DATA SCIENCE FUNDAMENTALS

(TDS 2101)

REPORT

FIFA 2022 DATA PREDICTION OF PLAYERS RELEASE CLUSE

AMIN AHMED MOHAMMEDELHASSAN

<1191302190>

1191302190@student.mmu.edu.my

TABLE OF CONTENTS

INTRODUCTION:

FIFA 22 is one of the most popular video games on the world, it tries to simulate the real-world football matches by collecting its data throw testing the player's physic, analyzing their performances, and gathering their contract information.

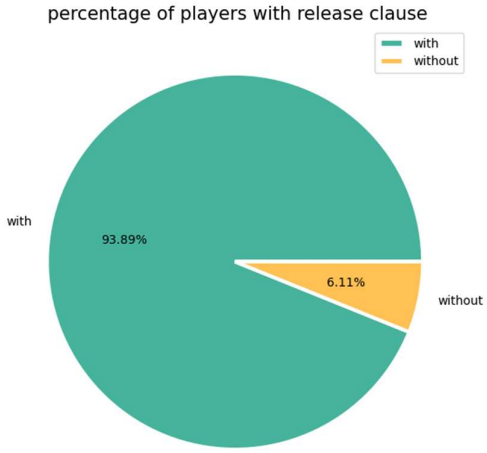
PROBLEM STATEMENT:

Release cluse is a condition in the player's contact which allow players to leave the club without its permission if that exact amount was delivered by any other club.

In this project we are trying to predict the missing value of the Release cluse for the player who doesn't have it in their contracts using multi liner regression

In the collected data some of the Release cluse value are missing, so the question is:

What is value of these missing data?



EXECUTIVE SUMMARY:

This project aims to predict the rating of FIFA 2022 players using data analysis and machine learning techniques. The dataset used in this project is the FIFA 2022 complete player dataset, which was obtained from Kaggle. The dataset includes various attributes such as player name, club, league, position, and various statistics related to the player's performance and contract.

.

DATA DESCRIPTION:

The dataset used in this project is the FIFA 2022 complete player dataset which was obtained from Kaggle. The dataset consists of 19,239 rows and 110 columns. It includes various attributes such as player name, club, league, position, and various statistics related to the player's performance, its type distributed as follow:

Type of columns	Number of columns
float64	16
int64	44
object	50

The memory usage of the data is 16.1+ MB

DATA CLEANING:

The data cleaning process involves identifying and handling missing values, removing duplicate observations. identifying and handling the outliers is not an option since the data is about player real information.

	number of null
value_eur	74
wage_eur	61
club_team_id	61
club_name	61
league_name	61
league_level	61
club_position	61
club_jersey_number	61
club_contract_valid_until	61
release_clause_eur	1176
goalkeeping_speed	17107

