np
from flask import Flask, request, jsonify, render_template
import pickle
app = Flask(__name__) #Initialize the flask App
model = pickle.load(open('model.pkl', 'rb'))
@app.route('/')
def home():
    return render_template('index.html')
@app.route('/predict',methods=['POST'])
def predict():
    '''
    For rendering results on HTML GUI
    '''
    int_features = [int(x) for x in request.form.values()]
    final_features = [np.array(int_features)]
    prediction = model.predict(final_features)
    output = round(prediction[0],2)
    return render_template('index.html', prediction_text='Player Value is € {}'.format(output))
if __name__ == "__main__":
    app.run(debug=True)
```

index.html:

```
<!DOCTYPE html>
<html >
<!--From https://codepen.io/frytyler/pen/EGdtg-->
<head>
  <meta charset="UTF-8">
  <title>ML API</title>
  <link href='https://fonts.googleapis.com/css?family=Pacifico' rel='stylesheet' type='text/css'>
```

```
<link href='https://fonts.googleapis.com/css?family=Arimo' rel='stylesheet' type='text/css'>
<link href='https://fonts.googleapis.com/css?family=Hind:300' rel='stylesheet' type='text/css'>
<link href='https://fonts.googleapis.com/css?family=Open+Sans+Condensed:300'
rel='stylesheet' type='text/css'>
<link rel="stylesheet" href="{{ url_for('static', filename='css/style.css') }}">
</head>
<body style="background-color:powderblue;">
 <div class="login">
          <h1>Fifa Player Value in Euros</h1>
    <!-- Main Input For Receiving Query to our ML -->
   <form action="{{ url_for('predict')}}"method="post">
          <input type="text" name="international_reputation"
placeholder="international_reputation" required="required" />
       <input type="text" name="overall" placeholder="overall" required="required" />
         <input type="text" name="potential" placeholder="potential" required="required" />
          <input type="text" name="mentality_composure" placeholder="mentality_composure"
required="required" />
       <input type="text" name="age" placeholder="age" required="required" />
        <input type="text" name="height_cm" placeholder="height_cm" required="required" />
         <input type="text" name="weight_kg" placeholder="weight_kg" required="required" />
          <input type="text" name="shooting" placeholder="shooting" required="required" />
           <input type="text" name="passing" placeholder="passing" required="required" />
           <input type="text" name="dribbling" placeholder="dribbling" required="required" />
       <button type="submit" class="btn btn-primary btn-block btn-large">Predict</button>
   </form>
  <br>
  <br>
  {{ prediction_text }}
  <p>Click the "Predict" button after entering all values. And it will predict the value of
player</p>
```

```
 </div>
</body>
</html>
```

3.10 Result Analysis:

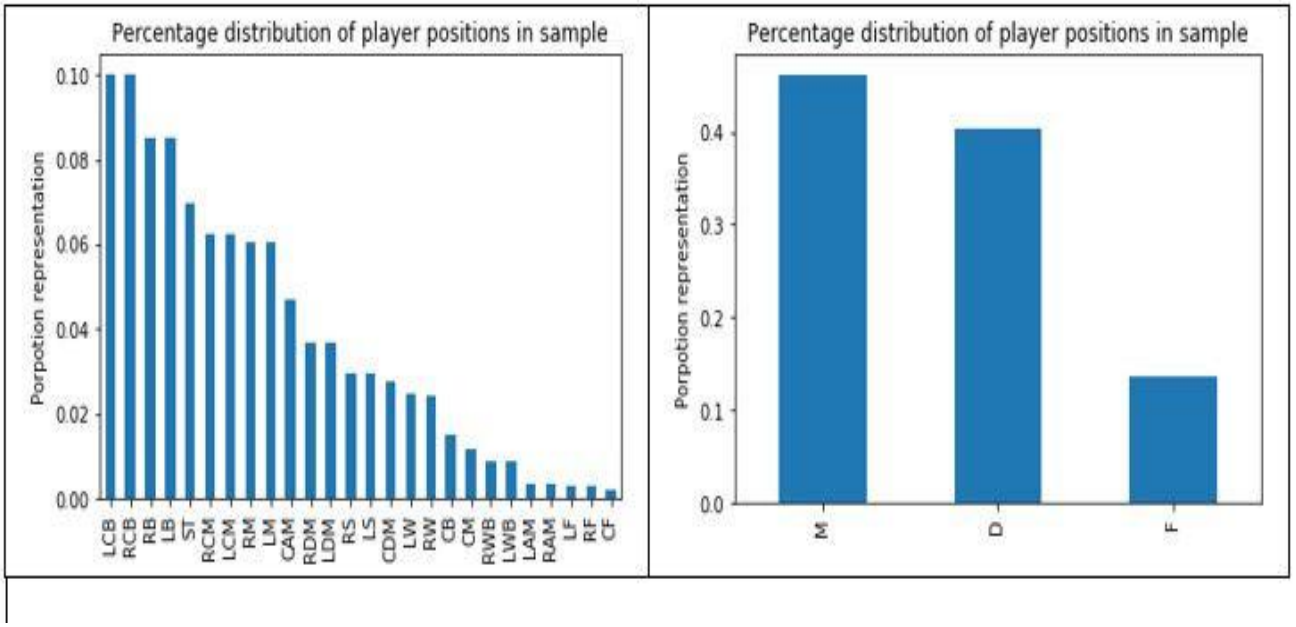


Fig:14 Distribution of players positions in the original dataset before and after grouped.

