NOVA IMS Information Management School

MDSAA

Master Degree Program in
Data Science and Advanced Analytics

Machine Learning: Applying Genetic Algorithms for Budget Management & Team Composition in Fantasy Premier League (FPL)

Ayotunde Anuoluwapo Aribo

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Machine Learning: Applying Genetic Algorithms for Budget Management & Team Composition in Fantasy Premier League (FPL)

by

Ayotunde Aribo

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Roberto André Pereira Henriques, PhD, NOVA Information Management School (NOVA IMS)

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Ayotunde Aribo
Lisbon, July 13 2024

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to those who have provided invaluable support and guidance throughout my master's thesis journey.

First, I would like to thank my supervisor, Dr. Roberto Henriques, for his continuous support, patience, and insightful feedback. His expertise and encouragement have been instrumental in guiding me through this research.

I am profoundly grateful to the faculty and staff of the Data Science Program at Nova IMS for creating a stimulating environment and providing the resources necessary for my research.

I would also like to extend my appreciation to my colleagues and fellow students for their discussions and collaborations during my program. Your support and encouragement have made this journey more enjoyable and enriching.

My heartfelt thanks go to my family, whose unwavering support and belief in my abilities have been a source of strength throughout this process. To my parents, Olawunmi and Adeola Aribo and also my brother, Abiodun Aribo, thank you for your constant encouragement and for instilling in me the value of education.

This thesis is dedicated to all of you who have played a part in this significant milestone in my academic journey. Thank you.

ABSTRACT

This thesis explores how genetic algorithms (GAs) can be used to optimize team selection in the Fantasy Premier League (FPL), an online multiplayer fantasy game. The main goal is to find the best possible teams that maximize overall performance, considering constraints like budget limits and player positions. The research looks at various configurations of GA operators, using different selection strategies, crossover methods, and mutation techniques. Specifically, it examines rank selection, tournament selection, and roulette wheel selection for selection strategies; single-point, double-point, and uniform crossover methods; and scramble mutation, swap mutation, and inversion mutation for mutation techniques. A total of 10 different GA configurations are tested to see which one works best for creating teams in FPL.

The results show that the best configuration involves tournament selection with uniform crossover and scramble mutation. This setup consistently achieved the highest fitness scores, indicating it performs best in team optimization. The genetic algorithm was tested using historical data from the 2023-2024 FPL season. The fitness function evaluated team configurations based on player performance (points per game), budget constraints, and other game rules like player positions.

The findings demonstrate that genetic algorithms are effective for the complex task of FPL team selection. This thesis adds to the field of Machine Learning in Fantasy Sports by comparing different GA setups and pointing out the most effective strategies for improving team performance. Future research could look into adding more constraints and performance metrics to further enhance the optimization process.

KEYWORDS

Genetic Algorithms (GAs); Fantasy Premier League (FPL); Machine Learning (ML); Optimization; Budget; Team Selection; Football

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. BACKGROUND

Fantasy Premier League (FPL) represents the pinnacle of sports management and fantasy gaming, captivating millions of enthusiasts worldwide. Originating from the English Premier League, one of the most prestigious football leagues globally, the FPL offers participants the unique opportunity to assume the role of team managers, selecting real-world football players to compose their fantasy teams (O'Brien et al., 2021).

The Fantasy Premier League (FPL) has experienced significant growth since its inception in the 2002-03 season. Initially starting with around 76,000 participants, the game quickly gained popularity, especially with the advent of digital technology which made it more accessible and engaging for a broader audience. By the 2012-13 season, FPL had already attracted millions of players, and as of the 2023-24 season, it boasts over 11 million participants globally. This growth is in line with the overall trend in fantasy sports, which has become a mass phenomenon (Conti, 2010). This exponential increase in participants and worldwide engagement can be attributed to various factors, including the introduction of new features such as chips (e.g., Bench Boost, Triple Captain, Free Hit) and improvements in the game's digital infrastructure, making it more user-friendly and reliable during peak times (All About FPL, 2021), (MATEJ ŠUĽAN, 2023).

The FPL, in particular, has emerged as a frontrunner in the fantasy gaming landscape, attracting a growing legion of participants with each passing season. Its global appeal transcends geographical boundaries, uniting football fans from various corners of the world under the common pursuit of virtual glory (Burton et al., 2013). Its user-friendly interface, comprehensive player database, and dynamic gameplay. The FPL has become synonymous with excitement, competition, and camaraderie among fantasy sports enthusiasts.