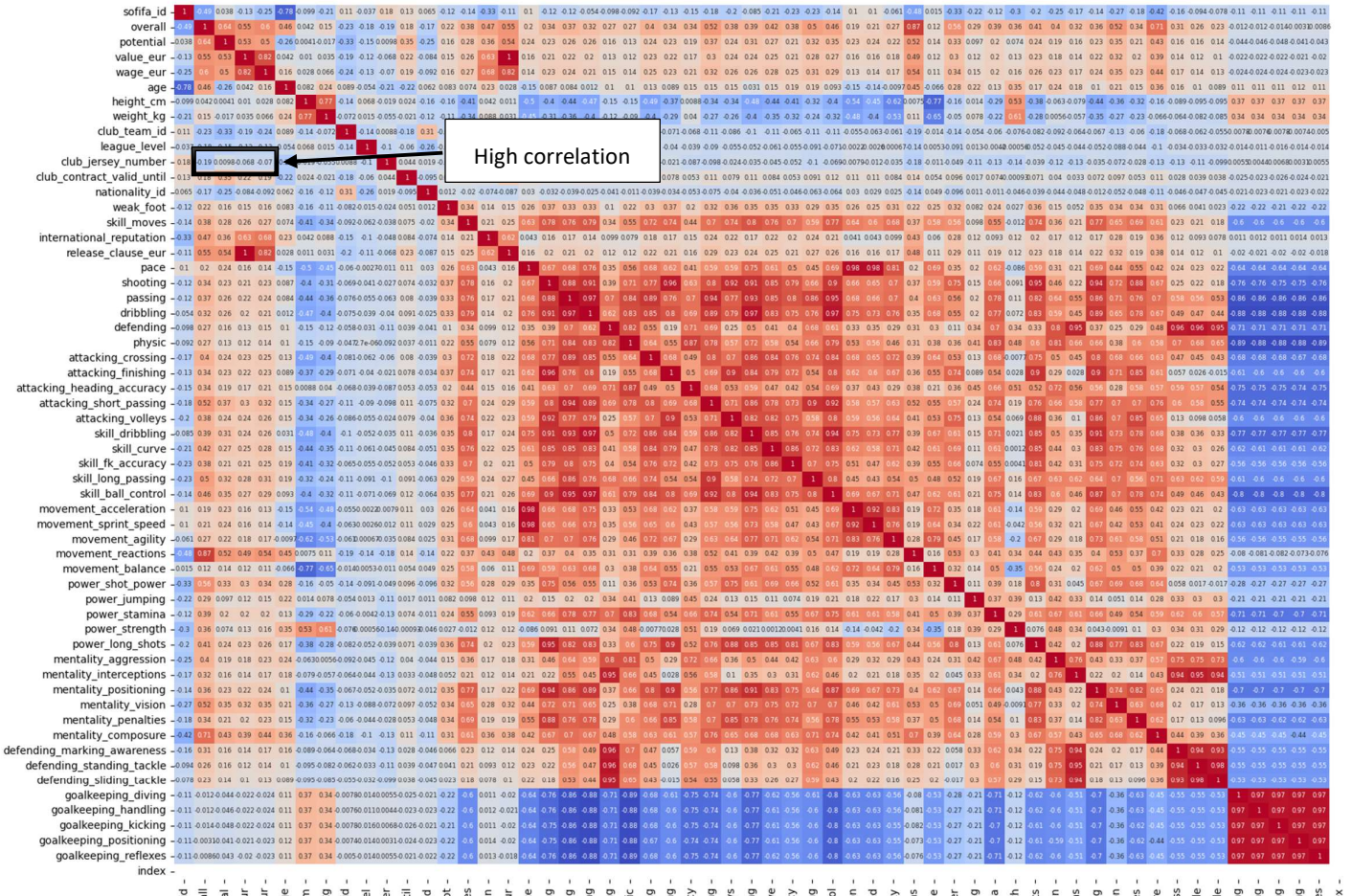
- The 61 player who did not have any club or league information were separated in a subset.
- The remaining vlaue_eur was filled with 0 assuming that the player does not have a value.
- Goalkeeping speed was solved in data transformation processes.
- Release clause is the target of the project.

DATA TRANSFORMATION:

The data transformation process involves normalizing the data and aggregating the data. The data was normalized by replacing the pace of all goalkeepers with their goalkeeping speed. The data was also aggregated by extracting league and club information from the dataset to identify who are the free players.

EXPLORATORY DATA ANALYSIS:

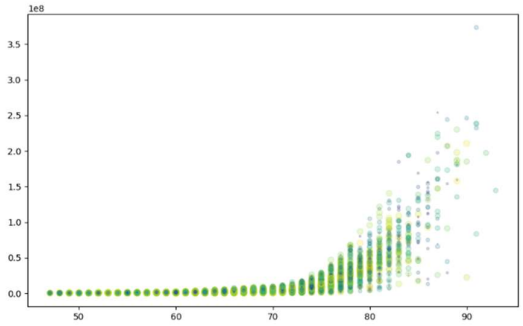
Exploratory data analysis was performed by using various visualization techniques, statistical methods, and machine learning models to summarize the dataset and identify relationships between variables. The correlation between different attributes and the release cluse of the players was analyzed using heatmaps and scatter plots. therefore, it was found that there are strong relationships between certain attributes and release cluse. The analysis also revealed that there are several players who do not have clubs, and these observations were removed from the dataset.



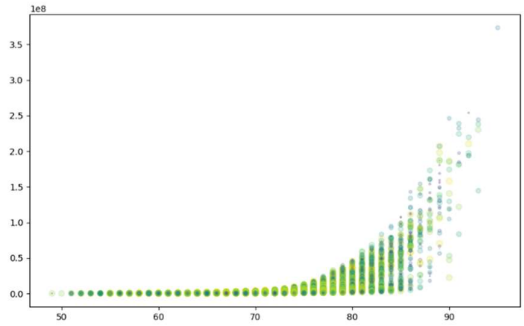
DATA MINING:

In this process we identified the high correlated columns with the target which is (overall, potential, wage, value) and discovering indirect correlation (age) by using (overall and potential).

