Predicting football player releases

Sriram Swaminathan, I6404161

12^{th} March 2025

1 Abstract

This project constructs a knowledge graph for football players by integrating real-world player statistics with FIFA video game data over two consecutive years (2021 and 2022). SPARQL queries were used to assess the quality of the knowledge graph and to derive inferences that would be challenging using traditional structured data. Additionally, logistic regression and a vanilla neural network were developed and compared for classifying players as released or retained. Both models achieved a similar test set accuracy of $\sim 67\%$.

2 Introduction

Football clubs and players engage in discussions about their future at the end of each season. However, most clubs (apart from a few like Liverpool, Brighton and Brentford) lack a data-driven approach to these decisions. Clubs often fail to identify weaknesses and secure suitable replacements, while agents struggle to predict layoffs and find the best opportunities for their players. These decisions are frequently influenced by short-term performances rather than long-term contributions, leading to inefficient team compositions and missed opportunities for players.
The objective of this project is to build a knowledge graph (KG) of football players enriched with FIFA video game data for two consecutive seasons of football. I will then use this KG to find insights that are hard to gain by querying structured data. I will also compare a simple logistic regression model and a vanilla neural network for classifying players as released or not. The ultimate goal is to set up a system (knowledge representation and classification model) that clubs and players can use for contract and salary negotiations, ensuring that clubs retain valuable players and that players find teams that align with their career goals.

3 Related work

In the last two years, Graph Neural Networks (GNNs) have been used in football for a different range of tasks. Jorris Bekkers and Amod Sahasrabudhe [1] used GNNs to predict successful counterattacks and tried to identify the factors that made them work in both men's and women's football. Aditya [3] on the other hand, used GNNs to automatically identify events from video footage, trying to keep up with the speed of data collection. In another application, Rakha et al. [2] used GNNs to make data-driven recommendations and predictions for football formations, assisting coaches with the tactical aspects of the game. Based on my brief literature review, it seems that graph-based approaches for football are becoming popular. From the papers mentioned above, it seems to be most applicable for on-field events that have spatial data, making a graph representation especially useful.

4 Methodology

4.1 Cleaning

For this project, I used player statistics from the 2021-2022 and 2022-2023 football seasons for players from the top five European leagues (Premier League, Ligue 1, La Liga, Bundesliga and Serie A). To enhance the dataset, I incorporated data from the FIFA video game (versions 22 and 23). The datasets were labeled as Stats 21.csv, Stats 22.csv, FIFA 21.csv, and FIFA 22.csv, since the majority of the season is played in the first year.
The reason FIFA data is useful because: (i) it provides a standardized rating (*overallrating*) for comparing players & their statistics across different leagues (ii) it includes metrics that are difficult to capture in statistics, like *skill moves, potential, work rate* etc.
Since the Stats data is only measured for players in Europe, I expected it to have far fewer players (rows) than the FIFA data, which is collected for players across the globe. On checking, I found that the FIFA data had about 19,000 and 18,500 players across both years, while the Stats data had only 2,900 and 2,700 players across both years.

The Stats data had 143 and 120 columns across both years. This was too many attributes for my analysis. I selected 14 statistics that would capture the most important aspects of a footballer's game. These were :

Property	Description	Property	Description	Property	Description
FullName	Full name	Born	Year of birth	MP	Matches played
Starts	Matches started	Min	Minutes played	90s	Matches played in 90-minute units
Goals	Goals scored	Assists	Assists provided	PasTotCmp%	Total pass completion percentage
Int	Interceptions made	Clr	Clearances made	AerWon	Aerial duels won
Touches	Total touches	PasLonCmp	Long passes completed	Fls	Fouls committed
Fld	Fouls drawn	SoT	Shots on target		

Table 1: Stats columns used

The FIFA data had 90 columns due to game design. The video game makers have to ensure that every player has a rating in every single position. This means gamers are allowed to assign any position to any player. A goalkeeper can play as a striker and vice versa. Each goalkeeper has a striker rating, and each striker has a goalkeeper rating. Generally, in real life, players stick to one of 4 broad positions: Goalkeeper, Defender, Midfielder, and Forward. Therefore, this data is just a good guess on how a player will perform in another position. This is not a feature that I need to include in my analysis. Hence, players will get only their position-specific ratings. The columns I chose to include (26) are:

Property	Description	Property	Description	Property	Description
ID	Unique sofifa ID	FullName	Full name	Height	Player's height (cm)
Nationality	Country of origin	Overall	Overall rating	Potential	Predicted improvement
BestPosition	Preferred playing position	Club	Current club	ValueEUR	Market value in Euros
ContractUntil	Contract expiration year	IntReputation	International reputation (1-5)	PreferredFoot	Dominant foot
SkillMoves	Dribbling skills (1-5)	AttackingWorkRate	Attacking effort (low, med, high)	DefensiveWorkRate	Defensive effort (low, med, high)
PaceTotal	Speed ability	ShootingTotal	Shooting ability	PassingTotal	Passing ability
DribblingTotal	Dribbling ability	DefendingTotal	Defensive ability	PhysicalityTotal	Physical build
GKDiving	Diving ability (goalkeepers)	GKHandling	Handling ability (goalkeepers)	GKKicking	Kicking ability (goalkeepers)
GKPositioning	Positioning ability (goalkeepers)	GKReflexes	Reflex ability (goalkeepers)		

Table 2: FIFA columns used

I converted the .csv files to pandas dataframes and began my data cleaning. I dropped duplicate entries from both data sources. Then I filtered out the columns identified in tables 1 and 2. I merged the 2021 Stats data & FIFA data on *FullName* and did the same for 2022. I named the merged data sets Data 21.csv and Data 22.csv, respectively. I found the players common to both Data 21.csv and Data 22.csv and kept only those players in the .csv files. In the end there were 1,145 players whose data was measured across both years. It is important for me to have data from both years since I am building a model that classifies players who were dropped after one season, players available in only one season won't help my model. The exact code used to do this can be found in the code/clean.ipynb file.

4.2 Building

I defined one main class for all players, Football Player. Since I had to store the values for the same attributes over 2 separate years, I chose to define the range for some of my properties to be rdf:Seq. This class can store multiple values in a manner where the numerical ordering of the values is significant. Properties that don't change over 2 years (static properties) had a Literal as their range (*name, nationality, height, position, prefferred foot, position, birthyear*). So all properties had a node of type football player as their domain and either an rdf:Seq or a Literal as their range.
I used 4 main namespaces for this project. (i) sofifa (ii) wikidata(iii) dbpedia (iv) sportsschema. The namespaces and the properties borrowed from them are visualized here. I created a local namespace called https://footballerontology.com/ and mapped the previously defined properties and classes to their equivalent fb:property, simplifying queries within the knowledge graph.
Each player does not have all properties assigned to them. In particular, the properties with a *GK* prefix in table 2 were added only to goalkeepers. There were some more properties from table 1 that were not associated with a goalkeeper. The case can be made that each position can be a different class. This is possible because dbpedia does indeed have different classes for each position. I chose not to do this because:

- Many statistics and properties are shared between both types of players (*club, matches played, starts*).
- For my use case, predicting which players will be released or retained by a club, the player's position shouldn't affect the predictions.
- Based on how I have parsed the data, there can be no inconsistencies; outfield players cannot have goalkeeper statistics and vice versa.

- There are only 67 goalkeepers in this dataset, so defining a separate class just for them feels like unnecessary effort.