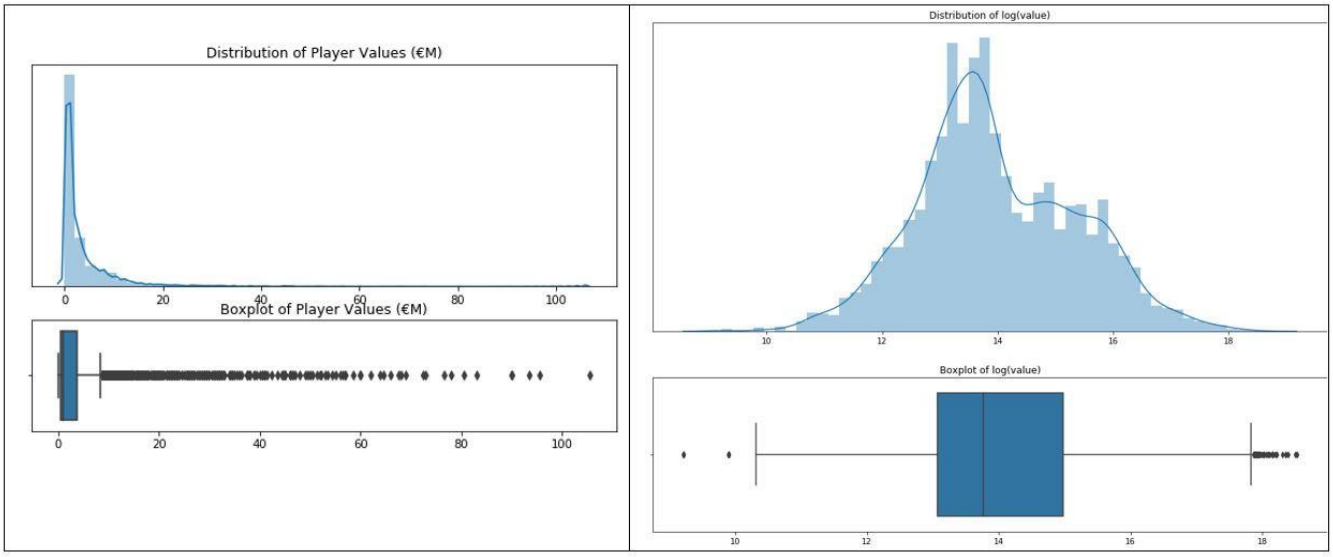


Fig:15 Football player value distribution before and after logarithmic transformation.

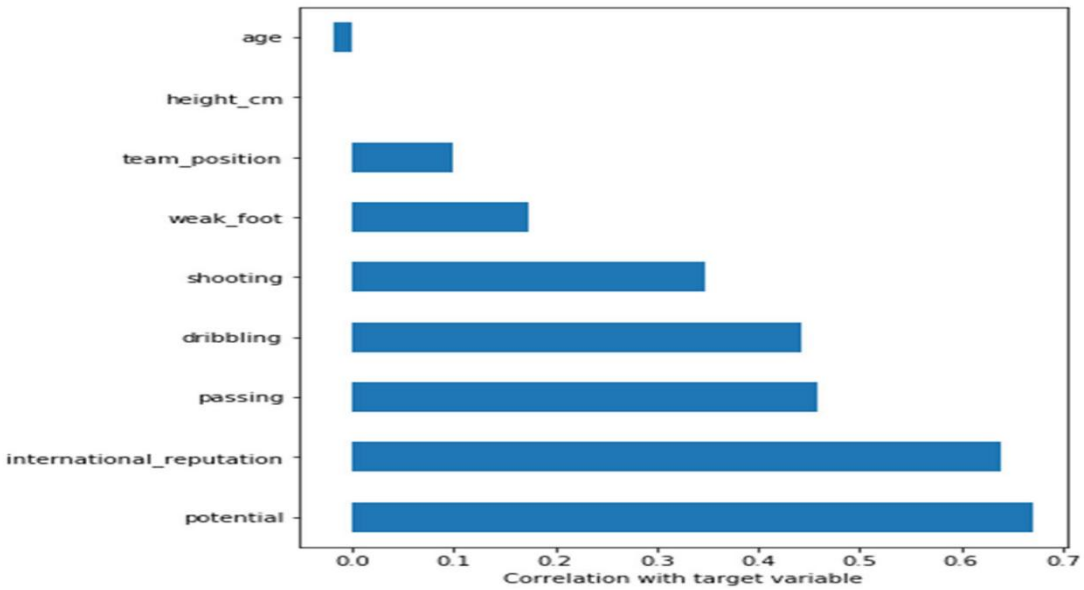


Fig:16 The numerical features correlation to the target variable. (player value)

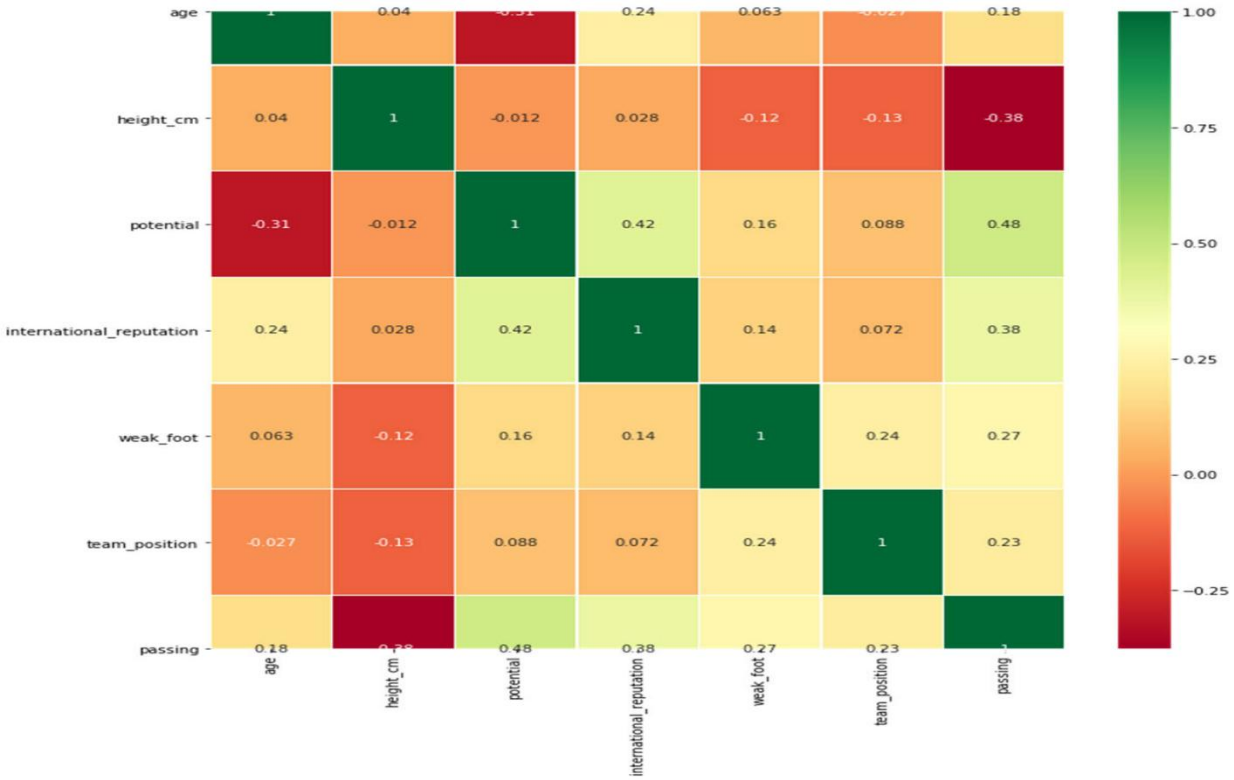


Fig:17 Heatmap for selected attributes.