Although machine learning approaches have been utilized to enhance the strategies of fantasy football game managers, the application of genetic algorithms in budget management and player composition is not readily apparent. In this paper, we explore the increasing acknowledgement of Fantasy Premier League (FPL) among users, emphasize the significance of financially optimizing the squad to achieve the best level of performance and examine genetic algorithms as a potential method for optimization within the FPL context.

As the FPL continues to evolve and innovate, it highlights the appeal of fantasy gaming, offering participants a platform to showcase their managerial prowess, tactical acumen, and football knowledge. Its significance extends beyond mere entertainment, shaping how fans engage and experience the beautiful game globally.

1.2. Problem Statement

A significant gap in the research on the Fantasy Premier League (FPL) is the insufficient detail on how genetic algorithms (GAs) can be optimally used for creating the best lineups and managing team strategies in fantasy sports. Despite extensive research on various aspects of fantasy sports leagues, there is a noticeable paucity of studies specifically addressing the application of genetic algorithms in the FPL. This gap highlights the need for an investigation into the utility of GAs in this context. Addressing this area could lead to tremendous benefits or even difficulties.

Studies of this sort gave in-depth assessments of fantasy sports and recognized the increase and changes related to players` engagement with this activity (Burton et al., 2013). In spite of that, a clear-cut genetic algorithm method as a part of the FPL is still a controversial subject. This study will innovatively utilize evolutionary programming techniques, such as genetic algorithms, to address the current lack of research in the field. By exploring these techniques in the context of a fantasy game environment, it aims to contribute valuable new knowledge and optimal navigation methods, showing how computational intelligence techniques perform in evolved playing fields with varying numbers of players and budgetary allocations.

1.3. Research Objectives

The research aims to apply genetic algorithms to optimize budget management in the Fantasy Premier League (FPL) by developing a computational framework that iteratively refines team compositions and budget allocations to maximize performance within budget constraints.

To reach this objective, we have outlined the following intermediate objectives:

- Perform exploratory data analysis on the key characteristics and performance indicators of FPL players that influence team selection.
- Understand and apply techniques for preprocessing FPL data and visualizing it graphically to identify patterns and insights.
- Apply Genetic Algorithm techniques to optimize the budget and team composition in FPL, aiming to maximize overall team performance.
- Evaluate the performance of the Genetic Algorithm implemented, and draw conclusions on its effectiveness in optimizing FPL team selection and budget management.

1.4. Significance of the Study

The significance of this research lies in its potential to validate sports management theories and apply them to fantasy sports, specifically the Fantasy Premier League (FPL). Its contributions to this field are multifaceted and impactful.

1.4.1. Academic Contribution

This research explores the innovative use of genetic algorithms in fantasy sports, focusing on the Fantasy Premier League (FPL). By exploring how these algorithms can optimize decision-making in FPL, the study bridges computer science and sports management. It shows that genetic algorithms are not only effective but also versatile for solving complex problems in various fields.

- Aims to develop and apply genetic algorithms within the realm of fantasy sports, particularly the Fantasy Premier League (FPL). By doing so, it extends research beyond traditional decision-making mechanisms in dynamic and unpredictable environments.
- This research examines the suitability of genetic algorithms, a commonly used AI technology, in the sports management field, which fosters interdisciplinary dialogue and the partnership between computer science and sports management researchers.
- This study presents clear evidence of the availability of genetic algorithms as the most suitable methods of all other systems of optimization for tackling complicated issues in a real-world environment, which leads to their verification in all other disciplines.

1.4.2. Practical Applications

FPL managers can leverage genetic algorithms to optimize team structures and manage budget constraints. By using evolutionary approaches, managers can identify and retain the most effective selection criteria while discarding inefficient ones, leading to better team compositions. This study enhances decision-making in FPL by providing a comprehensive understanding of player selection, promoting data-driven choices, and strengthening competitive advantage. The global application of this research offers valuable insights, improving the competitiveness and enjoyment of FPL for participants worldwide.

