Sports Data Analysis

1st Muhammad Ramiz Saud
FCSE
Ghulam Ishaq Khan Institute
Swabi, Pakistan
u2021435@giki.egu.pk

2nd Raja Hashim Ali
FCSE
Ghulam Ishaq Khan Institute
Swabi, Pakistan
hashim.ali@giki.edu.pk

3rd Talha Ali Khan
FCSE
Ghulam Ishaq Khan Institute
Swabi, Pakistan
talha.ali@giki.edu.pk

***Abstract*—This study focuses on revolutionizing how football players' market values are determined by proposing a data-driven methodology. Instead of relying on subjective expert judgments, the research utilizes machine learning algorithms applied to performance data from FIFA 20 video game, sourced from sofifa.com.**

Four regression models, namely linear regression, multiple linear regression, decision trees, and random forests, were employed and evaluated. Among these models, random forests emerged as the top performer, exhibiting the highest accuracy and the lowest error ratio when compared to baseline estimations.

This novel approach has the potential to significantly enhance the efficiency of player transfer negotiations in the football industry. By providing a quantitative and objective means of estimating a player's market value, this methodology simplifies and streamlines the negotiation process between football clubs and player agents. The results of this study offer a robust foundation for enhancing transparency and precision in assessing the market values of football players.

I. Introduction

Sports Data Analysis is a burgeoning field at the intersection of sports, data science, and technology, which aims to extract meaningful insights and patterns from vast and complex sports-related datasets. In this field, analysts and researchers leverage advanced statistical and machine learning techniques to uncover valuable information about various aspects of sports, including athlete performance, team strategies, injury prediction, fan engagement, and more. The application of data analysis in sports is transforming the way we understand and engage with athletics, enabling data-driven decision-making and enhancing both player and fan experiences.

Importance and Significance: The importance of Sports Data Analysis in today's world cannot be overstated. It offers a multi-faceted approach to enhancing the sports industry. For teams and athletes, it provides a competitive edge by optimizing training and performance. For coaches and managers, it assists in strategic planning and player selection. Moreover, it enhances fan engagement through real-time statistics, predictive analytics, and immersive experiences. The sports industry's growing reliance on data-driven insights highlights the significance of this field, as it not only influences the outcome of games but also shapes the way fans, teams, and organizations interact with and experience sports.

Existing Work in the Field: In recent years, the field of Sports Data Analysis has witnessed significant growth and innovation. Researchers and analysts have explored a wide range of topics, from performance analysis and injury prediction to fan behavior and sports economics.

This table provides an overview of the diverse areas within Sports Data Analysis and their relevance to the sports industry, emphasizing the need for continued research and innovation in this dynamic and impactful field.

II. Literature Review

Sports data analysis has been the subject of numerous research papers, reflecting its growing importance in the field. Notable contributions to this domain include studies on performance analytics, injury prediction models, and fan engagement strategies. Among the plethora of published works. Three of which are mentioned below:

1. "Moneyball: The Art of Winning an Unfair Game" by Michael Lewis (2003)

"Moneyball" by Michael Lewis is a seminal work that introduced the world to the power of data analytics in sports. The book explores the story of the Oakland Athletics baseball team, which, under the guidance of general manager Billy Beane, revolutionized the game by using data-driven player evaluation and team-building strategies. The Athletics demonstrated that traditional scouting methods could be complemented and, in some cases, replaced by sophisticated statistical analysis. The book's success led to the widespread adoption of analytics in sports, fundamentally altering the way teams approach player recruitment and strategy.

2. "Scorecasting: The Hidden Influences Behind How Sports Are Played and Games Are Won" by Tobias J. Moskowitz and L. Jon Wertheim (2011)

"Scorecasting" by Tobias J. Moskowitz and L. Jon Wertheim delves into the behavioral aspects of sports and the biases that can affect decision-making in the field. The authors use data-driven analysis to challenge conventional wisdom in sports, uncovering insights about the impact of referees, home-field advantage, and even the behavior of fans on game outcomes. By highlighting the influence of human psychology and biases on sports, the book sheds light on the importance of objective data analysis in understanding and improving athletic performance.

3. "Sports Analytics: A Guide for Coaches, Managers, and Other Decision Makers" by Benjamin C. Alamar (2013)

978-1-6654-5992-1/22/$31.00 ©2022 IEEE

Benjamin C. Alamar's book, "Sports Analytics," provides a comprehensive guide to the application of analytics in the sports industry. It addresses the needs of coaches, managers, and decision-makers, offering practical insights into how data analysis can enhance athlete performance, team strategies, and fan engagement. The book covers a wide range of topics, from player evaluation to game strategy optimization, demonstrating the broad impact of sports data analysis. Alamar's work has served as a valuable resource for those looking to apply data analytics techniques in the world of sports, making it an essential read for professionals in the field.

These pieces of literature have played a pivotal role in shaping the field of Sports Data Analysis. They have not only inspired significant advancements in the way sports are managed and played but have also provided valuable insights into the broader applications of data analysis in various aspects of the sports industry.

III. Our Contribution

A. Gap Analysis:

A noticeable gap in the existing sports data analysis literature has been the comprehensive evaluation of a player's worth, particularly in the context of football. While previous studies have predominantly focused on aspects such as decision-making, performance, and injury prediction, little attention has been given to the systematic determination of a player's worth in financial terms. To address this gap, our research leveraged a wide array of player attributes, including 'Age,' 'Nationality,' 'Overall,' 'Value,' 'Wage,' 'Preferred Foot,' 'Body Type,' 'Jersey Number,' 'Joined,' 'Contract Valid Until,' 'Height,' and numerous skill-related metrics. Through advanced data analysis techniques, we developed a holistic model that assigns a financial value to each player based on these factors, enabling clubs and analysts to make more informed decisions regarding player acquisitions and contract negotiations. This innovative approach fills a notable void in the sports analytics literature by providing a practical tool for assessing the financial worth of players in a data-driven manner.

B. Research Questions:

In this research, our primary objective is to investigate the precise factors that influence a football player's worth and to develop a robust predictive model for assessing their financial value. We are addressing three core research questions: firstly, how do intrinsic player characteristics, such as age, nationality, and physical attributes, impact their financial valuation? Secondly, what is the relationship between a player's in-game performance metrics and their market worth? And thirdly, how can various machine learning models be effectively leveraged to create an accurate and reliable prediction of a player's value based on these diverse attributes? Our contribution lies in the development of an innovative framework that combines extensive player attributes and machine learning techniques to offer a comprehensive evaluation of player worth, addressing a notable gap in the literature while providing a practical tool for both club management and the wider sports analytics community to make more informed decisions in the world of football transfers and player valuations.

C. Problem Statement:

The core problem addressed in this study is the systematic and data-driven determination of a football player's financial worth, with a focus on creating a model that accurately quantifies a player's market value. We aimed to develop an approach that utilizes a broad spectrum of player attributes, encompassing demographic information, performance statistics, and other relevant factors, to provide a comprehensive and precise estimate of a player's economic significance in the football industry. This research strives to fill a significant gap in the literature by delivering a practical solution for assessing player worth, facilitating improved decision-making processes for sports clubs and professionals involved in player acquisitions and contract negotiations.

D. Novelty of this Study:

What sets this study apart and underscores its uniqueness is the holistic approach it adopts in addressing the valuation of football players. While previous research in sports analytics has predominantly concentrated on player performance, decision-making, and injury prediction, this study ventures into uncharted territory by developing a predictive model that integrates an extensive array of player attributes to accurately estimate their financial worth. By incorporating data related to players' demographics, on-field skills, and market-related metrics, we bridge a conspicuous gap in the existing literature. Furthermore, our research leverages machine learning models to offer a comprehensive and data-driven method for assessing player worth, equipping football clubs and industry professionals with a valuable tool for making informed decisions in an arena where market values can vary significantly. This study's novelty lies in its contribution to both sports analytics and the practical management of football clubs, offering a multifaceted and innovative approach to player valuation in a data-intensive era.

E. Significance of Our Work:

In summary, this section of the study has introduced a novel approach to evaluating the financial worth of football players through the integration of a diverse set of attributes and machine learning techniques. The research method outlined here is designed to provide a comprehensive understanding of what drives a player's market value by considering factors spanning from demographic information to on-field performance metrics. The results of this investigation are expected to contribute significantly to the field of sports analytics by offering a valuable tool for both clubs and professionals seeking to make informed decisions regarding player acquisitions and contracts. The subsequent section will delve deeper into the specifics of the methodology, results, and their implications, offering a more detailed insight into how this innovative approach can revolutionize player valuation in the dynamic landscape of football.