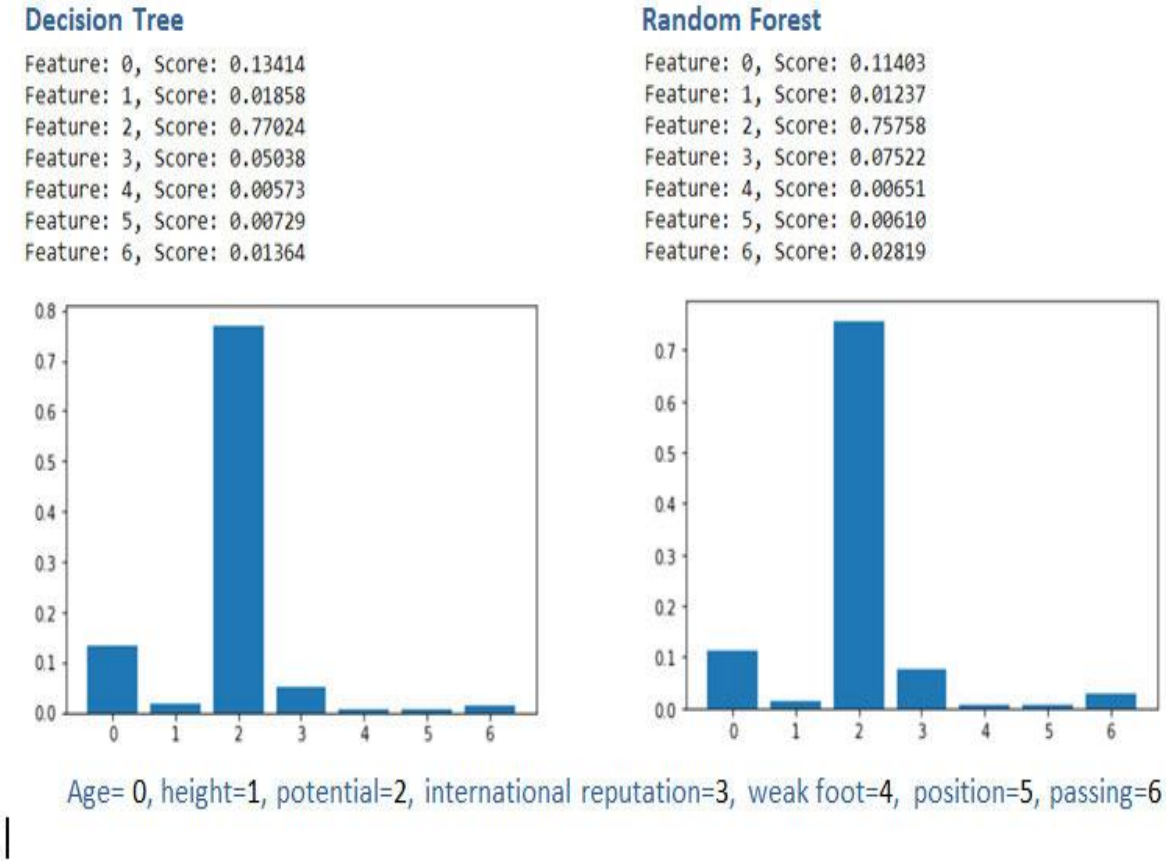


Fig:18 The importance of predictors according to models of decision trees and random forests

N	Classifier	MAE	RMSE	R^2
1	Linear Regression (Baseline)	5,468,144	5,468,144	0.47
2	Multiple Linear Regression	2,618,108	4,662,630	0.61
3	Regression Tree	835,935	2,713,452	0.87
4	Random Forest Regression	576,874	1,649,921	0.95

Table 1. Shows mean absolute errors (MAE), Root mean square errors (RMSE) and the coefficient of determination (R2) for all models.