The exact properties associated with both types of players are illustrated with an example below.

```
<https://footballerontology.com/player/246350> a fb:player ;
    fb:aerialswon <https://footballerontology.com/aerialswonSeq/246350> ;
    fb:assiststotal <https://footballerontology.com/assiststotalSeq/246350> ;
    fb:attackingworkrate <https://footballerontology.com/attackingworkrateSeq/246350> ;
    fb:birthyear 2000 ;
    fb:clearancessuccessful <https://footballerontology.com/clearancessuccessfulSeq/246350> ;
    fb:club <https://footballerontology.com/clubSeq/246350> ;
    fb:contractuntil <https://footballerontology.com/contractuntilSeq/246350> ;
    fb:defending <https://footballerontology.com/defendingSeq/246350> ;
    fb:defensiveworkrate <https://footballerontology.com/defensiveworkrateSeq/246350> ;
    fb:dribbling <https://footballerontology.com/dribblingSeq/246350> ;
    fb:foulscommitted <https://footballerontology.com/foulscommittedSeq/246350> ;
    fb:foulssuffered <https://footballerontology.com/foulssufferedSeq/246350> ;
    fb:fts <https://footballerontology.com/ftsSeq/246350> ;
    fb:goalstotal <https://footballerontology.com/goalstotalSeq/246350> ;
    fb:height "170"^^xsd:float ;
    fb:interceptions <https://footballerontology.com/interceptionsSeq/246350> ;
    fb:intreputation <https://footballerontology.com/intrepSeq/246350> ;
    fb:matchesplayed <https://footballerontology.com/matchesplayedSeq/246350> ;
    fb:minutes <https://footballerontology.com/minutesSeq/246350> ;
    fb:name "Enzo Le Fée" ;
    fb:nationality "France" ;
    fb:overallrating <https://footballerontology.com/overallratingSeq/246350> ;
    fb:pace <https://footballerontology.com/paceSeq/246350> ;
    fb:passescompletelong <https://footballerontology.com/passescompletelongSeq/246350> ;
    fb:passescompletepercentage <https://footballerontology.com/passescompletepercentageSeq/246350> ;
    fb:passing <https://footballerontology.com/passingSeq/246350> ;
    fb:physicality <https://footballerontology.com/physicalitySeq/246350> ;
    fb:position "CAM" ;
    fb:potential <https://footballerontology.com/potentialSeq/246350> ;
    fb:preferredfoot "Right" ;
    fb:shooting <https://footballerontology.com/shootingSeq/246350> ;
    fb:shotsongoaltotal <https://footballerontology.com/shotsongoaltotalSeq/246350> ;
    fb:skillmoves <https://footballerontology.com/skillmovesSeq/246350> ;
    fb:starts <https://footballerontology.com/startsSeq/246350> ;
    fb:touches <https://footballerontology.com/touchesSeq/246350> ;
    fb:value <https://footballerontology.com/valueSeq/246350> .
```

(a) Outfield player

```
<https://footballerontology.com/player/248478> a fb:player ;
    fb:aerialswon <https://footballerontology.com/aerialswonSeq/248478> ;
    fb:birthyear 2001 ;
    fb:clearancessuccessful <https://footballerontology.com/clearancessuccessfulSeq/248478> ;
    fb:club <https://footballerontology.com/clubSeq/248478> ;
    fb:contractuntil <https://footballerontology.com/contractuntilSeq/248478> ;
    fb:foulscommitted <https://footballerontology.com/foulscommittedSeq/248478> ;
    fb:foulssuffered <https://footballerontology.com/foulssufferedSeq/248478> ;
    fb:fts <https://footballerontology.com/ftsSeq/248478> ;
    fb:gkdiving <https://footballerontology.com/gkdivingSeq/248478> ;
    fb:gkhandling <https://footballerontology.com/gkhandlingSeq/248478> ;
    fb:gkkicking <https://footballerontology.com/gkkickingSeq/248478> ;
    fb:gkpositioning <https://footballerontology.com/gkpositioningSeq/248478> ;
    fb:gkreflexes <https://footballerontology.com/gkreflexesSeq/248478> ;
    fb:height "186"^^xsd:float ;
    fb:intreputation <https://footballerontology.com/intrepSeq/248478> ;
    fb:matchesplayed <https://footballerontology.com/matchesplayedSeq/248478> ;
    fb:minutes <https://footballerontology.com/minutesSeq/248478> ;
    fb:name "Diant Ramaj" ;
    fb:nationality "Germany" ;
    fb:overallrating <https://footballerontology.com/overallratingSeq/248478> ;
    fb:passescompletelong <https://footballerontology.com/passescompletelongSeq/248478> ;
    fb:passescompletepercentage <https://footballerontology.com/passescompletepercentageSeq/248478> ;
    fb:position "GK" ;
    fb:potential <https://footballerontology.com/potentialSeq/248478> ;
    fb:preferredfoot "Right" ;
    fb:starts <https://footballerontology.com/startsSeq/248478> ;
    fb:touches <https://footballerontology.com/touchesSeq/248478> ;
    fb:value <https://footballerontology.com/valueSeq/248478> .
```