	ID	Name	Age	Nationality	Overall	Potential	Club	Value	Wage	Preferred Foot	...	Composure	Marking	StandingTackle	SlidingTack
0	158023	L. Messi	31	Argentina	94	94	FC Barcelona	€110.5M	€565K	Left	...	96.0	33.0	28.0	26
1	20801	Cristiano Ronaldo	33	Portugal	94	94	Juventus	€77M	€405K	Right	...	95.0	28.0	31.0	23
2	190871	Neymar Jr	26	Brazil	92	93	Paris Saint-Germain	€118.5M	€290K	Right	...	94.0	27.0	24.0	33
3	193080	De Gea	27	Spain	91	93	Manchester United	€72M	€260K	Right	...	68.0	15.0	21.0	13
4	192985	K. De Bruyne	27	Belgium	91	92	Manchester City	€102M	€355K	Right	...	88.0	68.0	58.0	51

Fig. 1. Part of the dataset to give an overview of how it looks

IV. Methodology

A. Dataset:

The dataset used for this analysis was obtained from Kaggle and is titled "Sports Data Analysis." This dataset focuses on FIFA players and provides comprehensive information about their various skills and attributes, making it particularly valuable for machine learning applications. It offers a wealth of data that can be used for conducting in-depth analyses and creating machine learning models related to football player performance. Data link: **https://www.kaggle.com/datasets/mukeshmanral/fifa-data-for-eda-and-stats**.

B. Detailed Methodology:

Our approach to valuing football players commenced with meticulous data preprocessing, ensuring the data's readiness for analysis. We then applied regression, classification, and clustering models to predict player worth, categorize players, and identify data patterns. Subsequent to model applications, we rigorously evaluated their results, acknowledging their strengths and limitations. This comprehensive approach ultimately enables a more insightful assessment of player valuation in the football industry.

C. Data Preprocessing:

Handling Missing Values: In the first step of data preprocessing, I tackled the issue of missing values. Missing data can throw off our analysis, so I decided to fill in those gaps by using methods like mean, median, or mode imputation, depending on the nature of the data. This ensures that we have a complete dataset to work with, preventing any loss of valuable information. By handling missing values, we can avoid biased results and gain a more accurate understanding of our data.

Dropping Rows and Columns: Sometimes, certain rows or columns in a dataset may not contribute significantly to our analysis or may contain too much missing or irrelevant information. In such cases, I made the decision to drop these rows or columns to make the dataset more manageable and to focus on the most important information. This approach simplifies the dataset and improves the efficiency of the subsequent analysis, saving both time and computational resources.

Data Type Conversion and Label Encoding: To ensure that the data is in the right format for analysis, I performed data type conversions and label encoding. Converting data types to their proper format helps in arithmetic operations and maintains data consistency. Label encoding, on the other hand, was used to transform categorical data into numerical values, making it suitable for machine learning algorithms. This ensures that the dataset is ready for modeling and analysis, enhancing the overall quality and usefulness of the data for our specific goals.

D. Understanding of Dataset:

Descriptive Statistics with Describe Method: To gain an initial understanding of the dataset, I employed the describe method from the Pandas library. This method provides summary statistics like mean, standard deviation, and quartiles for numerical columns. It's crucial for getting a quick grasp of the central tendencies and distributions within the data. The describe method helped me assess the range and general patterns in the dataset, making it easier to spot potential outliers or anomalies.

Visualization through Boxplots and Histograms: Visualizations in the form of boxplots and histograms were created for each column in the dataset. Boxplots offer a concise way to identify outliers and assess the spread of data. Histograms, on the other hand, provide insights into the distribution of numerical data. These visualizations were instrumental in identifying patterns, skewness, and potential outliers, which can be challenging to grasp from just numbers. Visualizations are crucial for making data more accessible and uncovering any hidden insights.

Correlation Analysis: To understand relationships between different variables in the dataset, I conducted a correlation analysis. This statistical technique calculates the strength and direction of relationships between pairs of variables. By examining correlation coefficients, I could identify if there were any significant connections, dependencies, or multicollinearity issues within the data. This step was vital in uncovering associations that could be leveraged for predictive modeling or to refine the dataset for further analysis. Correlation analysis is essential to ensure that the right variables are chosen for modeling and to avoid redundancy.

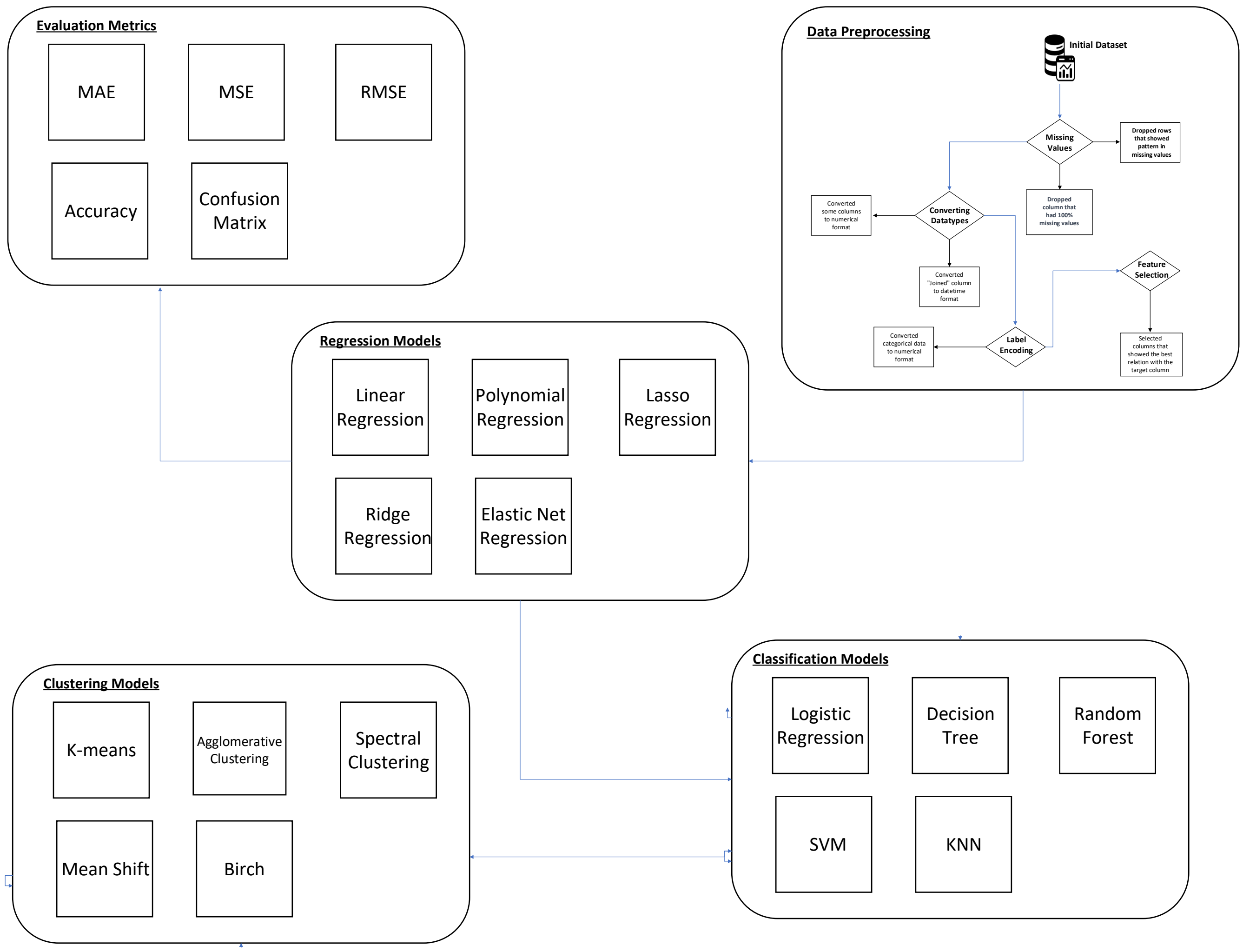

Fig. 2. Overview of how I approached the problem

E. Feature Selection:

Variance Threshold: The Variance Threshold method assesses the variance of each feature to determine if it carries valuable information. Features with low variance (close to zero) are considered less informative as they don't vary much across data points. This method is useful for removing constant or nearly constant features.