4.Output Screens

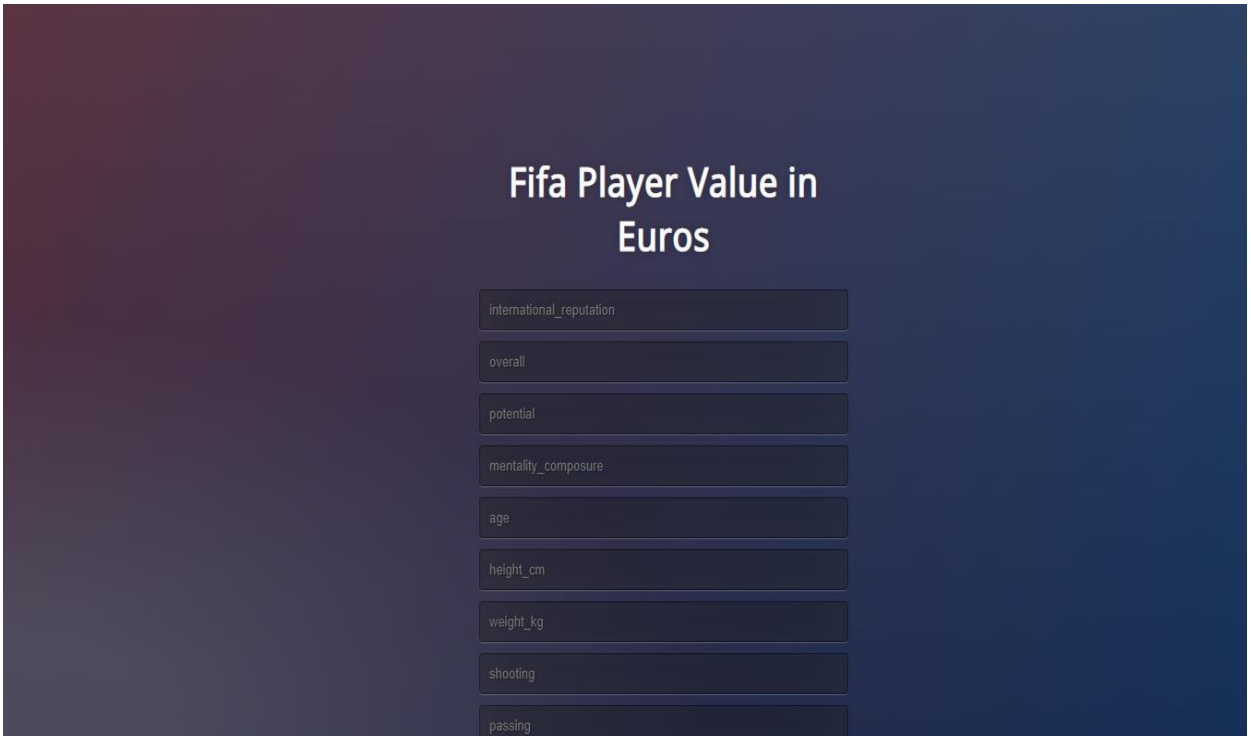


Fig:19 Fifa player value in Euros

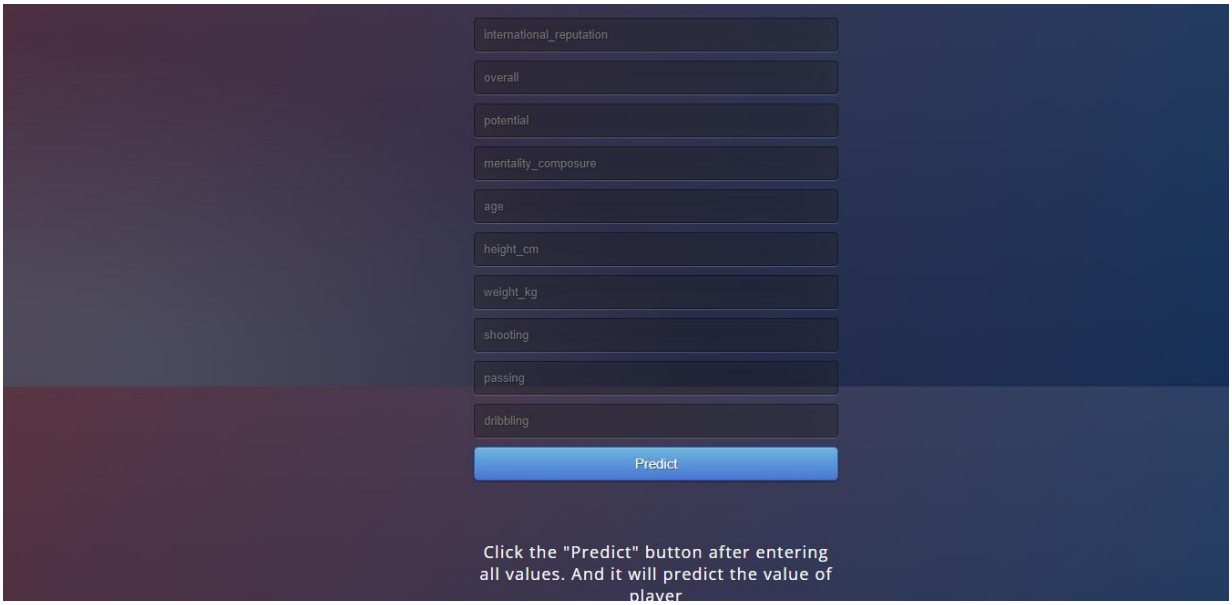


Fig : 20 Attributes

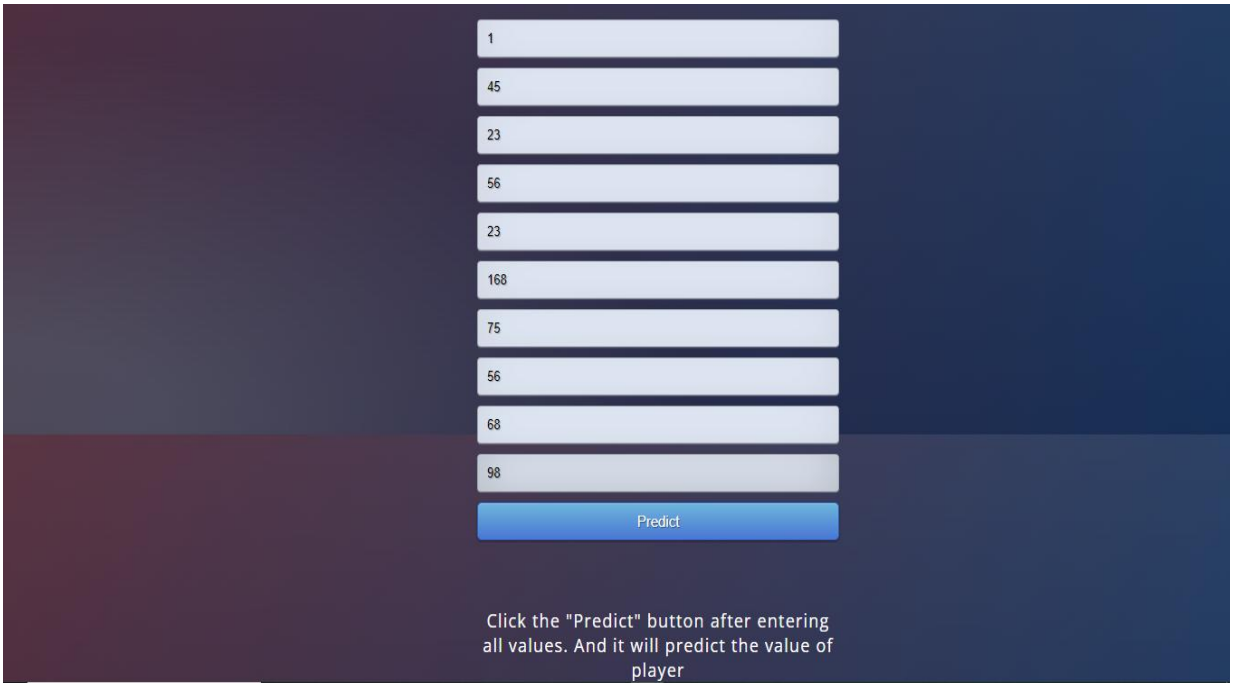


Fig:21 Attributes with values

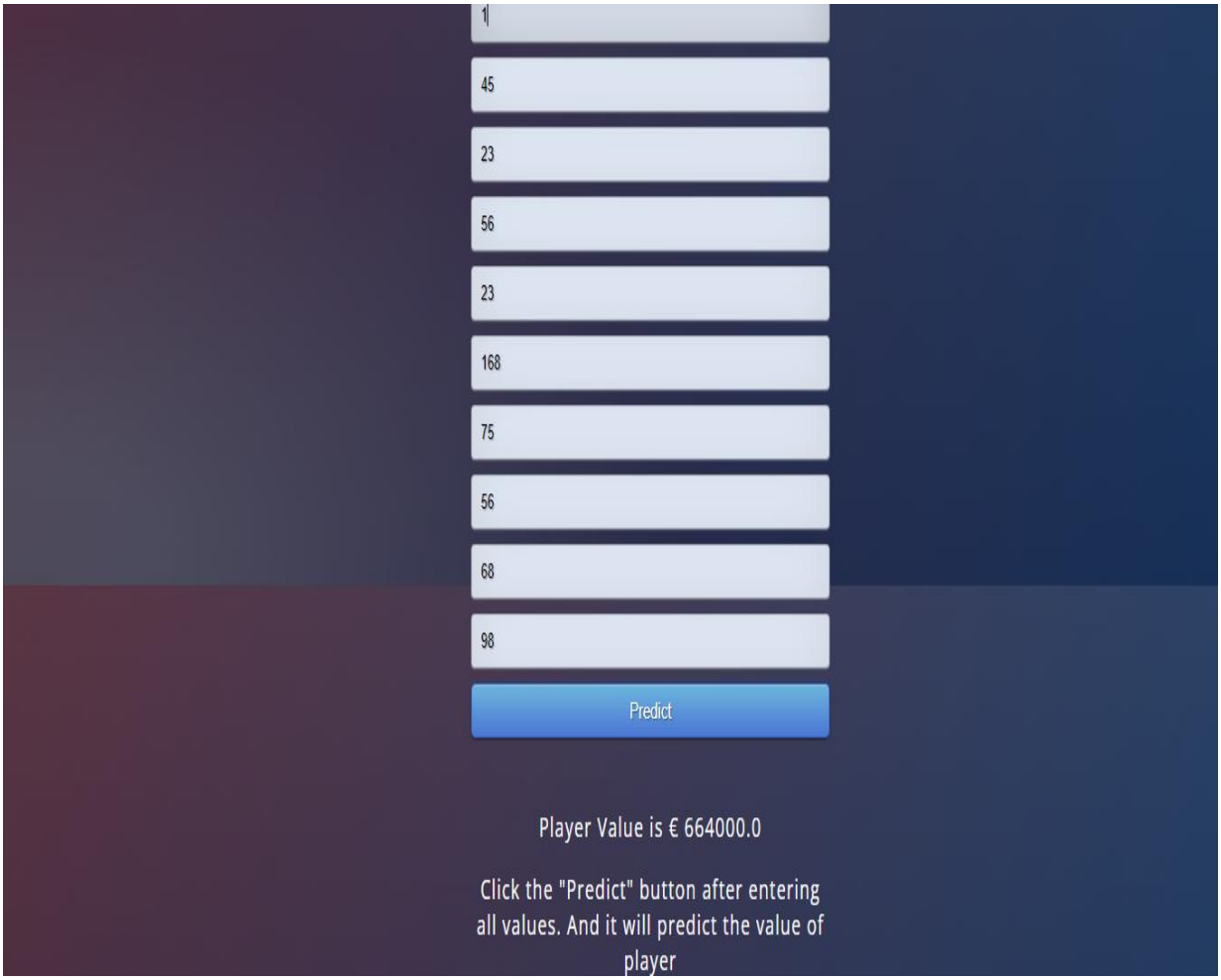


Fig:22 Predicting the player value

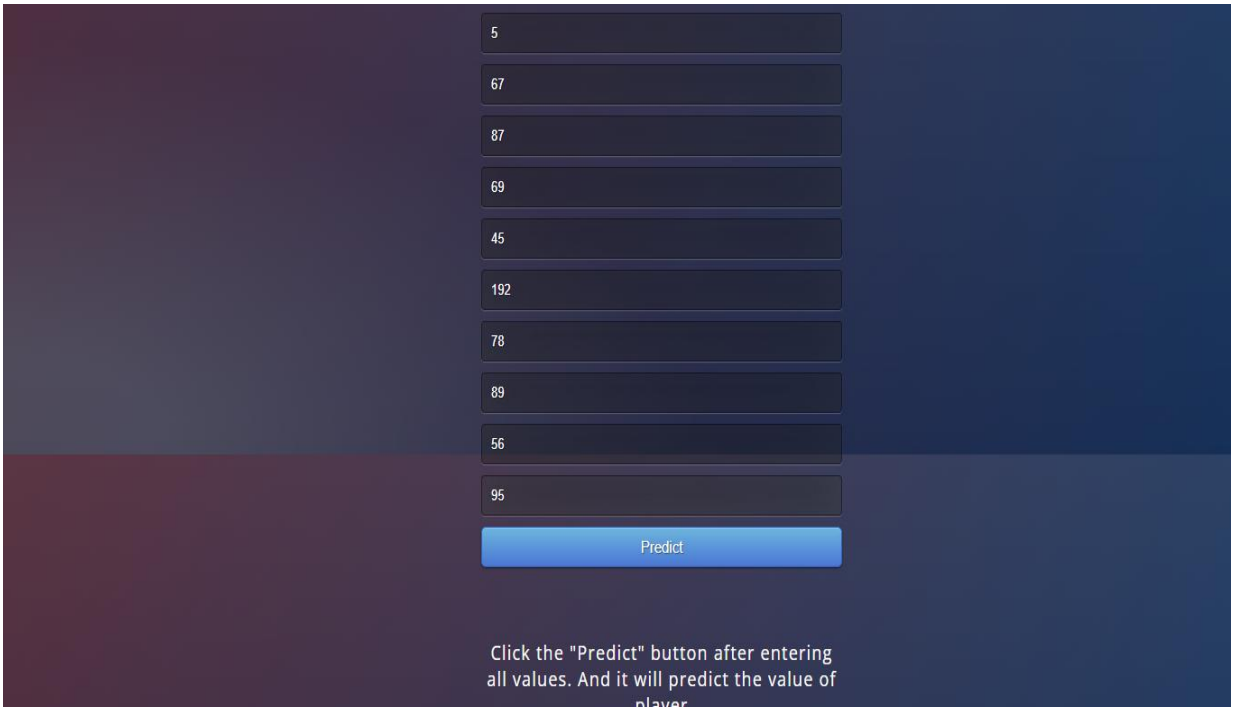


Fig.23 Attributes with values

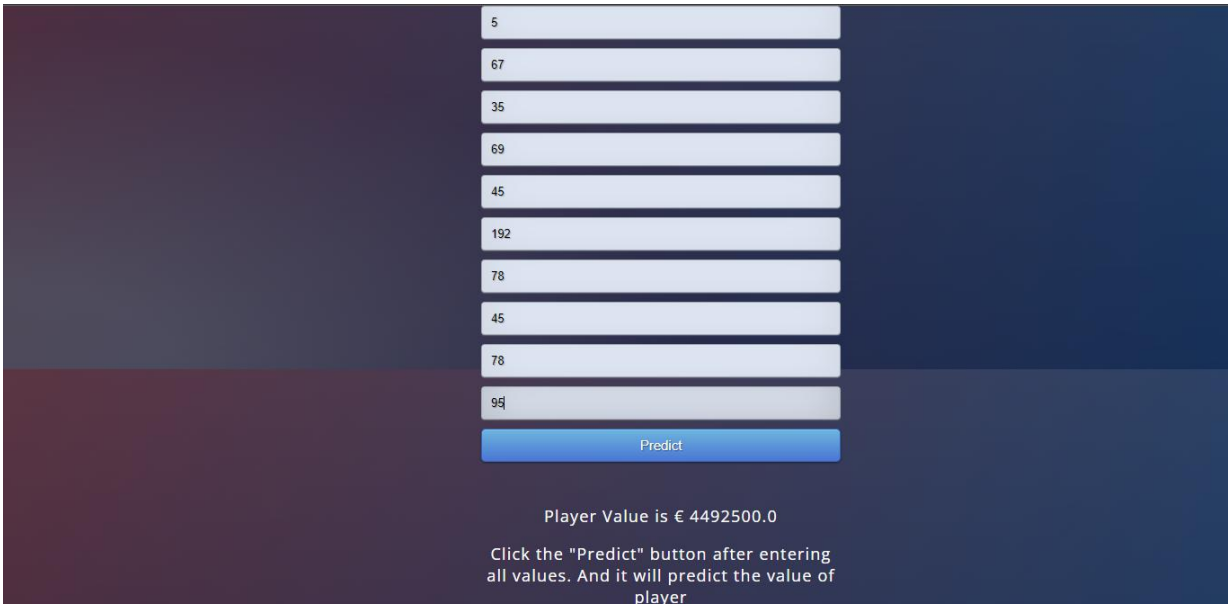


Fig.24 Predicting the player value

5.Conclusion

We have used 3 algorithms like Linear Regression, Decision Trees, Random Forest in- order to predict the value of football players. The accuracy varies for different algorithms. The accuracy for Linear Regression algorithm is 61.06. The accuracy of Random Forest algorithm is 99.9 when correlation are applied. The highest accuracy for Decision trees algorithm is 99.6%. Hence, we conclude Random forest is the best suitable algorithm.

6.Future scope

To develop more accuracy using machine learning algorithms and advanced techniques . The work can be extended and improved for the value prediction of football players by using peep.

7.Bibliography