(b) Goalkeeper

Figure 1: Example players and their properties

```
<https://footballerontology.com/clearancessuccessfulSeq/246242> a rdf:Seq ;
    rdf:_1 "0.28"^^xsd:float ;
    rdf:_2 "0.74"^^xsd:float .

<https://footballerontology.com/clearancessuccessfulSeq/246350> a rdf:Seq ;
    rdf:_1 "0.78"^^xsd:float ;
    rdf:_2 "1.36"^^xsd:float .

<https://footballerontology.com/clearancessuccessfulSeq/246431> a rdf:Seq ;
    rdf:_1 "0.22"^^xsd:float ;
    rdf:_2 "0.0"^^xsd:float .
```

Figure 2: Example of property with rdf:Seq as range

I populated the KG by parsing Data 21.csv and Data 22.csv. I made a player type node for each player, with the sofifa ID used as the URI. Then, I added the static properties to the player node. I checked if the player was a goalkeeper or not, and then created sequence types with URI *propertynameSeq/playerID* for properties that change yearly, and added them to the KG. I then appended the player data row by row to the appropriate sequence. Finally, I serialized the graph into two turtle files (kg/schema.ttl and kg/players.ttl). The final KG has 143,673 triples with 1,145 players. The exact code used for this can be found in the code/build.ipynb file.

4.3 Mining

4.3.1 SPARQL queries

I wrote 5 queries to check for proper KG quality.

1. **Completeness:** Check for players with missing attribute links
2. **Consistency:** Find all literals whose datatype does not match the range of the predicate
3. **Semantic:** Find inconsistent birth years
4. **Completeness:** Find any attributes with missing values
5. **Conciseness:** Find any duplicate players

Only query 4 returned a result. There was one player (*András Schäfer*) who had a missing *fb:contractuntil* value for 2021. The other queries produced no results, showing that the KG is of good quality.
I also wrote 3 queries that would be hard to implement in the original structured data format. I think these queries highlight the effectiveness of a KG representation for football players. In particular, I searched for things that combined both FIFA and Stats data while also incorporating the temporal nature of the data.

6. Find the most improved players by passing rating and their passes completed percentage
7. Find players whose shooting rating in FIFA dropped despite improvement in shots on goals
8. Find the teams with players whose value dropped the most, and compare difference in overall rating

All three of these queries can be modified slightly to find results across various dimensions based on the user's information need. Some of them also return counterintuitive results.

4.3.2 Classification models

The first step for this task was vectorizing my data. Since I have values over 2 consecutive years, I had to compress this to a single value so that I could store it in a feature matrix. I chose to record the difference in player performance across all attributes. I reason that players whose performance drops significantly from the previous year are let go by their club, while at the same time, greatly improved players are picked up by bigger and better clubs. It is then not absolute performance that matters, but the growth (or lack thereof) that a player demonstrates. There is also a standardization in the data. Each player is compared to themselves and the magnitude and direction of growth is retained.
I started by creating a dictionary called players to store each player's ID along with their performance changes over two seasons. For each player, I calculated the difference in performance for each attribute (*aerialswon, assiststotal*, etc.) between the two seasons. This singular value represents the player's improvement or decline in that attribute. I also added a dropped attribute to indicate whether the player was released by their club (1 if dropped, 0 if not), determined by checking if the player's *fb:club* property had the same value over both years.
I created a sorted list of all possible attributes to structure the feature matrix, excluding the dropped attribute since it was the target variable. Using this list, I built a feature matrix $\mathbf{X}$. Here, each row corresponds to a player, and each column corresponds to an attribute. For players missing certain attributes, I filled in 0 for that attribute.
My target vector $\mathbf{y}$ was the dropped attribute. After dropping a player (*András Schäfer*) with a NaN value for the contract attribute, I split $\mathbf{X}$ and $\mathbf{y}$ into training and testing sets in the ratio 80:20, respectively.
I trained a logistic regression model on the training data. I used my trained model to predict probabilities for the test set and applied a threshold of 0.5 to convert these probabilities into binary predictions.
Next, I trained a simple neural network with one hidden layer containing 8 neurons and an output layer with 1 neuron. The hidden layer used the ReLU activation function, and the output layer used the Sigmoid activation function. I trained the neural network using Binary Cross-Entropy Loss (BCELoss) and the Adam optimizer. After training for 50 epochs, I used the model to predict probabilities for the test set and applied the same threshold (0.5) to convert these probabilities into binary predictions. The SPARQL queries and code for vectorization and classification can be found in the code/mine.ipynb file.