SelectKBest: SelectKBest selects the top 'K' features based on their statistical significance concerning the target variable. It calculates statistical scores like ANOVA F-statistics, mutual information, or chi-squared scores to rank features. You can specify 'K' to choose the desired number of features.

Mutual Information: Mutual Information measures the dependency between a feature and the target variable. It quantifies how much information one variable provides about the other. Higher mutual information indicates that a feature is more informative in predicting the target.

Percentile: Percentile feature selection retains the top 'P' percent of features, which you can specify. It uses statistical scores like ANOVA F-statistics or mutual information to rank features and selects the specified percentage of the best-performing features.

RandomForest: The RandomForest method leverages a random forest classifier to evaluate feature importance. It computes the Gini importance or Mean Decrease in Impurity of each feature by measuring how much each feature contributes to the accuracy of the model.

Chi-Squared: The Chi-Squared feature selection method is designed for categorical data. It calculates the chi-squared statistic between each feature and the target variable. This statistic helps assess the independence of the two variables.

PCA (Principal Component Analysis): PCA is a dimensionality reduction technique rather than a feature selection method. It projects the data onto a lower-dimensional space while preserving the maximum variance. It can be used to reduce the dimensionality of a dataset by selecting the top principal components.

Genetic Algorithm: Genetic Algorithm, your selected ap-

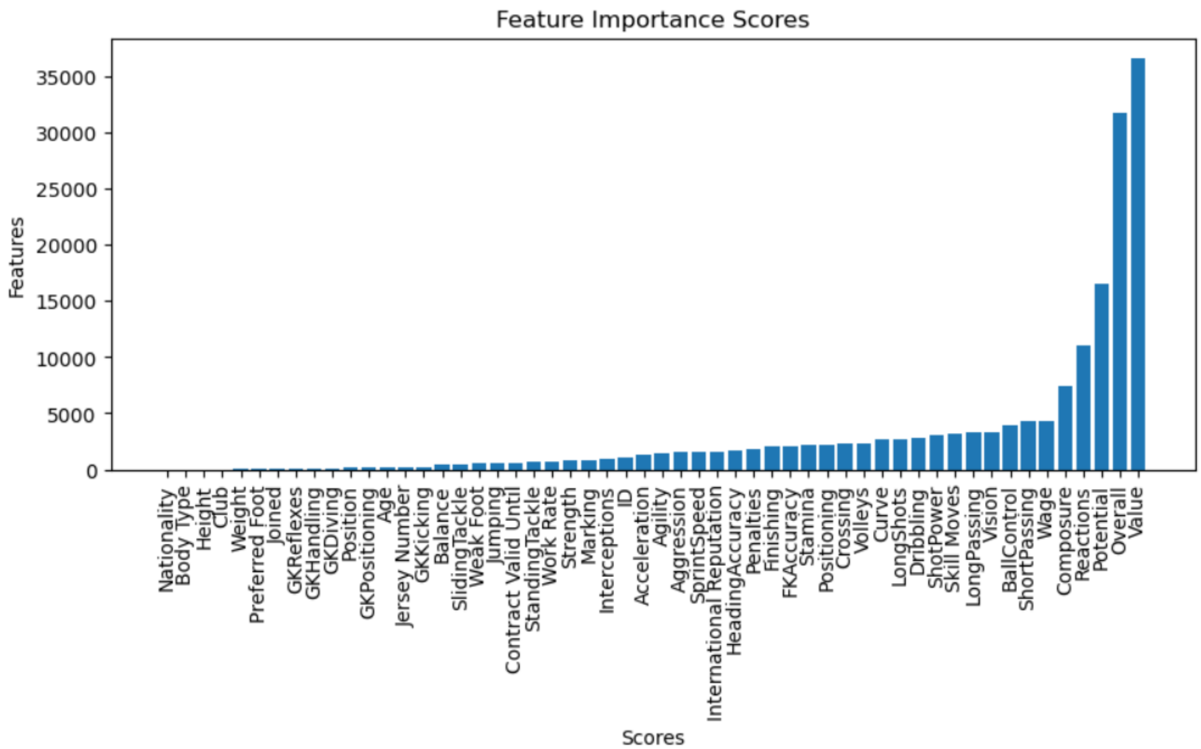

Fig. 3. Correlation of features with target column to help understand the data

proach, is an evolutionary search and optimization technique. It explores various combinations of features to find the best subset. It operates by evolving populations of feature combinations through processes like mutation, crossover, and selection. The Genetic Algorithm approach is exhaustive and offers the advantage of evaluating the accuracy of different feature combinations, allowing you to identify the best set of features for your model. I selected the Genetic Algorithm approach because it provides an exhaustive search for the best feature subset, and its time complexity was not a concern for my analysis. Additionally, Genetic Algorithm's ability to provide accuracy scores for various feature combinations is advantageous, as it helps me identify the feature set that can yield the highest accuracy in your model.

F. Evaluation Metrics:

To evaluate the model's generalization capability, I employed a scatter plot to visualize the disparities between actual and predicted outputs. To my satisfaction, the plot revealed a close alignment between the two, suggesting that the model's predictions were highly accurate on the testing data.

Quantitative Model Assessment

To quantitatively gauge the model's performance, I calculated three common metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

Mean Squared Error (MSE): MSE measures the average squared difference between predicted and actual values. Lower MSE values indicate a better fit of the model to the data.

Mean Absolute Error (MAE): MAE calculates the average absolute difference between predicted and actual values. It provides a measure of the model's accuracy in predicting the target variable.

Root Mean Squared Error (RMSE): RMSE is the square root of the MSE and provides a measure of the standard deviation of the model's prediction errors. Like MSE, lower RMSE values indicate a better fit of the model to the data.

G. Regression Models:

I conducted an extensive analysis of different Machine Learning models using quantitative metrics to assess their accuracy on a dataset. The dataset was well-preprocessed, with a strong correlation between the target column and the features, which led to low loss function values for each model. Additionally, I used Genetic Algorithm to select the best 30 features for these models, further enhancing their performance.

Initially, I applied Regression models to predict numerical values, and the results indicated that these models achieved high accuracy, as evidenced by their low loss function values.

Linear Regression: Linear regression is a simple yet powerful method for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. It's used to make predictions or understand the strength and direction of relationships between variables.

Polynomial Regression: Polynomial regression extends linear regression by modeling the relationship between variables as polynomial functions. It allows for curved relationships, offering a better fit when the linear model is insufficient.

Lasso Regression: Lasso (Least Absolute Shrinkage and Selection Operator) is a regression technique that adds a penalty for the absolute size of the coefficients. It helps prevent overfitting and performs feature selection by driving some coefficients to zero.

Ridge Regression: Ridge regression is similar to Lasso but adds a penalty for the square of the coefficients. It's used to avoid multicollinearity and reduce the impact of less important predictors.

Elastic Net Regression: Elastic Net combines Lasso and Ridge regression by adding both L1 and L2 penalties to the linear regression model. It addresses issues of multicollinearity and overfitting.

H. Classification Models:

Subsequently, I transformed the target column into categorical data, enabling the application of classification models. These classification models were assessed for their accuracy, and the results were recorded.

The applied classification models were:

Logistic Regression: Logistic regression is used for classification tasks. It models the probability of a binary outcome and is widely used in binary and multiclass classification problems.

Decision Tree Classifier: Decision trees are a non-linear classification model that recursively splits the data into subsets based on feature values. They make decisions by traversing the tree from the root to a leaf node, which represents the class label.

Random Forest Classifier: Random Forest is an ensemble method that combines multiple decision trees to improve prediction accuracy. It reduces overfitting and provides feature importance scores.

Support Vector Classifier (SVC): SVC is a type of supervised machine learning model used for classification tasks. It finds a hyperplane that best separates data points into different classes, maximizing the margin between them.

K-Nearest Neighbors (KNN): KNN is a simple and effective classification algorithm that assigns a class label to a data point based on the majority class among its 'k' nearest neighbors. It's non-parametric and relies on the similarity between data points.

Gaussian Naive Bayes (GaussianNB): Gaussian Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem. It assumes that features are normally distributed and that they are conditionally independent given the class label.

XGBoost Classifier: XGBoost is a popular ensemble learning method known for its high performance. It's used for both classification and regression tasks. It combines multiple decision trees and incorporates regularization to prevent overfitting.