[1] S. Leone. (2020). *FIFA 20 Complete Player Dataset*. [Online]. Available: https://www.kaggle.com/stefanoleone992/fififa-20-complete-player dataset

[2] C.J. Willmott and K. Matsuura, ‘‘Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,’’ *Climate Res.*, vol. 30, no. 1, pp. 79–82, 2005.

[3] J. Dufour, ‘‘Coeffificients of determination,’’ Dept. Econ., McGill Univ., Montreal, QC, Canada, 2011, pp. 1–14. [Online]. Available: https://monde.cirano.qc.ca/~dufourj/Web_Site/ResE/Dufour_1983_R2W.pdf

[4] S. Kiefer, ‘‘The impact of the Euro 2012 on popularity and market value of football players,’’ Inst. Org. Econ., Berlin, Germany, 2012. [Online]. Available: https://www.wiwi.uni-muenster.de/io/forschen/downloads/DPIO_07_2020

[5] R.Poli, L. Ravenel, and R. Besson, ‘‘CIES football observatory,’’ *Annu. Rev.*, pp. 1–83, 2014. [Online]. Available: https://www.footballobservatory.com/IMG/pdf/ar2013_exc-2.pdf

8. Conference Paper

AN INTELLIGENT FRAMEWORK FOR PREDICTING THE VALUE OF FOOTBALL PLAYERS

G.Bodhini 1, K.Vengamamba 2, V.Sreevani 3, M.Satyam Reddy 4

1,2,3 Student, Department of CSE, Narasaraopeta Engineering College,Narasaraopeta,Guntur(D.T) A.P, India

4 professor, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur(D.T) A.P,India

golibodhini2002@gmail.com[1], kandulavengamba@gmail.com, vallurisreevani @gmail.com[3],

---

Abstract

As we all know that football is a very popular and a trending game across the globe and the football players like Christiano Ronaldo,Lionel Messi are became very popular in recent games. We all know their names and origin of the famous players, but many of us don't know their net value.Market values also play a vital role. Generally the market values are predicted by football experts. Actually the expert decisions are incorrect and not transparent. Now, we are going to propose a method to determine the football player's net value. This is completely based on machine learning algorithms. Here we are going to use a fifa 20 dataset, which is collected from kaggle.com. In this approach, we are going to use 4 models like Linear Regression, Multiple Regression, Decision Tree, Random forest. Here, we will take the most important factors that will help in predicting the player's market value. The results will be highly accurate, good performance and less errors. These results will help in between the foot ball clubs and player's agents. Hence, from this we can predict the football player's market value.