2. LITERATURE REVIEW

The literature review chapter is a crucial guide that navigates us through a vast array of information, encompassing discussions on fantasy sports management with focus on the Fantasy Premier League (FPL). We will explore a variety of research from domains such as sports analytics, machine learning, and optimization algorithms. Our objective is not only to update ourselves with the current level of knowledge but also to gain awareness of the shared elements that connect different parts of the study. By meticulously examining the available literature, we aim to identify the most exceptional concepts and any gaps or overlooked aspects that need additional investigation. This expedition also enhances our comprehension of the thrills that can be derived from fantasy sports, ultimately providing us with valuable insights.

2.1. GENETIC ALGORITHMS IN OPTIMIZATION

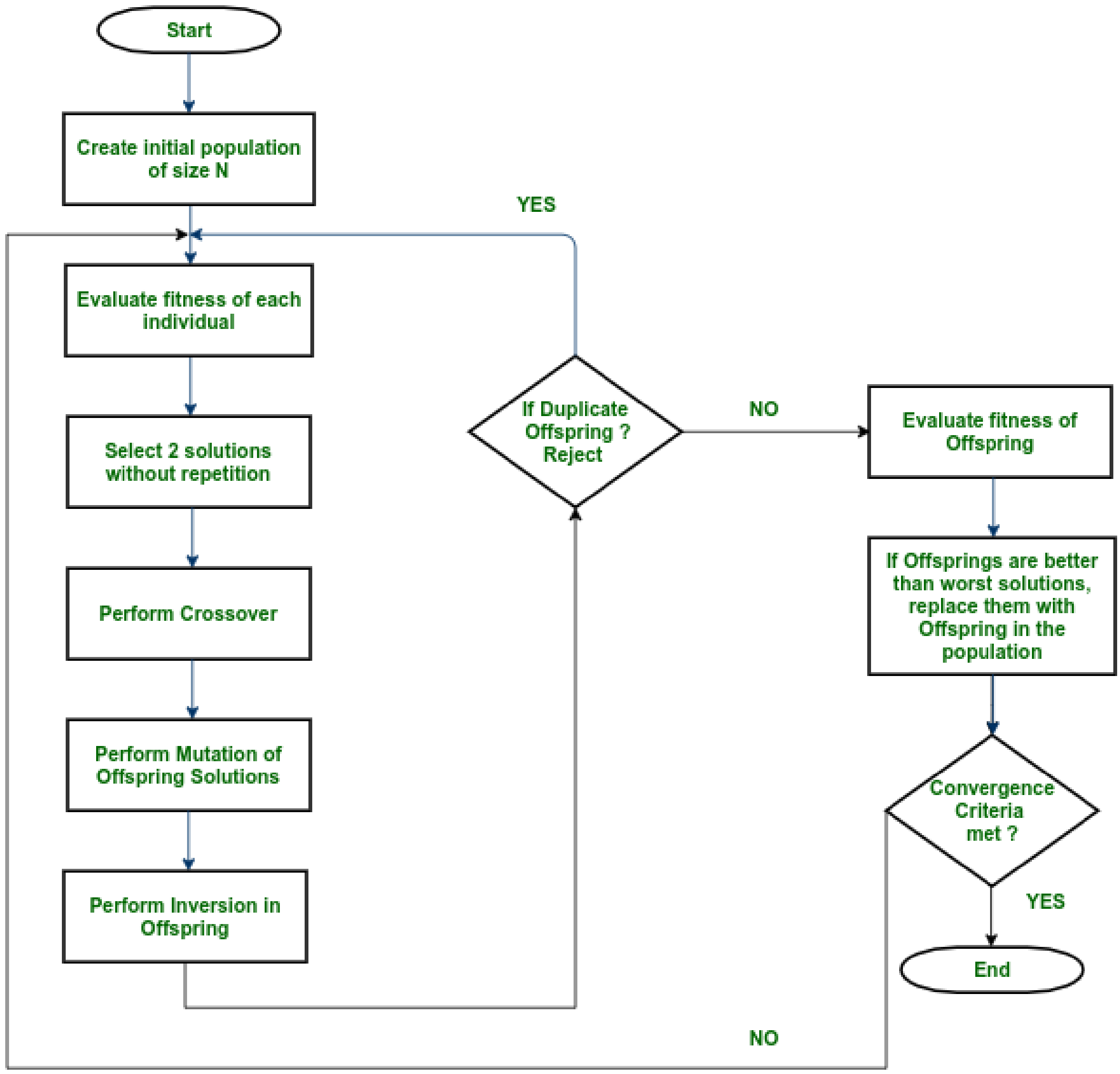
Genetic algorithms (GAs) are the foundational component of optimization algorithms, drawing inspiration from the principles of natural selection and genetics. They are inspired by the theory of natural evolution and are based on the idea of survival of the fittest (Batista et al., 2009). These algorithms work by encoding potential solutions to a problem on a chromosome-like data structure and applying recombination operators to preserve critical information (Mirjalili, 2018). The process begins with a population of random chromosomes, which are then evaluated and given reproductive opportunities based on their fitness. This allows for the evolution of a solution to the problem over time.