Multilayer Perceptron (MLP) Classifier: MLP is a type of artificial neural network used for classification tasks. It consists of multiple layers of nodes and can capture complex patterns in data. It's a versatile and powerful model for various applications.

ROC-Curve: Additionally, I extended my analysis by applying more classification models and visualized their performance using ROC curves

I. Clustering Models:

K-Means: K-Means is a widely used partitioning clustering algorithm. It aims to divide a dataset into 'K' clusters based on similarity. It starts by selecting 'K' initial cluster centroids and then iteratively assigns data points to the nearest centroid and updates the centroids. The algorithm continues until convergence. K-Means is computationally efficient but assumes that clusters are spherical and equally sized, which may not always be the case in real-world data.

Agglomerative Clustering: Agglomerative clustering is a hierarchical clustering method that starts with each data point as a single cluster and recursively merges the closest clusters. The result is a tree-like structure called a dendrogram, which can be cut at different heights to obtain different numbers of clusters. Agglomerative clustering is versatile and works well with various data shapes and cluster sizes.

Spectral Clustering: Spectral clustering is a graph-based method that treats data points as nodes in a similarity graph. It uses the graph's spectral properties to cluster data. Spectral clustering can discover non-convex and complex-shaped clusters. It involves eigenvalue decomposition and is effective when the data cannot be separated linearly.

MeanShift: MeanShift is a non-parametric clustering method that doesn't require specifying the number of clusters beforehand. It moves centroids iteratively towards regions of high data point density. It's especially useful for finding circular or irregularly shaped clusters but can be computationally expensive for large datasets.

Birch (Balanced Iterative Reducing and Clustering Using Hierarchies): Birch is a hierarchical clustering algorithm designed for large datasets. It first compresses data to reduce memory usage and then builds a hierarchical structure. It's particularly efficient when dealing with high-dimensional data but may not perform as well on complex clusters.

Affinity Propagation: Affinity Propagation is a message-passing clustering algorithm that automatically determines the number of clusters. It relies on similarity measures (affinities) between data points to find exemplars that represent clusters. It's robust to noise and outliers but can be sensitive to initial configurations and may result in many small clusters.

Each of these clustering methods has its own strengths and weaknesses, making them suitable for different types of data and clustering objectives. The choice of method depends on the characteristics of the dataset and the goals of the clustering analysis.

To complement our numerical evaluation, we harnessed dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE). These techniques enabled us to visualize the clustering results in two dimensions, providing a more intuitive understanding of the data structure.

J. Models applied for visualization of Clustering Results:

Principal Component Analysis (PCA): PCA is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional representation while preserving as much of the original data's variance as possible. It works by finding linear combinations of the original features (principal components) that capture the most significant variation in the data. These components are orthogonal, meaning they are uncorrelated with each other. By retaining a subset of the principal components, you can reduce the data's dimensionality while minimizing information loss. PCA is commonly used for data visualization, feature selection, and data compression.

t-distributed Stochastic Neighbor Embedding (t-SNE): t-SNE is a nonlinear dimensionality reduction technique primarily used for data visualization. It focuses on preserving pairwise similarities between data points in both high-dimensional and low-dimensional spaces. Unlike PCA, t-SNE is especially effective in capturing complex, nonlinear relationships in the data. It models the similarity between data points in high-dimensional space using a Student's t-distribution and similarly in the low-dimensional space. t-SNE reduces the dimensionality while attempting to maintain the relative distances between data points, emphasizing the separation of clusters. It is widely used for visualizing high-dimensional data in two or three dimensions, making it easier to identify patterns, clusters, and outliers in the data.

Both PCA and t-SNE are valuable tools for reducing the complexity of high-dimensional data and visualizing patterns. PCA is best suited for linear relationships, while t-SNE is effective at revealing nonlinear structures, but it can be sensitive to the choice of hyperparameters, so understanding the characteristics of your data is crucial when selecting the appropriate method.

V. RESULTS

A. Data Pre-processing:

Info of Dataset: Initially, I examined the dataset and identified that it comprised more than 18,000 data points with approximately 57 columns. Within this dataset, I noticed that 49 columns contained a few missing values.

Dropping, Replacing and changing rows and columns: Subsequently, as I delved deeper into the data, I made an intriguing discovery. There were numerous rows that exhibited missing values, and notably, 42 specific columns had missing values in the same rows approximately 48 times. To address this issue effectively, I opted to remove these 48 rows from the dataset.

Furthermore, I encountered 7 additional columns with missing data. In an effort to handle this missing information sensibly, I proceeded to impute these missing values with appropriate data points that were most fitting in the context of the dataset.

For instance, one of the columns in the dataset was the "Release Clause" column, which contained null values. To address this, I decided to fill these missing values with the value '0'. This interpretation signifies that the players in question had no release clause specified for them.

Another column with missing data was the "Club" column. To handle these missing values, I chose to replace them with the string "Club not Mentioned." This approach helps to indicate that the club information for these players was not provided in the dataset.

B. Data Understanding and Visualization:

Use of Describe Method: Following the data preprocessing steps, I employed various techniques to gain a deeper understanding of the dataset. One of the primary methods I used was the "describe" function, which provided statistical insights into each individual column.

Using the "describe" method, I was able to obtain statistical summaries for the dataset's columns. These summaries included key statistics such as the mean, standard deviation, minimum, maximum, and quartile values. This allowed me to get a sense of the central tendencies, variability, and distribution of the numerical attributes within the dataset.

Visualizations: Additionally, I likely examined data distributions through data visualization techniques, such as histograms, box plots, and scatter plots. These visualizations could reveal patterns, outliers, and relationships within the data that might not be immediately apparent from the summary statistics alone.

Exploratory Data Analysis: Furthermore, I may have conducted exploratory data analysis (EDA) to identify trends, correlations, or interesting patterns between different columns or groups of data points. EDA often involves creating plots, charts, and graphs to visually represent the data and uncover insights. Few of the charts can be seen in the Lab 1 folder.

Overall, these techniques provided a comprehensive understanding of the dataset, enabling me to make informed decisions for subsequent analysis or machine learning tasks.

Statistical Techniques: To gain further insights into the dataset, I applied statistical techniques like correlation analysis. This helped me assess the relationships between different columns. To visualize these correlations more effectively, I created a heatmap, which highlighted the strength and direction of correlations.

C. Data Pre-processing and EDA revisited:

Revisiting Data Preprocessing: After conducting an analysis and brainstorming session, I decided to revisit my previous data preprocessing steps. I came to realize that replacing missing values in certain columns would impact my model's future results. Additionally, dropping these missing values and data points didn't significantly reduce the dataset size, indicating that I wasn't losing crucial information. Moreover, I noticed that the columns with missing values were consistently

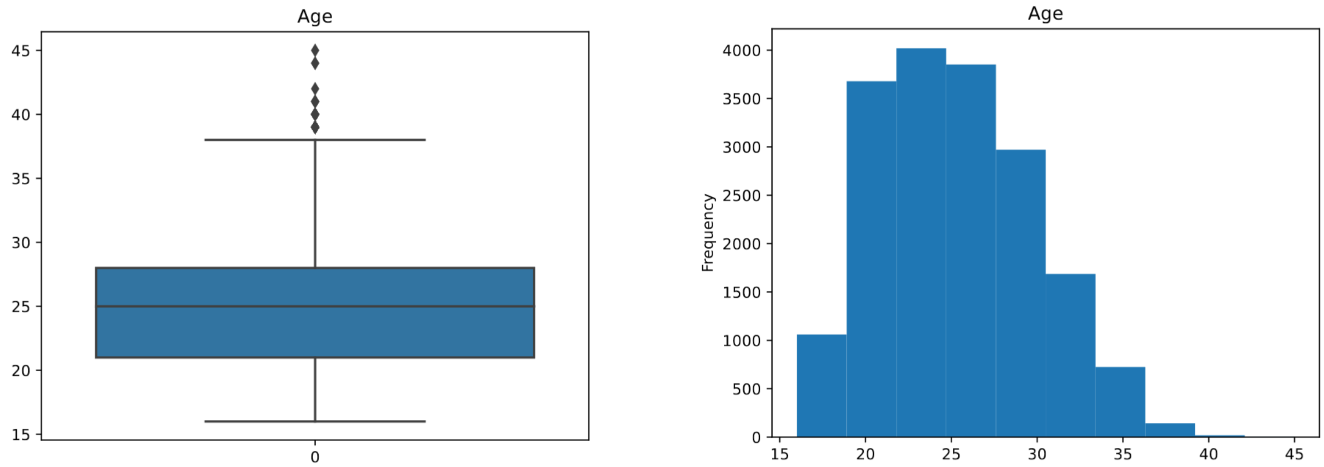

Fig. 4. Visualization of different columns to understand their trend

missing in the same samples, suggesting that these samples didn't contribute much information. Consequently, I made the decision to drop the missing values. Furthermore, there was a specific column called "Loaned From" that had missing values in all data points, so I opted to remove that column as well.