Keywords - player's value prediction, Linear Regression, Multiple Regression, Decision Tree, Random forest, machine learning.

I. INTRODUCTION

The football is one of the tremendous game in the world.The popularity for football players are increasing drastically day by day. The experts are paying keen observation on the market value of the players.So to determine the value we are taking different categories such as player characteristics, player performance and player popularity. Nowadays machine Learning is used in every domain, such as finance, disease prediction, value prediction etc. Here we are using FIFA 20 data set collected from kaggle.com.In this dataset, we have approximately more than 17,000 players. By using this dataset, we can predict the value of the player accurately and efficiently.Now, we only consider the attributes that will help in estimating the net value of football players such as height, weight,age, passing etc. We are using four models such as Linear Regression, Multiple Regression, Decision Tree and Random forest.After processing the data with various models, we conclude that Random forest is the best model. It requires less inputs and gives best result. The results are accurate and efficient.

II. EXISTING SYSTEM

Previously, judgments are made by agents and experts based on their experience and knowledge. This will result in leading many errors, takes a lot of time to calculate and more expensive which affects the value of football players. Football experts will calculate the value based on the player's characteristics, player's performance and player's popularity. These expert decisions are sometimes incorrect and not efficient. They also consume lot of time.Hence, lacking of accuracy may results in lack of value of football players.

III. PROPOSED SYSTEM

Now, we are going introducing machine learning in our approach. Here, we are using four machine learning algorithms like Linear Regression, Multiple Regression,Decision Tree and Random Forest. Here we will consider three important factors such as player characteristics, player performance and player popularity.

1. Player Characteristics:

This is one of the most important characteristic which is to consider. These include Age, player height, weight and player position. Age is an attribute which reflects in experience and ability.

Height, which helps in increase the score and preventing the goals. Weight of the player will help in estimating the value of the football player. Player position like defender, midfielder, goal keeper etc are useful in predicting the value of football player.

2. Player Performance:

Player performance which includes passing, shooting, dribbling and yellow and red

cards.Passing, it represents the passing of a ball intentionally from one player to another player in the same team. Shooting represents the hitting of a ball in an attempt to score the goal. Dribbling I t represents the passing of a ball in a given direction and avoiding the defender's attempts to intercept the ball. Yellow card and Red card represents the number of warnings and mistakes they have committed.

3.Player Popularity :

The popularity will also helps in determining the value of football player.It means the crowd pulling power and the image of the player they show on the pitch. Hence, we can say international reputation will play a major role in determining the market value of a player.By using these algorithms we can reduce the complexity in predicting the value the football players, reduce errors, enhance accurate results. It is easier to predict the value and it will also help the club agents to make quick decisions. The main advantages are Generate accurate and efficient results, Computation time is greatly reduced, Reduces manual work, Efficient and transparent results.

IV. ALGORITHMS

Here, we are using 4 algorithms to predict the value of a football players. The four algorithms are Linear Regression, Multiple Regression, Decision Tree and Random Forest.

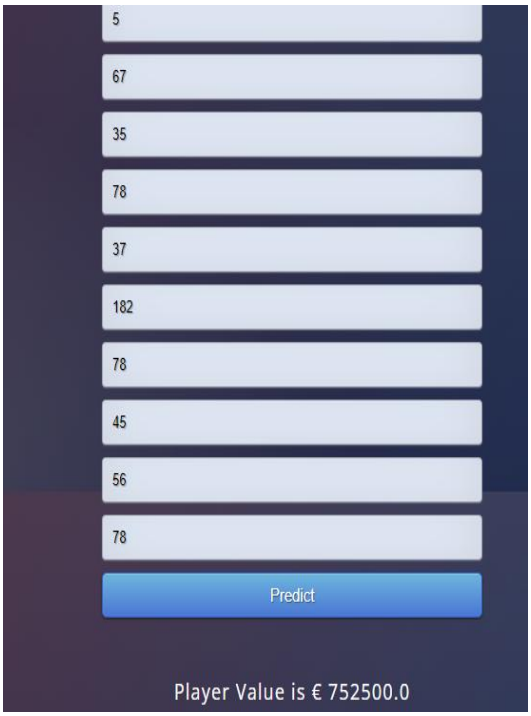
- Linear Regression is the simplest form of algorithm, it is used to predict the relationship between two variables. There are 2 types of variables such as Independent and dependent variable. Suppose Y=a+X, here X is Independent variable. Since for a given value of X, Y is changing accordingly. Hence Y is a dependent variable. These two variables are used to predict the target variable.

- Multiple Regression is the extension of Linear Regression. It is used to find the relationship between two or more number of independent variables and one dependent variable. For example, we are making survey for finding the reasons for lung cancer. The factors include smoking, drinking etc. Here, we are having two or more variables as the reason for lung cancer. In this scenario, we will use multiple regression.

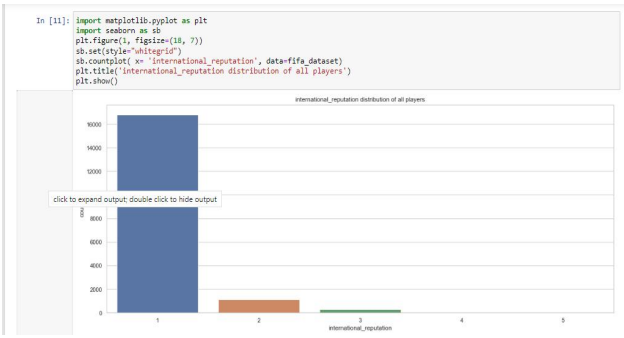
- Decision Tree is a supervised learning technique. It is having a flow-chart structure. It is having internal nodes, branches and leaf nodes.Internal nodes represents the attributes of the dataset, branches represent the decision rules and leaf nodes tells about the outcome. The decisions rules are taken from the dataset. Here the attributes are compared among the decisions, if it matches it will show you the outcome else it will skip the respective condition and jump to the next node.