Famed for his pioneering work in the 1970s, John H. Holland significantly advanced the field of genetic algorithms (GA) with his seminal book, Adaptation in Natural and Artificial Systems (Holland, 1992). Holland's work introduced genetic algorithms as a robust problem-solving approach, utilizing a population of candidate solutions that evolve over time through processes analogous to natural selection and genetic recombination. His contributions have led to widespread adoption and continuous improvement of genetic algorithms, making them a prevalent method for tackling complex adaptive systems.

Genetic algorithms (GAs) offer several advantages for optimization problems. They are global, parallel, and robust, making them effective in dealing with non-convexity, locality, and complexity (Lam & Yin, 2001). GAs can be enhanced with other algorithms, such as Sequential Quadratic Programming, to improve search quality and prevent premature convergence (Al-Milli, 2014). They are particularly useful for optimizing complex, high-dimensional, and non-linear systems, where conventional methods may struggle to converge (Rezende et al., 2008). GAs are also inspired by natural evolution, providing a robust solution for a given problem (Janga Reddy & Kumar, 2012).

The core premise of genetic algorithms is the implementation of evolutionary concepts where a group of candidate solutions chromosomes, enduring selection, crossover, and mutations resemble the natural selection process—generation after generation, individuals with higher fitness (compared to the given performance objective). A fitness landscape tends to survive and is able to contribute offspring for the next generation, which in turn creates a convergence toward the best solution (Beasley et al., 1993)

Figure 1 – Genetic Algorithm Steps, image sourced from (geeksforgeeks, 2021)



Learning how the biology of genetic algorithms works requires comprehending their distinctions from their evolutionary equivalents by learning their initialization, selection, reproduction, crossover, mutation, and termination criteria. The operators combined constitute the main push-stick of evolutionary operation, entrusted with accurately exploring the solution space, maintaining it diversity-laden, and eventually focusing on assessment areas that terminate at promising solutions as observed by (Forrest, 1996).

Empirical studies with real-world usage have revealed not only how genetic algorithms display adaptability but also offer the capacity to boost throughput. The main contribution of (Yin,

2000) to the optimization of transportation issues using genetic algorithms was the development of an efficient method for solving bilevel programming models. This approach significantly improved the modeling of complex transportation network logistics by providing a simpler, computationally efficient solution that more effectively achieves global optimization compared to previous heuristic algorithms. Another instance in point, like (Al-Milli, 2010), hybrid genetic algorithms combined with Great Deluge algorithms were utilized to tackle limited optimization problems, and they were superior in solving infringing behavior within tight limitations. Moreover, (Pontes et al., 2016) optimized a large-scale industrial reactor using a genetic algorithm (GA) that integrated process knowledge into the initial population. This optimization led to a significant increase in process efficiency and economic performance, with profits rising by up to 25% compared to traditional industrial practices.

Genetic algorithms have been used in a number of disciplines, including sports analytics, techniques to arrange football teams, and others. (Deb, 1999) utilized genetic algorithms to introduce and solve complex optimization problems by providing a comprehensive overview of their principles and applications, specifically highlighting their effectiveness in handling constrained, multi-modal, and multi-objective optimization tasks. (Costa et al., 2019) released a framework for sorting and positioning digital people appropriately by means of genetic algorithms for the sake of increased performance in the competition. Moreover, (Imbrie et al., 2020) found research success in democratizing high-performing educational groups through the genetic optimization of teams for heterogeneity, seeking an environment where diversity and collaboration among all team members are balanced.

In a nutshell, genetic algorithms function as a strong and systematic instrument capable of accomplishing a large scope of optimization projects within a number of enterprise disciplines. They flawlessly perform search operations in difficult situations and identify relevant subsets of solutions. They can properly transform dynamic and changing surroundings and pick optimum or substandard solutions.