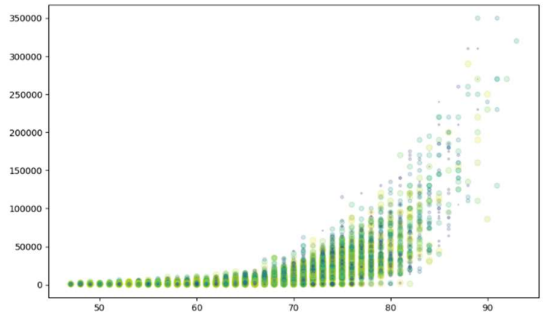
Overall vs Release cluse



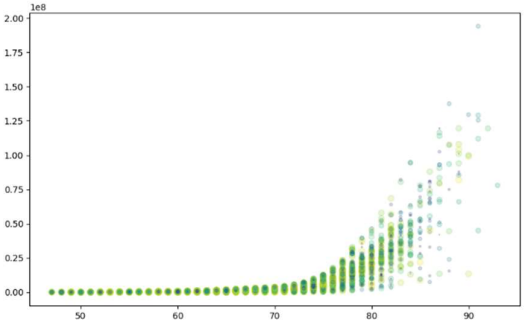
Potential vs Release cluse



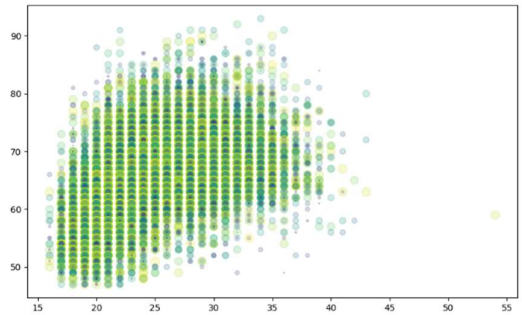
Overall vs wage



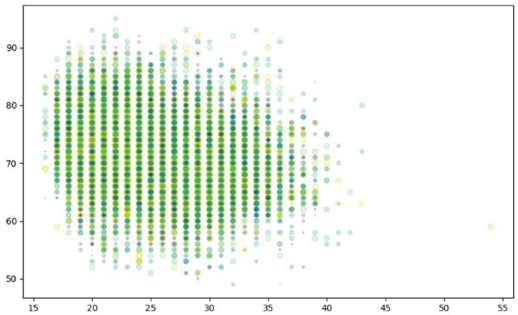
Overall vs value



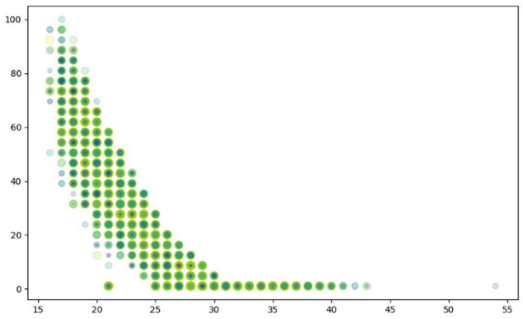
Age vs Overall



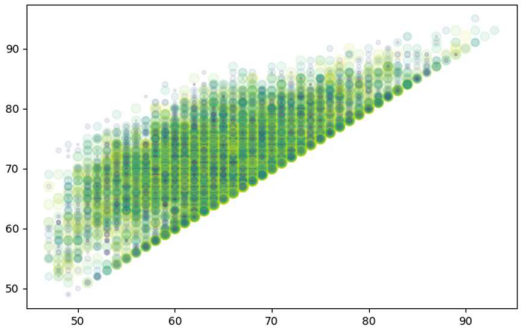
Age vs Potential



Age vs Potential diffrance (Potential - overall)



Overall vs Potential



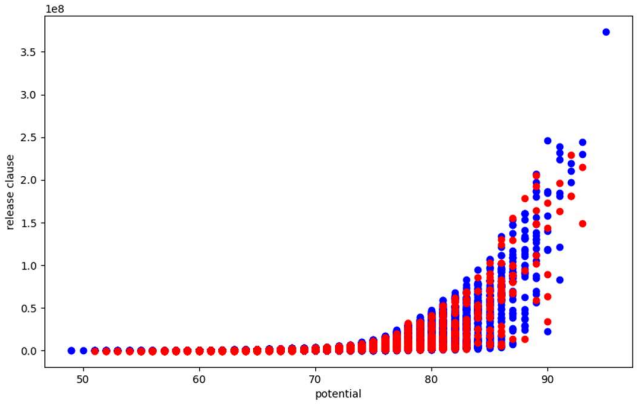
TRAINING THE MODEL:

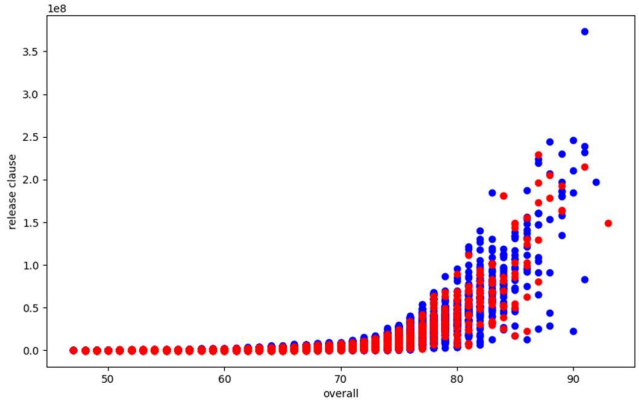
Multilinear regression was used to analyze the data and make prediction model to predict player release cluse by feeding it the 5 related columns and split the data into training and testing sets, the ratio is 8 : 2 , so 80 % for training and 20% testing.