5 Results

5.1 Inference

```
Most improved players by passing rating and their growth in passes completed percentage:
        name        |              overallRatingGrowth              |                    passesCompletedGrowth
--------------------------------------------------------------------------------------------------------------------------------------------
"Anthony Gordon"    |"12"^^<http://www.w3.org/2001/XMLSchema#integer>|"14.599999999999994"^^<http://www.w3.org/2001/XMLSchema#float>
"Castello Lukeba"   |"15"^^<http://www.w3.org/2001/XMLSchema#integer>|"-1.9000000000000057"^^<http://www.w3.org/2001/XMLSchema#float>
"Gianluca Scamacca" |"12"^^<http://www.w3.org/2001/XMLSchema#integer>|"-7.8999999999999915"^^<http://www.w3.org/2001/XMLSchema#float>
"Hugo Ekitike"      |"13"^^<http://www.w3.org/2001/XMLSchema#integer>|"11.800000000000011"^^<http://www.w3.org/2001/XMLSchema#float>
"Luca Ranieri"      |"16"^^<http://www.w3.org/2001/XMLSchema#integer>|"8.700000000000003"^^<http://www.w3.org/2001/XMLSchema#float>
"Mattia Viti"       |"12"^^<http://www.w3.org/2001/XMLSchema#integer>|"0.5"^^<http://www.w3.org/2001/XMLSchema#float>
"Nadir Zortea"      |"13"^^<http://www.w3.org/2001/XMLSchema#integer>|"-4.299999999999997"^^<http://www.w3.org/2001/XMLSchema#float>
"Nico Schlotterbeck"|"18"^^<http://www.w3.org/2001/XMLSchema#integer>|"3.799999999999997"^^<http://www.w3.org/2001/XMLSchema#float>
"Patrick Osterhage" |"12"^^<http://www.w3.org/2001/XMLSchema#integer>|"-13.599999999999994"^^<http://www.w3.org/2001/XMLSchema#float>
"Warmed Omari"      |"12"^^<http://www.w3.org/2001/XMLSchema#integer>|"0.0"^^<http://www.w3.org/2001/XMLSchema#float>
```

Figure 3: Results of query 6

While most players with improved ratings have actually improved their performance, there are some whose rating has improved even though their real-life performance has gone down.

Players whose FIFA rating dropped despite improvement in real life (shooting attribute):

name	fifaRatingDiff	shotsOnGoalDiff
"Alexander Isak"	"-3"^^<http://www.w3.org/2001/XMLSchema#integer>	"0.5599999999999998"^^<http://www.w3.org/2001/XMLSchema#float>
"Antoine Griezmann"	"-2"^^<http://www.w3.org/2001/XMLSchema#integer>	"1.49"^^<http://www.w3.org/2001/XMLSchema#float>
"Arkadiusz Milik"	"-2"^^<http://www.w3.org/2001/XMLSchema#integer>	"0.24"^^<http://www.w3.org/2001/XMLSchema#float>
"Fabio Quagliarella"	"-3"^^<http://www.w3.org/2001/XMLSchema#integer>	"0.71"^^<http://www.w3.org/2001/XMLSchema#float>
"Lars Stindl"	"-5"^^<http://www.w3.org/2001/XMLSchema#integer>	"0.24"^^<http://www.w3.org/2001/XMLSchema#float>
"Lionel Messi"	"-3"^^<http://www.w3.org/2001/XMLSchema#integer>	"0.77"^^<http://www.w3.org/2001/XMLSchema#float>
"Marcel Sabitzer"	"-3"^^<http://www.w3.org/2001/XMLSchema#integer>	"1.0"^^<http://www.w3.org/2001/XMLSchema#float>
"Marcus Rashford"	"-2"^^<http://www.w3.org/2001/XMLSchema#integer>	"0.4300000000000005"^^<http://www.w3.org/2001/XMLSchema#float>
"Romelu Lukaku"	"-2"^^<http://www.w3.org/2001/XMLSchema#integer>	"0.1800000000000005"^^<http://www.w3.org/2001/XMLSchema#float>
"Youssef En-Nesyri"	"-3"^^<http://www.w3.org/2001/XMLSchema#integer>	"0.61"^^<http://www.w3.org/2001/XMLSchema#float>

Figure 4: Results of query 7

Here we see that players whose statistics have improved have seen their FIFA ratings reduced, which is unfair to their performances.