2.2. Budget Optimization in FPL And Its Importance to Team Selection

A range of studies have explored different approaches to optimizing Fantasy Premier League team performance and financial strategy. Budget optimization in Fantasy Premier League (FPL) is a critical factor in team performance (Pantuso, 2017). It ensures that the team has the required mix of skills, meets competition regulations, and respects budget limits. This is particularly important in a stochastic environment, where the uncertainty in player career development must be considered (Pantuso, 2017). The use of optimization models can significantly improve decision-making and lead to steady growth in team value. However, achieving optimal efficiency under budget constraints can be complex, requiring careful management and potentially unequal treatment of team members (López-Pintado & Moreno-Ternero, 2011).

2.3. MACHINE LEARNING IN FANTASY SPORTS

Machine learning (ML) technologies have drawn increasing attention in fantasy sports as they can be promising to pluck players carefully, design the most efficient squad, and complete extensive performance analysis. To begin with, this portion elaborates on the current literature that analyzes the employment of ML algorithms in varied functions of fantasy sports leagues, notably those of benefit to the public, like the Fantasy Premier League (FPL).

Regardless of the technique reflected in the literature, the essential conclusion is that artificial intelligence for fantasy sports analytics is wanted and currently applied by many. (Papageorgiou et al., 2024) introduced a novel approach that ensures accurate NBA player predictions and lineup optimizations in daily fantasy sports, demonstrating that machine learning is indeed an effective tool for predicting player outcomes. (Koseler & Stephan, 2017) give a thorough literature analysis regarding ML in baseball, highlighting the necessity of employing ML approaches to determine player quality and game performances.

(Gupta, 2019) developed a model using a hybrid of Autoregressive Integrated Moving Average (ARIMA) and Recurrent Neural Networks (RNNs) for time series prediction of player points in the Fantasy Premier League. The predictions were then optimized using Linear Programming (LPP) to maximize total points while considering constraints such as player roles and budget limits. The model's effectiveness was validated based on the predicted performance of players in the subsequent season, demonstrating its feasibility and potential for application in real-world scenarios.

(Kotrba, 2020) investigated the use of heuristic strategies, finding that users often choose squads based on players' past performances and the team's status as a favorite, although the latter may be overestimated while (Wang et al., 2022) proposed an improved GA for team formation, focusing on person-job matching and collaboration.

In their paper (Muniz & Flamand, 2023), a thorough examination of sports analytics due to balanced decision-making for team-building illustrates ML's involvement in optimizing the team composition for better performance. (Dhanday & Ranjan, 2024) came up with an analytical method to pick potential winner IPL Dream11 teams using a clever analysis of data, showcasing how data science is being applied realistically in fantasy cricket tournaments. Table 1 summarizes numerous ML models used in the fantasy sports area, containing techniques, algorithms, outcomes, and restrictions.

Table 1 - Summary of Studies on Machine Learning Applications in Fantasy Sports

Author Name & Year	**Methodology**	**Algorithm**	**Findings**	**Limitations**
(Papageorgiou et al., 2024)	Data preprocessing,	Machine learning	Developed an accurate method	Limited to NBA data; may not

	predictive modeling	ensemble models	for NBA player performance forecasting and lineup optimization	generalize to other sports leagues
(Koseler & Stephan, 2017)	Systematic literature review	Various ML techniques	Reviewed applications of machine learning in baseball, highlighting predictive modeling capabilities	Limited to a review; does not provide original findings
(Muniz & Flamand, 2023)	Analytical modeling, optimization techniques	CNNs	Propose sports analytics methods for balanced team-building decisions	Lack of specificity regarding the techniques and algorithms used
(Dhanday & Ranjan, 2024)	Analytical approach	Not specified	Developed an approach for crafting probable winning Dream11 teams using IPL data insights	Lack of detail on the specific methods and algorithms employed
(Pretorius & Parry, 2016)	Comparison of human and AI decision-making in sports	Machine learning models	Compared human and AI prediction accuracy in sports, indicating AI's potential for improvement	Lack of detailed analysis on specific AI techniques
(Bäck, 1996)	compares genetic algorithms, evolution strategies, and evolutionary programming	genetic algorithms, evolution strategies, and evolutionary programming	Mutation is crucial for genetic algorithms. A meta-evolution experiment combines ESs and GAs for mixed-	Some theoretical results need further empirical validation.