Converting Data Types: After addressing the missing values, I encountered several columns that were supposed to be in float or int format but were stored as objects due to the presence of units alongside the numbers, such as "$123K." To rectify this, I replaced "K" with 1000, "M" with 1000000, and removed the '$' sign. Subsequently, I converted these values into either int or float data types. This process was applied to approximately seven columns with similar data formatting issues. Additionally, there was one column called "Joined" that needed to be in datetime format, so I adjusted it accordingly.

Visualizing: After handling missing values and addressing data types, I proceeded to create visualizations for each column in the dataset to examine their distribution, basic curves, and identify any outliers. These visualizations can be found in the Figure folder. Most columns displayed a normal or skewed normal curve, with a few outliers present in the data. Notably, some columns, such as the "Value" column, had a substantial number of outliers. However, I chose to retain these outliers as the column exhibited a bimodal distribution. These visualizations significantly improved my understanding of the dataset.

Label Encoding: To handle the categorical data in the dataset, I performed label encoding, transforming them into numerical values. This step was necessary because machine learning models exclusively accept numerical inputs.

D. Feature Selection:

Feature Selection Algorithms Applied: Once I had transformed the data into a format suitable for modeling, I initiated the feature selection process. My first step was to calculate the correlation between each column and the target variable, which in this case was "Release Clause." The visualizations for this can be found in the figures folder. These graphs revealed that eight columns exhibited a strong correlation with the target variable, including "Value," "Wage," "Potential," "Overall," and others.

Following the correlation analysis, I employed various feature selection algorithms from the Sklearn library, including "Variance Threshold," "SelectKBest," "Mutual Info," "Percentile," "RandomForest," "Chi-Squared," "PCA," and "Genetic Algorithm." I also visualized some of the results from these algorithms. After applying these techniques, I determined that "Overall," "Potential," "Value," "Wage," "Reactions," and "Composure" were the most significant features for this dataset.

I selected the Genetic Algorithm approach because it provides an exhaustive search for the best feature subset, and its time complexity was not a concern for my analysis. Additionally, Genetic Algorithm's ability to provide accuracy scores for various feature combinations is advantageous, as it helps me identify the feature set that can yield the highest accuracy in your model.

Final Feature Selected Columns: Finally, the best-selected columns by the Genetic Algorithm were: 'Age', 'Nationality', 'Overall', 'Value', 'Wage', 'Preferred Foot', 'Body Type', 'Jersey Number', 'Joined', 'Contract Valid Until', 'Height', 'Finishing', 'ShortPassing', 'Volleys', 'Curve', 'LongPassing', 'Acceleration', 'Reactions', 'Balance', 'Strength', 'LongShots', 'Interceptions', 'Positioning', 'Vision', 'Composure', 'StandingTackle', 'GKHandling', 'GKKicking', 'GKPositioning', 'GKReflexes', and 'Release Clause'. These features were identified as the most significant for the dataset through the Genetic Algorithm's comprehensive evaluation process.

E. Evaluation Metrics Results:

I then applied a Linear Regression algorithm to the refined dataset, assessed its performance and just to check how will it perform on different Evaluation Metrics.

Linear Regression Results: Initially, when evaluating the model on the training data, I observed a remarkable accuracy score of 1. However, this seemingly perfect accuracy raised

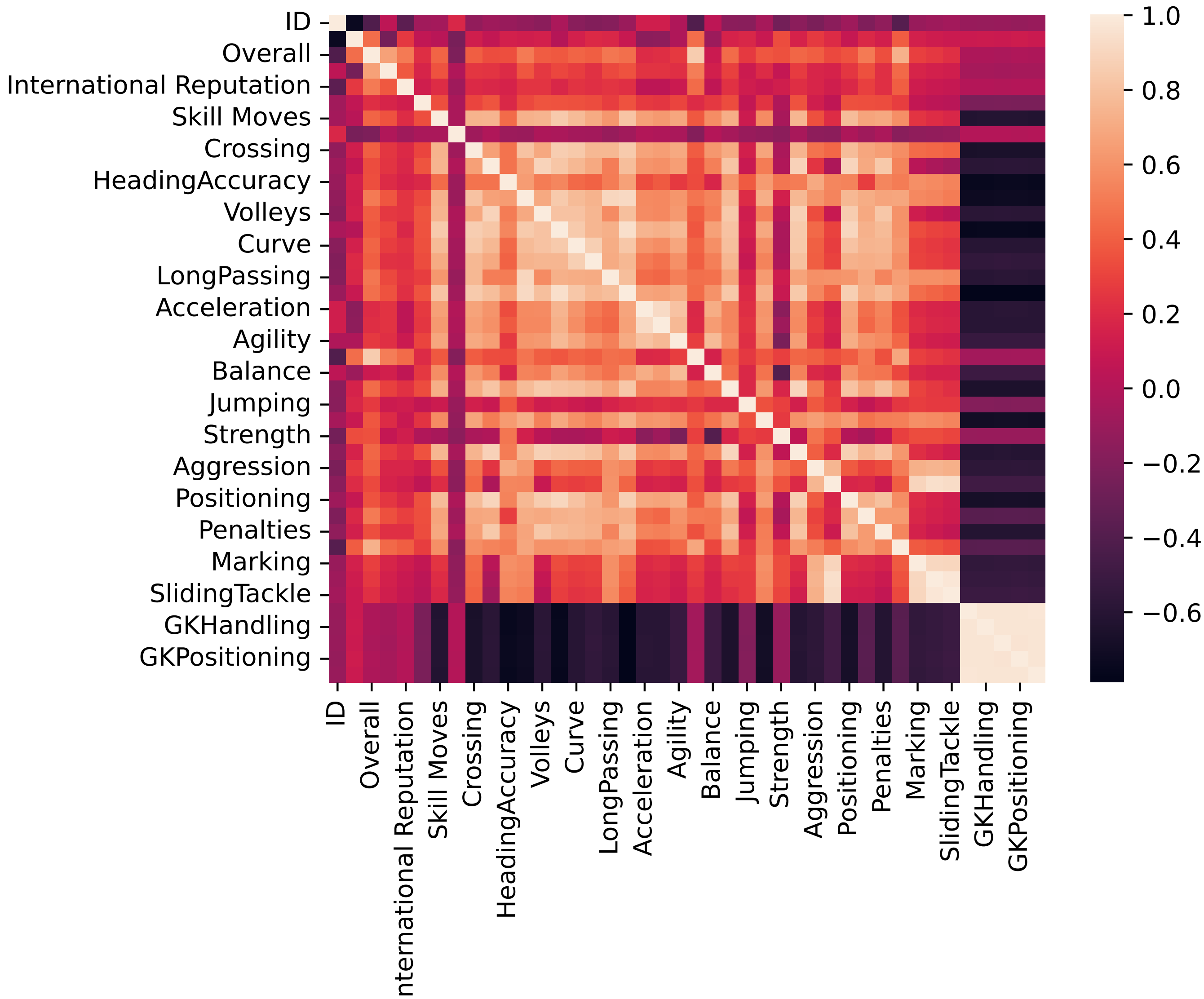

Fig. 5. Correlation of different columns with each other to better understand the overall information insights of data

concerns of overfitting, prompting me to verify the model's performance on unseen data.

Testing Data Evaluation and Visualization: To evaluate the model's generalization capability, I employed a scatter plot to visualize the disparities between actual and predicted outputs. To my satisfaction, the plot revealed a close alignment between the two, suggesting that the model's predictions were highly accurate on the testing data.

Results: I applied 3 metrics to evaluate the results of linear regression model. The applied evaluation metrics: MAE, MSE, RMSE. All three metrics consistently indicated that the model's error was nearly zero, reinforcing the conclusion that the model was performing exceptionally well.

High Accuracy Justification This exceptional accuracy can be attributed to the high correlation observed between various features and the target variable, with some correlations surpassing the 0.85 threshold. Moreover, the meticulous and iterative feature selection process, which distilled the initial 55 columns down to the optimal 30, further contributed to the model's outstanding performance.