Players who lost most value and their overall rating difference:

name	initialTeam	valueDiff	ratingDiff
"Dayot Upamecano"	"FC Bayern München"	"27000000.0"^^<http://www.w3.org/2001/XMLSchema#float>	"1"^^<http://www.w3.org/2001/XMLSchema#integer>
"Jadon Sancho"	"Manchester United"	"55000000.0"^^<http://www.w3.org/2001/XMLSchema#float>	"3"^^<http://www.w3.org/2001/XMLSchema#integer>
"Marc-André ter Stegen"	"FC Barcelona"	"30500000.0"^^<http://www.w3.org/2001/XMLSchema#float>	"2"^^<http://www.w3.org/2001/XMLSchema#integer>
"Marcel Sabitzer"	"FC Bayern München"	"27000000.0"^^<http://www.w3.org/2001/XMLSchema#float>	"4"^^<http://www.w3.org/2001/XMLSchema#integer>
"Marcus Rashford"	"Manchester United"	"40500000.0"^^<http://www.w3.org/2001/XMLSchema#float>	"4"^^<http://www.w3.org/2001/XMLSchema#integer>
"N'Golo Kanté"	"Chelsea"	"28000000.0"^^<http://www.w3.org/2001/XMLSchema#float>	"1"^^<http://www.w3.org/2001/XMLSchema#integer>
"Raheem Sterling"	"Manchester City"	"35000000.0"^^<http://www.w3.org/2001/XMLSchema#float>	"2"^^<http://www.w3.org/2001/XMLSchema#integer>
"Raphaël Varane"	"Manchester United"	"29500000.0"^^<http://www.w3.org/2001/XMLSchema#float>	"2"^^<http://www.w3.org/2001/XMLSchema#integer>
"Robert Lewandowski"	"FC Bayern München"	"35500000.0"^^<http://www.w3.org/2001/XMLSchema#float>	"1"^^<http://www.w3.org/2001/XMLSchema#integer>
"Ángel Di María"	"Paris Saint-Germain"	"29500000.0"^^<http://www.w3.org/2001/XMLSchema#float>	"3"^^<http://www.w3.org/2001/XMLSchema#integer>

Figure 5: Results of query 8

We see that the player's ratings has gone up (although not a lot) in spite of their reduction in value. The players here are all superstars of the game, and even slight drops in performance affect them a lot. The difference metric is incorrect in their case, since they perform at a very high level and have very little room for growth (especially in FIFA ratings).
Queries 6 and 7 seem to have some contradicting results, and further investigation is required to confirm the validity of my results. I can think of 2 reasons for this to be the case: (i) These players have attempted more risky passes/shots, and while their accuracy has reduced, the output (reward) for the team is increased (ii) Data quality issues in either the Stats or FIFA data sources.

5.2 Classification

The logistic regression model has a test set classification accuracy of 67%, and the vanilla neural network has a test set classification accuracy of 69%. Since both models had similar performance, I recommend using the logistic regression model due to its interpretability and the ease with which the impact of individual features can be adjusted and their effects on the output can be observed.

6 Discussion

The main goal of this project was to set up a KG and a classification model that would benefit both players and the clubs they play for. I was able to set the KG up successfully. I was able to compare two prediction models for predicting whether or not players will be retained. My prediction system can also be used in the middle of a season (provided that the data is available) since it is based on the difference in player performance. It can be used as a reminder for players to improve their performance, compare themselves against the competition, and make clubs more informed for salary negotiations. However, I could not implement graph embeddings correctly on time. I had to sacrifice the quality and time spent on individual aspects (building/mining) to get the full set of results on time.
Some improvements and future work:

1. **Design:** Players can change clubs mid-season, example: Ronaldo played for Juventus and Manchester United in 2021. Since I am storing each player's club year-wise, I will not be able to store this information. I will only capture the first club that a player has played for in these instances. And drop the duplicate. Any changes in the club are only recorded after the full season. In my representation, Ronaldo played for Manchester United in 2021 and went to Juventus only in 2022.