- Random Forest is the group of decision trees.It

will split the data into subparts and solve the complex problem.It will predict the accuracy of the dataset. The more number of decision trees in the forest leads to higher accuracy,good transparency and it avoids the problem of overfitting.The Random Forest is having the highest coefficient of determination where as linear regression is having the lowest coefficient of determination. Hence, we conclude Random Forest is the best model.



V. CODE IMPLEMENTATION



VI. CONCLUSION

N	Classifier	MAE	RMSE	R^2
1	Linear Regression (Baseline)	5,468,144	5,468,144	0.47
2	Multiple Linear Regression	2,618,108	4,662,630	0.61
3	Regression Tree	835,935	2,713,452	0.87
4	Random Forest Regression	576,874	1,649,921	0.95

Here, we have used 4 algorithms. We make use of 3 metrics such as, Mean Absolute Error(MAE) , Root Means Square Error (RMSE), Coefficient of Determination(R2). The above table shows the errors between the actual and predicted values . From the above table the random forest shows the least root mean square between the actual and predicted values, where as Linear Regression provided the highest root mean square values.Here we also calculated the coefficient of determination. The value which is close to 1 indicates with zero

error. If the value is close to 0, It means it shows the error.From the above table we can say that the random forest algorithm provides the highest coefficient of determination and linear regression provides the least coefficient of determination. Hence we conclude the Random forest is the best suitable mode for modelling.Ultimately, we can say that expert judgments are not accurate and it also consumes lots of time. The results are not transparent and inefficient. So, By using these machine learning algorithms we will predict the value of football players accurately and efficiently.

VII.REFERENCES

[1] R. Asif, ‘‘Football (soccer) analytics: A case study on the availability and limitations of data for football analytics research,’’ Int. J. Comput. Sci. Inf. Secur., vol. 14, no. 11, p. 516, 2016.

[2] R. Poli, L. Ravenel, and R. Besson, ‘‘CIES football observatory,’’ Annu. Rev., pp. 1–83, 2014. [Online]. Available: https://www.footballobservatory.com/IMG/pdf/ar2013_exc-2.pdf.

[3] S. Kiefer, ‘‘The impact of the Euro 2012 on popularity and market value of football players,’’ Inst. Org. Econ., Berlin, Germany, 2012.[Online].Available:https://www.wiwi.uni muenster.de/io/forschen/downloads/DPIO_07_2020.

[4] S. Leone. (2020). FIFA 20 Complete PlayerDataset. [Online]. Available: https://www.kaggle.com/stefanoleone992/fifa-20-complete-playerdataset.

[5] J. Dufour, ‘‘Coeffificients of determination,’’ Dept. Econ., McGill Univ., Montreal, QC, Canada, 2011, pp. 1–14. [Online]. Available:https://monde.cirano.qc.ca/~dufourj/Web_Site/ResE/Dufour_1983_R2_ W.pdf

[6] J. L. Felipe, A. Fernandez-Luna, P. Burillo, L. E. de la Riva, J. Sanchez-Sanchez, and J. Garcia-Unanue, ‘‘Money talks: Team variables and player positions that most inflfluence the market value of professional male footballers in Europe,’’ Sustainability, vol. 12, no. 9, p. 3709, May 2020.

[7] L. Cotta, ‘‘Using fififa soccer video game data for soccer analytics,’’ in Proc. Workshop Large Scale Sports Anal., 2016, pp. 1–4.

[8] P. Siuda, ‘‘Sports gamers practices as a form of subversiveness–the example of the FIFA ultimate team,’’ Crit. Stud. Media Commun., vol. 38, no. 1, pp. 75–89, 2021.

[9] P. Awasthi, A. Beutel, M. Kleindessner, J. Morgenstern, and X. Wang, ‘‘Evaluating

fairness of machine learning models under uncertain and incomplete information,'' in Proc. ACM Conf. Fairness, Accountability, Transparency, Mar. 2021, pp. 206–214.

[10] J. Gareth, An Introduction to Statistical Learning: With Applications in R. Springer, 2013.

Similarity Check Report

DG-5

ORIGINALITY REPORT

14%	2%	13%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Mustafa A. AL-ASADI, Sakir Tasdemir. "Predict the Value of Football Players Using FIFA video game data and Machine Learning Techniques", IEEE Access, 2022 Publication	10%
2	S. Siva Sunayna, S. N. Thirumala Rao, M. Sireesha. "Chapter 25 Performance Evaluation of Machine Learning Algorithms to Predict Breast Cancer", Springer Science and Business Media LLC, 2022 Publication	1%
3	www.researchgate.net Internet Source	1%
4	"Proceedings of Second International Conference on Advances in Computer Engineering and Communication Systems", Springer Science and Business Media LLC, 2022 Publication	1%
5	yorkspace.library.yorku.ca Internet Source	1%

6	www2.mdpi.com Internet Source	1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		


NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)
Approved by AICTE, Permanently Affiliated to JNTUK, Kakinada, NIRF Ranking (251-300 Band), Accredited by NBA (Tier-I) & NAAC with 'A+' Grade
Kotappakonda Road, Yellamanda (Post), Narasaraopet - 522601, Palnadu Dist., Andhra Pradesh, INDIA. Website:www.nrtec.in
PAPER ID
NECICAIEA2K23080
International Conference on
Artificial Intelligence and Its Emerging Areas
NEC-ICAIEA-2K23
17th & 18th March, 2023
Organized by Department of Computer Science and Engineering in Association with CSI
Certificate of Presentation
This is to Certify that Goli Bodhini , Narasaraopeta Engineering College has presented the paper title Predicting the value of football players in the International Conference on Artificial Intelligence and Its Emerging Areas-2K23 [NEC-ICAIEA-2K23], Organized by Department of Computer Science and Engineeringin Association with CSI on 17th and 18th March 2023 at Narasaraopeta Engineering College, Narasaraopet, A.P., India.
Convenor
Dr.S.V.N.Srinivasu
Chief-Convenor
Dr.S.N.Tirumala Rao
Principal, Patron
Dr. M. Sreenivasa Kumar
MHRD'S INNOVATION CELL (GOVERNMENT OF INDIA)