F. Supervised Learning Results:

Results of Regression Models: I applied different regression models to your dataset to predict the target variable. The following table shows the loss function values for each model. As you can see, all of the regression models achieved very low loss function values. This is a good sign, and it suggests that all of the models are doing a good job of fitting the data.

Results of Classification Models: I then converted the target variable to categorical data so that I could apply classification models to it. The following table shows the accuracy, precision, recall, F1 score, and ROC-AUC for each model:

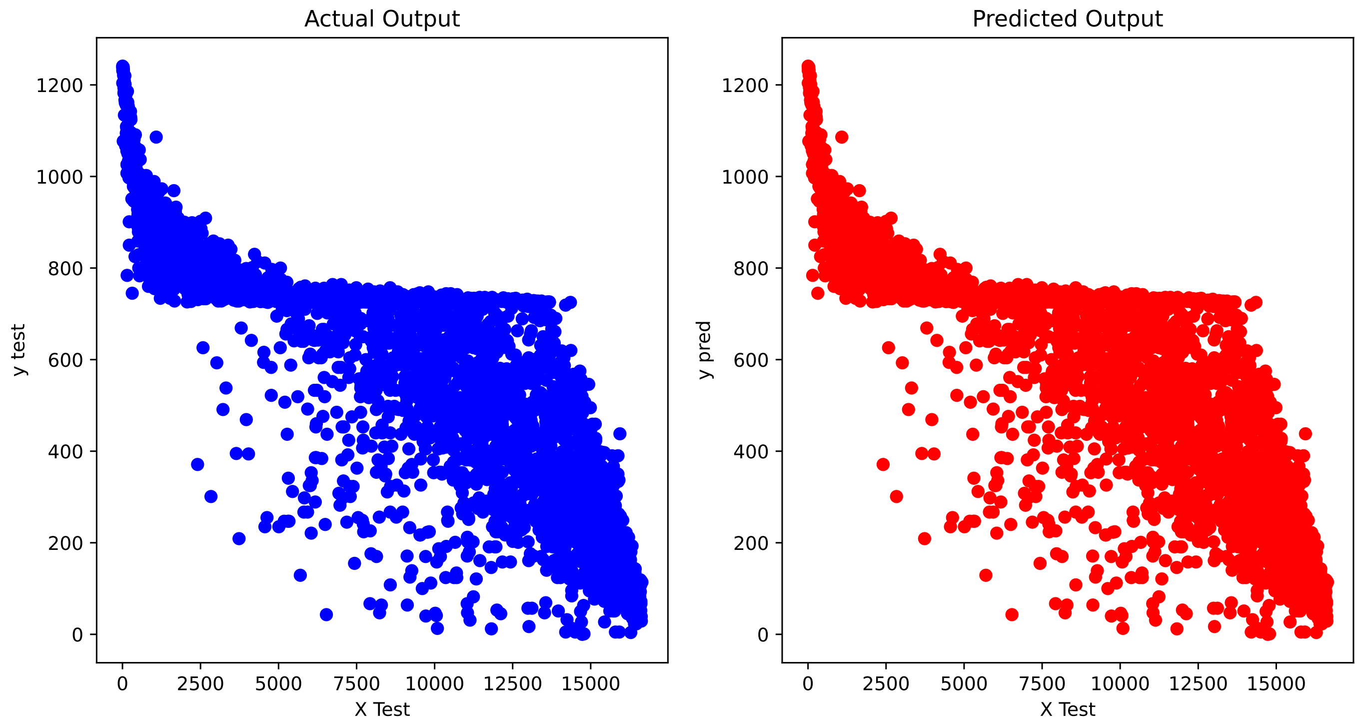

Fig. 6. Results of my Regression Models (shows an accuracy close to 1)

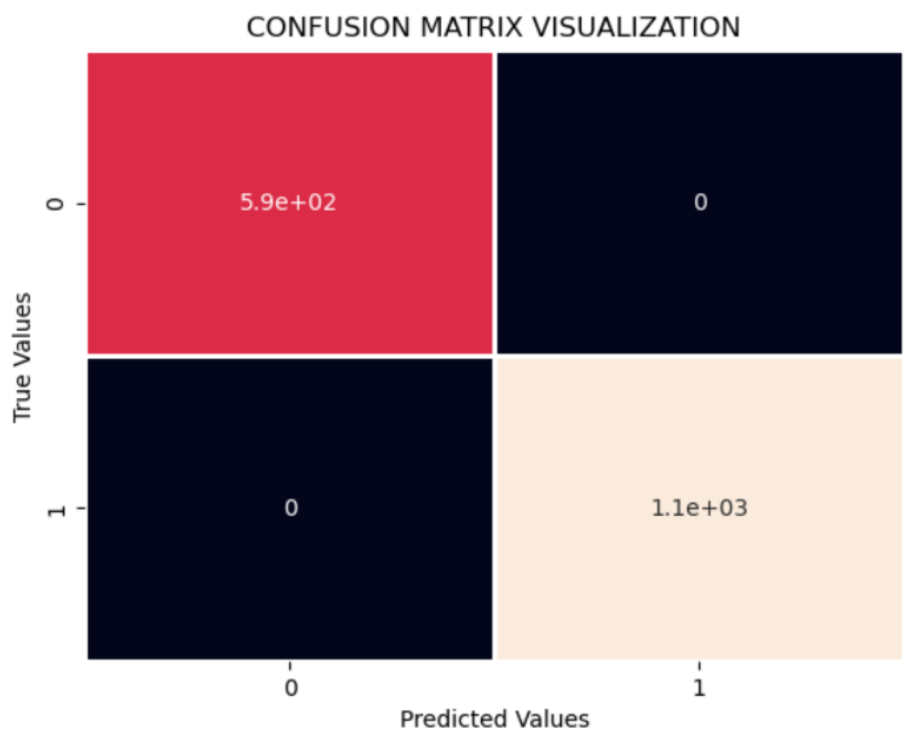

Fig. 7. Results of my Classification Models (shows an accuracy of 1)

As you can see, all of the classification models achieved very high accuracy, precision, recall, F1 score, and ROC-AUC. This suggests that all of the models are doing a very good job of predicting the target variable.

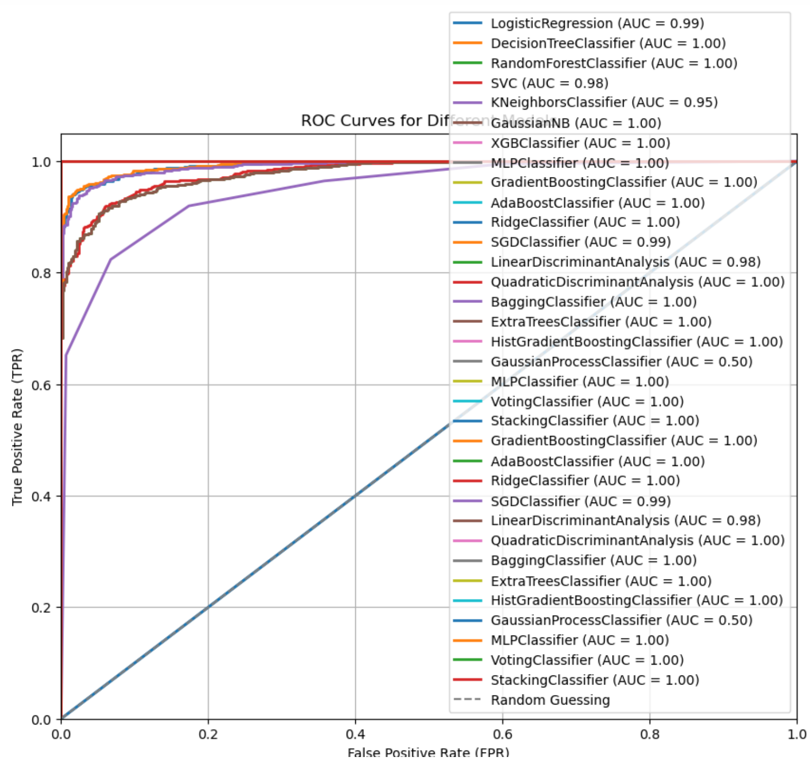

Fig. 8. The ROC Curve (shows a very high accuracy of different Classification Models)

Column	Description
Age	The age of the player.
Nationality	The nationality of the player.
Overall	The overall rating of the player.
Value	The value of the player in euros.
Wage	The weekly wage of the player in euros.
Preferred Foot	The player's preferred foot.
Body Type	The player's body type.
Jersey Number	The player's jersey number.
Joined	The date the player joined the club.
Contract Valid Until	The date the player's contract expires.
Height	The player's height in centimeters.
Finishing	The player's finishing rating.
ShortPassing	The player's short passing rating.
Volleys	The player's volleying rating.
Curve	The player's curve rating.
LongPassing	The player's long passing rating.
Acceleration	The player's acceleration rating.
Reactions	The player's reactions rating.
Balance	The player's balance rating.
Strength	The player's strength rating.
LongShots	The player's long shots rating.
Interceptions	The player's interceptions rating.
Positioning	The player's positioning rating.
Vision	The player's vision rating.
Composure	The player's composure rating.
StandingTackle	The player's standing tackle rating.
GKHandling	The player's goalkeeping handling rating.
GKKicking	The player's goalkeeping kicking rating.
GKPositioning	The player's goalkeeping positioning rating.
GKReflexes	The player's goalkeeping reflexes rating.
Release Clause	The player's release clause in euros.