2. **Data Quality:** The Stats data uses inconsistent naming practices. This makes it harder to merge on the FullName. The FIFA data has two separate name columns called short name (initial. last name) and full name (first name middle name last name). The Stats data has some combination of this. This makes the matching tricky. I tried to use fuzzy matching to do this, but that ended up degrading the FIFA data. Hence, I chose to forgo names that aren't an exact match due to time constraints and worked with about 60% of the total data available.

3. **Small n:** Deep Learning methods work well when there are a lot of observations. I had a total of 1144 players, of which only 30% (340) changed their clubs in these years. If I had more observations and more transfers, I am confident the neural network would have outperformed logistic regression.

4. **Graph embeddings:** Players are related across many dimensions, such as preferred foot, statistics, play styles, etc. I can infer that Messi, Salah and Saka are similar players since they all play on the right wing with their left foot and are top performers for club and country. Learning graph embeddings that can accurately capture these aspects can positively enhance the classification process.

5. **Domain Knowledge:** Working on this project made me realize the importance of domain knowledge for both building and mining the KG. The scope of my project is limited by my expertise in football, which was built largely by playing the sport growing up :) I can see how a domain expert's input can improve the quality of my implementation a lot.

I gained the ability to apply a wide range of topics covered in the course to my project. Building and mining a knowledge graph on a topic of interest helped me fulfill the following course objectives: (i) define and describe what a Knowledge Graph is (ii) identify and describe the components of a Knowledge Graph (iii) distinguish between different representations of Knowledge Graphs, and identify their strengths and weaknesses (iv)describe and execute approaches to construct Knowledge Graphs (v) construct and query Knowledge Graphs to answer questions about their content using open standards such as RDF and SPARQL (vi) describe KG quality metrics and evaluate the quality of a KG (vii) develop your own KG solution for a problem of interest.
Beyond the course objectives, undertaking an end-to-end project independently, especially under time constraints, taught me about project management, setting realistic expectations, and communicating my findings clearly and concisely.

7 Conclusions

I successfully constructed a KG integrating real-world football player statistics with FIFA video game data across two consecutive seasons. I validated it's quality using SPAQRL queries. Lastly, I built and compared two classification models for predicting football player releases.
The findings from the SPARQL queries revealed insights into player performance and valuation trends, highlighting discrepancies between real-world statistics and FIFA ratings. For instance, some players showed improved FIFA ratings despite declining real-world performance, suggesting potential biases or data quality issues. While other players saw a decrease in their ratings in spite of improved performance in the real world. These inferences show that a KG representation is an effective tool for football analytics.
The $\sim 70\%$ accuracy achieved in the classification task provided validation that the features I selected from the original data sources effectively and reasonably capture the relevant information about a football player.. The fact that logistic regression had similar results to a neural network indicates that it is a problem that can be solved with explainable tools guided by a domain expert.

8 Disclosure on the use of AI

Grammarly was used to correct grammatical mistakes in the report. ChatGPT was used in the initial stages of the project for idea generation. In later stages of the project, GPT-4o was used for code completion, debugging, and also writing some helper functions. I also used ChatGPT to receive feedback on my writing. I uploaded my report and the grading metric and asked it to score my writing, highlighting the strong and weak points. The output was used to improve my final submission.

9 Appendix

- The code for this project is available here.
- The Stats data used was collected from Vivo Vinco on Kaggle under the Attribution 4.0 International License.
- The FIFA data used was collected from Alex on Kaggle under the Attribution-NonCommercial-ShareAlike 4.0 International License.

References

[1] Bekkers, J. and Sahasrabudhe, A. [2024], 'A graph neural network deep-dive into successful counterattacks'.
URL: *https://arxiv.org/abs/2411.17450*

[2] Rakha, A. and Torralba, A. [2024], 'Graph neural networks in football: A new era of formation recommendations'.

[3] Rana, A. S. S. [2023], 'Event detection in football using graph convolutional networks'.
URL: *https://arxiv.org/abs/2301.10052*