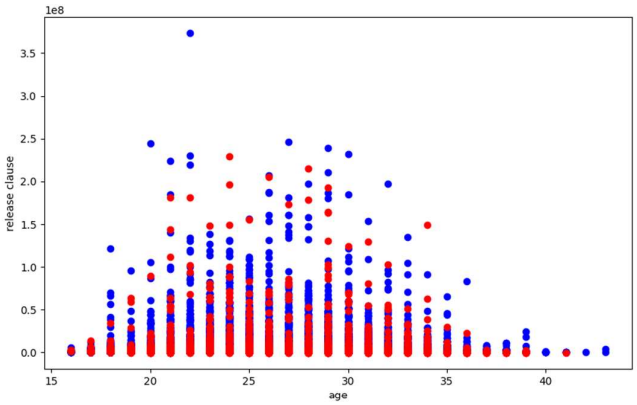
RESULTS:

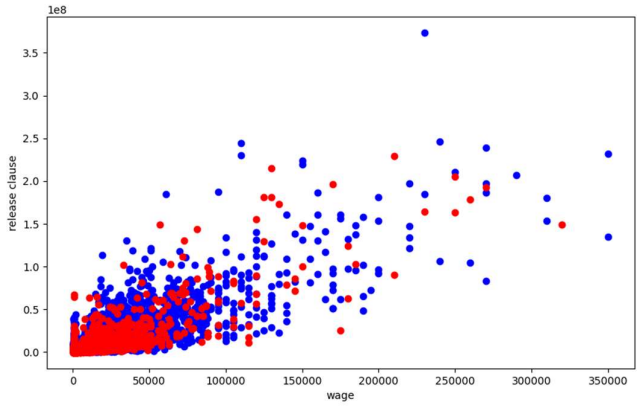
The results of this project showed that the linear regression model was able to predict the rating of the players with a high degree of accuracy. The correlation between different attributes and the release cluse of the players was also analyzed, which showed that attributes such as overall rating, potential, and age had the highest correlation with the rating of the players. The model was able to predict the rating of the players with an R-squared value of 1.0, which indicates a strong correlation between the predicted and actual values.

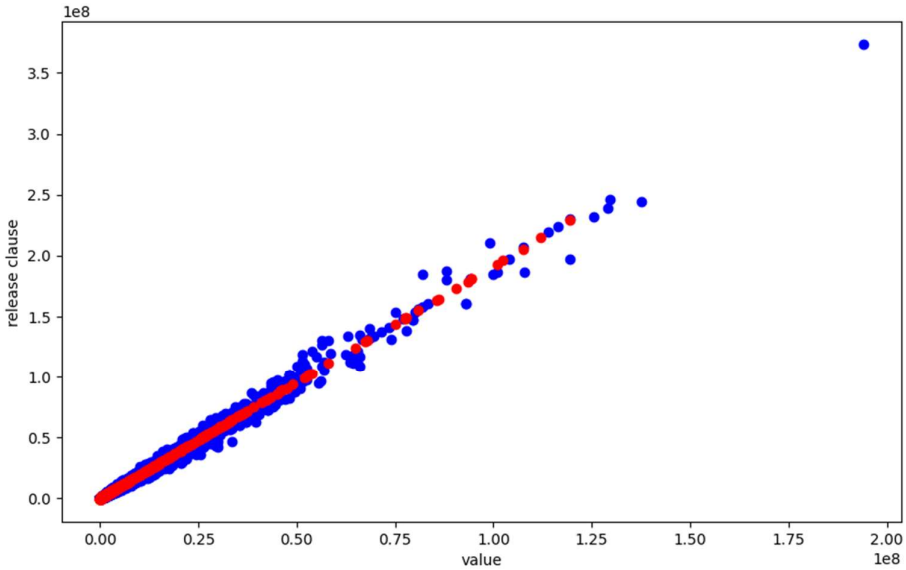
original data vs Predicted data











CONCLUSION AND FUTURE WORK:

The main conclusion of this project is that the linear regression model was able to predict the rating of the players with a high degree of accuracy. The results also showed that attributes such as overall rating, potential, and age had the highest correlation with the rating of the players. In the future, more advanced machine learning models such as Random Forest or XGBoost could be used to improve the accuracy of the predictions. Additionally, more data and attributes could be used to increase the complexity of the model and improve the predictions.

REFERENCES:

- FIFA 2022 complete player dataset, obtained from Kaggle (https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset?select=players_22.csv)