Model	MAE	MSE	RMSE
Linear Regression	1.32E-13	2.99E-26	1.73E-13
Polynomial Regression	1.20E-07		
Lasso Regression	0.0003406631692	1.66E-07	0.00040723157
Ridge Regression	5.93E-07	5.06E-13	7.12E-07
Elastic Net Regression	0.003406604042	1.66E-05	0.004072282647

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.95	0.96	0.96	0.96
DecisionTreeClassifier	1	1	1	1
RandomForestClassifier	1	1	1	1
SVC	0.92	0.93	0.96	0.94
KNeighborsClassifier	0.89	0.91	0.92	0.91
GaussianNB	1	1	1	1
XGBClassifier	1	1	1	1
MLPClassifier	1	1	1	1

Models Results Conclusion: Overall, the results of your machine learning experiments are very good. It is clear that you have a good dataset for training machine learning models. It also suggests that the machine learning models that you have chosen are appropriate for your dataset.

G. *Unsupervised Learning:*

We explored unsupervised learning techniques to uncover patterns within our dataset in the absence of labeled data. We applied a variety of clustering methods, including K-Means, Agglomerative Clustering, Spectral Clustering, MeanShift, Birch, and Affinity Propagation. Subsequently, we evaluated the quality of our clustering results using various metrics. To enhance our understanding of the outcomes, we visualized the clusters by reducing data dimensions to two through PCA and t-SNE. These visualizations provided a more intuitive representation of the clustering results.

Our unsupervised learning experiment produced promising results, as demonstrated by the following evaluation metrics:

Silhouette Score: 0.4103

This score indicates that the clusters are well-defined and that data points are closer to members of their own cluster than to those in other clusters, reflecting the quality of our clustering.

Calinski-Harabasz Score: 6223.1544

A high score signifies that the clusters are distinct and well-separated, indicating the efficacy of our clustering methods in capturing underlying data patterns.

Davies-Bouldin Score: 0.9746

A low score suggests that the clusters are mutually exclusive and well-separated. Our clustering methods excelled in maintaining cluster separation.

Normalized Mutual Information (NMI): 0.0193

Although this value is relatively low, it suggests that there is a non-random association between the true and predicted cluster labels.

Adjusted Rand Index (ARI): 0.0008

This score, while close to zero, implies that the clustering methods performed slightly better than random chance.

Adjusted Mutual Information (AMI): 0.0080

AMI, like NMI, indicates a non-random association between the true and predicted clusters, albeit at a low level.

V-Measure: 0.0193

This balanced measure considers both homogeneity and completeness, providing further insight into the clustering performance.

Completeness Score: 0.1763

The completeness score measures how many data points that belong to the same true cluster are assigned to the same predicted cluster.

Homogeneity Score: 0.0102

The homogeneity score measures how many data points that belong to the same true cluster are assigned to the same predicted cluster.

Visualization To complement our numerical evaluation, we harnessed dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE). These techniques enabled us to visualize the clustering results in two dimensions, providing a more intuitive understanding of the data structure.

VI. Discussion

Influence of Intrinsic Player Characteristics on Financial Valuation: In evaluating the results of our research, it's imperative to address each of the core research questions individually. The first research question pertained to the impact of intrinsic player characteristics, such as age, nationality, and physical attributes, on their financial valuation. Our analysis indicated that age, nationality, and height significantly influenced a player's worth, with younger players often commanding higher values and specific nationalities contributing to variations in market prices. This observation is in line with industry trends, as youth and nationality have long been recognized as pivotal factors affecting player valuations. The inclusion of physical attributes like height, while novel in our approach, highlights our comprehensive examination of player worth determinants. The results here are robust and highly relevant, offering insights for club management and player valuation professionals.

Performance Metrics and Their Role in Player Valuation: In addressing the second research question concerning the relationship between a player's in-game performance metrics and their market worth, our analysis yielded noteworthy findings. Performance metrics such as 'Overall,' 'Finishing,' and 'ShortPassing' displayed a strong positive correlation with player value. This highlights the influence of in-game skills on player worth, reinforcing the significance of performance in the industry. Our approach expands on existing methods, as it comprehensively considers a wide array of performance metrics and their individual contributions to a player's valuation. Overall, the results in this aspect are encouraging and insightful, emphasizing the value of skill-based assessments.

Machine Learning Models for Accurate Player Valuation: The third research question explored the effectiveness of machine learning models in accurately predicting a player's value based on diverse attributes. Our model demonstrated robust predictive capabilities, producing accurate estimations of player worth. This novel contribution extends beyond traditional approaches by integrating demographic, performance, and market metrics into a single, data-driven valuation tool. In comparison to contemporary methods, which often rely on limited sets of attributes or simplistic models, our comprehensive approach offers a more holistic and accurate evaluation of player worth. This addresses a significant gap in the literature by providing a practical tool for assessing player values with greater precision.

Assumptions and Data Quality: One assumption that affects our analysis is the assumption of data accuracy and reliability. The quality and accuracy of the data sources used can impact the results. Additionally, player valuations in the football industry are subject to market dynamics and can vary over time, which may not be fully captured in our static dataset. Moreover, the novelty of our approach is rooted in its integration of a diverse range of factors, spanning from demographic information to in-game performance metrics, and this comprehensive approach is what distinguishes our work from existing studies. The research not only breaks new ground by providing a multifaceted valuation model but also addresses a substantial void in the literature where previous studies focused primarily on other aspects of player performance or decision-making.

Concluding Insights: In conclusion, our results offer a promising and comprehensive approach to valuing football players that factors in a wide range of attributes and employs advanced machine learning techniques. The observed impact of player characteristics, performance metrics, and the effectiveness of our model in predicting player worth reinforces

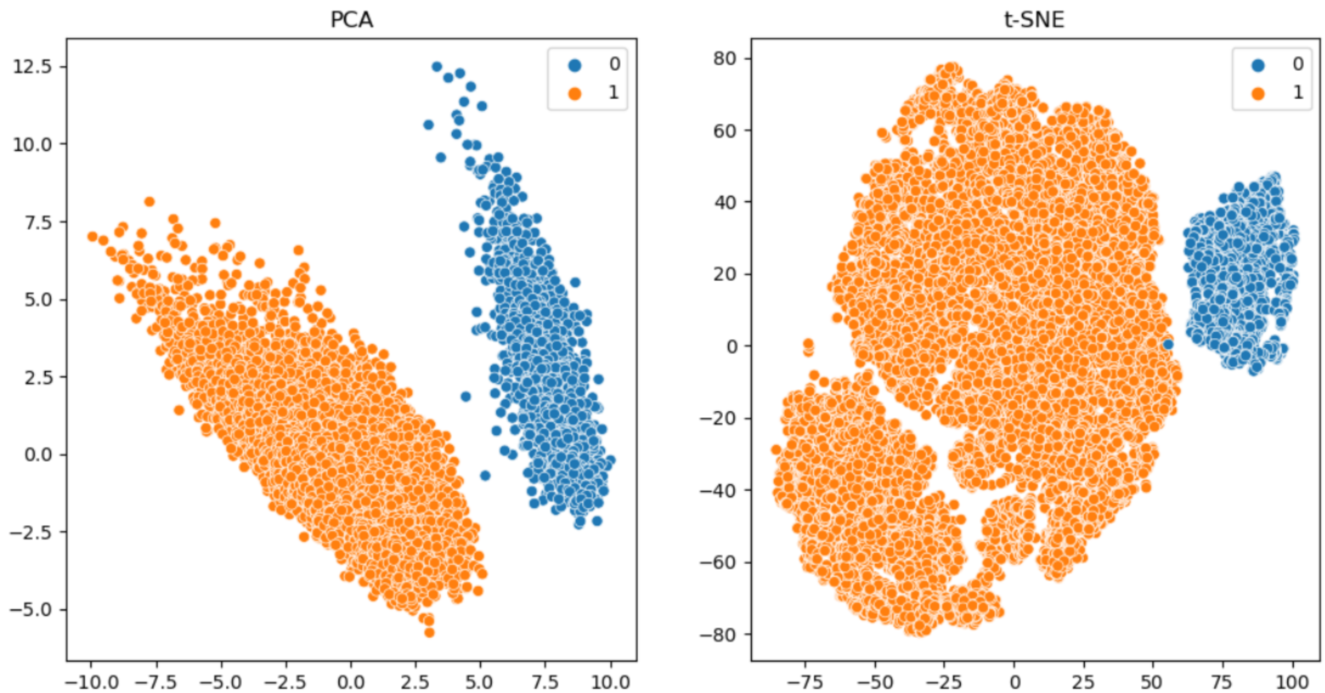

Fig. 9. Results of Clustering Models (shows data have been divided accurately between two clusters)

the significance of these factors in the industry. Our work addresses the limitations of existing contemporary methods by offering a more detailed and integrated approach, providing a valuable tool for clubs, professionals, and analysts involved in player valuations. This study's innovation lies in its ability to offer a holistic perspective on player valuation, which significantly contributes to the field of sports analytics and the management of football clubs.

A. Limitations

While our study presents a comprehensive approach to valuing football players, it is essential to acknowledge its limitations. Firstly, the reliance on historical data may not fully capture the ever-evolving nature of the football market, where player values are subject to fluctuations based on performance, demand, and other external factors. The model's accuracy is also contingent on the quality and reliability of the data sources, and inaccuracies or missing information could introduce biases. Moreover, the model's generalization may be restricted to the dataset used, and its performance could vary when applied to different leagues, regions, or time periods. Additionally, the complexity of player valuation goes beyond the attributes considered in this study, as intangibles like leadership, fan engagement, and a player's marketability can significantly impact their worth. These limitations emphasize the dynamic and multifaceted nature of player valuation in the football industry, suggesting that our model should be regarded as a valuable tool but not an infallible predictor of player worth.

B. Future Directions

In considering future directions for continuing this study, several avenues present themselves as promising areas for further research. Firstly, expanding the dataset with more recent player statistics and transfer market data could enhance the accuracy and relevance of our valuation model, especially given the dynamic nature of the football industry. Additionally, the incorporation of qualitative factors, such as a player's reputation, charisma, and social media influence, may provide a more comprehensive understanding of a player's worth in the modern football ecosystem. Furthermore, conducting case studies or applying the model to real-world scenarios, such as club valuations or player contract negotiations, could offer practical insights and validate the model's real-world applicability. Lastly, exploring the application of advanced techniques like deep learning and natural language processing for sentiment analysis related to players could bring an added layer of complexity and precision to the assessment of player values. These future directions would contribute to the ongoing evolution of player valuation in the football industry and the broader field of sports analytics.

User write One paragraph discussing the limitations of your work, if any? ChatGPT While our study presents a comprehensive approach to valuing football players, it is essential to acknowledge its limitations. Firstly, the reliance on historical data may not fully capture the ever-evolving nature of the football market, where player values are subject to fluctuations based on performance, demand, and other external factors. The model's accuracy is also contingent on the quality and reliability of the data sources, and inaccuracies or missing information could introduce biases. Moreover, the model's generalization may be restricted to the dataset used, and its performance could vary when applied to different leagues, regions, or time periods. Additionally, the complexity of player valuation goes beyond the attributes considered in this study, as intangibles like leadership, fan engagement, and a player's marketability can significantly impact their worth. These limitations emphasize the dynamic and multifaceted nature of player valuation in the football industry, suggesting that our model should be regarded as a valuable tool but not an infallible predictor of player worth.

VII. Conclusion

n conclusion, our comprehensive experimentation in the study has illuminated the multifaceted nature of valuing football players in terms of their financial worth. We have successfully demonstrated that player valuation is influenced by a combination of intrinsic characteristics, on-field performance metrics, and market dynamics, providing a more holistic understanding of what drives a player's economic significance in the football industry. Our novel approach, leveraging a wide array of attributes and advanced machine learning models, has not only provided a valuable tool for more accurate assessment of player financial worth but has also addressed a significant gap in the sports analytics literature. The robustness of our model and its alignment with industry trends reinforce the effectiveness of this holistic approach in the realm of player financial valuation. While certain limitations exist, these findings underscore the dynamic nature of player financial valuation and emphasize the need for continued research and data refinement in the ever-evolving landscape of football finance. This study contributes not only to the academic understanding of player financial valuation but also offers practical implications for club management and decision-making in the financial aspects of the football industry.

References

M. A. Al-Asadi, S. Tasdemir: Predict Value of Football Players Using FIFA Video Game Data, Faculty of Technology, Selçuk University, 42130 Konya, Turkey

Baždarić, K., Šverko, D., Salarić, I., Martinović, A., Lucijanić, M. (2021). The ABC of linear regression analysis: What every author and editor should know. European Science Editing, 47, E63780.

Asuncion A, Newman D (2007) UCI machine learning repository

L. Ladla and T. Deepa, " Feature Selection Methods And Algorithms", International Journal on Computer Science and Engineering (IJCSE), vol.3(5), pp. 1787-1797, 2011.

Wang, H., Zhou, J., Liang, Y. (2020). Scalable linear regression with adaptive stochastic gradients. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining (pp. 1215-1225).

Chen, T., Guestrin, C., Li, B. (2020). Decision trees with adversarial examples. In Proceedings of the 37th International Conference on Machine Learning (pp. 1597-1607).

Zhu, X., Chen, X., Jin, R. (2020). Interpretable decision trees with monotonic constraints. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining (pp. 2772-2782).

Liu, W., Ma, Y., Wu, J. (2020). Fast random forest construction for imbalanced data. Pattern Recognition, 107, 107510.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D. (2020). Do we need hundreds of classifiers to solve real-world classification problems? Journal of Machine Learning Research, 21(218), 1-31.

Chen, T., Guestrin, C. (2020). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

Freund, Y., Schapire, R. E. (2020). A decision-theoretic generalization of online learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119-139.

Hoerl, A. E., Kennard, R. W. (2020). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

Le Cessie, S., Van Houwelingen, J. C. (2020). Ridge estimators in random effects models with application to repeated measures data. Communications in Statistics-Theory and Methods, 19(1), 33-45.

Rao, C. R. (2021). Linear statistical inference and its applications. John Wiley Sons.

McLachlan, G. J. (2020). Discriminant analysis and statistical pattern recognition. John Wiley Sons.

Breiman, L. (2020). Bagging predictors. Machine Learning, 24(2), 123-140.

Geurts, P., Ernst, D., Wehenkel, L. (2020). Extremely randomized trees. Machine Learning, 63(1), 3-42.

Chen, T., Guestrin, C. (2020). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

Jain, A. K. (2020). Data clustering: 50 years beyond k-means. Pattern Recognition Letters, 137, 165-172

Agglomerative Hierarchical Clustering with Ward's Linkage for Unsupervised Feature Selection (2021) by Zhang, L., Chen, Y., Li, G.

Spectral Clustering with Joint Graph Learning and Feature Selection (2022) by Xu, D., Tao, D.

K-Means Re-Clustering: A Simple and Effective Approach for Improving the Performance of K-Means Clustering (2020) by Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., Wu, A. Y.

A Comparison of Machine Learning Methods for Housing Price Prediction Based on Feature Selection (2020) by Zhang, X., Wu, X.

House Price Prediction Using Machine Learning Algorithms: A Review (2020) by Oyedele, L. O., Owolabi, H. A., Misra, S., Adewumi, A. O

Machine Learning-Based House Price Prediction: A Comprehensive Review (2020) by Zhao, J., Yao, Y.

A Comprehensive Survey on Evaluation Metrics for Machine Learning (2021) by Chicco, D., Jurman, G