	within a unified framework		integer optimization	
(Hussain et al., 2024)	Context-aware game strategy	RunsGuard Framework	Proposed a framework for cricket game strategy for field placement and score containment	Limited to cricket; may not be applicable to other sports
(Rajesh et al., 2022)	Player recommendation system	Machine learning models	Developed a player recommendation system for the Fantasy Premier League	Lack of information on the specific machine learning algorithms used
(Spiros Valouxis, 2023)	Machine learning techniques	Not specified	Explored the application of machine learning in the Fantasy Premier League	Lack of detail on the specific machine learning techniques applied
(Bangdiwala et al., 2022)	ML models for point prediction	Machine learning models	Used machine learning models to predict points in the Fantasy Premier League	Lack of information on the specific machine learning algorithms utilized
(Ramdas, 2022)	Convolutional Neural Networks	CNNs	Utilized CNNs to predict footballers' performance in the Fantasy Premier League	Limited to CNNs; may not explore the full range of machine learning techniques applicable to FPL

In addition, in (Pretorius & Parry, 2016), the authors look at human decisions against AI approaches employed in sports prediction, which helps create a circumstance where machine learning algorithms are more suitable since they are correct. (Rajesh et al., 2022) offered a player recommendation system for the Fantasy Premier League, which is based on ML, approaches and gives solid information for the selection team and strategy selection.

Besides this, (Bangdiwala et al., 2022) and (Ramdas, 2022) are also important contributors. They employ machine learning models together with a convolutional neural network to forecast the performance and points of a fantasy Premier League player each week. (Valouxis, 2023) speaks about employing ML in efficiently determining the players' lineups, while (Hussain et al., 2024) put out their innovative RunsGuard Framework for creating cricket game strategies. Additionally, the study done by the two researchers (Kolb & Kolb, 2010) and (Bhatt et al., 2019) illustrates the educational qualities of fantasy sports and how crowds can make smarter selections. (Vidgen et al., 2023) utilized a machine learning model to analyze over 2.3 million tweets which highlighted the abuse on Twitter directed at football players during the 2021-22 Premier League season.

The adoption of ML technologies in fantasy sports has played an unusual role and shown numerous techniques and approaches that are targeted at enhancing decision-making, squad makeup, and player performance prediction. These investigations accurately show how ML approaches contribute to everything about current fantasy sports analytics.

2.4. EXISTING STUDIES IN FANTASY PREMIER LEAGUE

Numerous studies on the Fantasy Premier League (FPL) have explored various techniques, algorithms, and insights. These papers have been carefully reviewed to understand their methodologies, the algorithms they employ, and the insights they uncover.

(O’Brien et al., 2021) conducted a thorough analysis of skill identification in the Fantasy Premier League (FPL). They examined several factors contributing to player success, including decision-making, player selection, and the balance between skill and luck. Their study highlights the intricate and diverse elements that lead to sustained success in the FPL, offering significant insights into what drives player performance.

In a related study, (Rajesh et al., 2022) presented a sophisticated player suggestion system that leverages machine learning to address some of the most challenging issues in FPL today. This system automates current player selection process, providing football fans with an enhanced user experience. (Valouxis, 2023) not only emphasizes the application of machine learning techniques to optimize team performance in Fantasy Premier League (FPL) but also introduces innovative methodologies for predicting Expected Value (EV) and generating strategic recommendations for team selection and transfers. Also, (Bangdiwala et al., 2022) and (Ramdas, 2022) undertake detailed studies of machine learning models for forecasting the points of an FPL player, bringing valuable approaches and contributions to problem-solution management.

The papers examined here encompass diverse strategies, ranging from traditional statistical approaches to modern deep learning techniques applied to neural networks and evolutionary algorithms. These methodologies surrounding FPL offer a wide idea of outcomes-based decision-making processes contributing to correct player suggestions, predictive models, and unique approaches to sport-style decision-making in FPL administration.

2.5. SUMMARY AND RESEARCH GAP IDENTIFICATION

In reviewing the extensive literature on Fantasy Premier League (FPL) management and optimization, it becomes evident that while various machine learning techniques have been explored, there remains a notable research gap in applying genetic algorithms (GA) specifically for optimizing FPL team management strategies. Existing studies by (Burton et al., 2013), (O'Brien et al., 2021), (Rajesh et al., 2022), and others have predominantly focused on skill identification, player recommendation systems, and predictive modeling using machine learning algorithms. While these approaches have shown promise in enhancing FPL decision-making, the utilization of genetic algorithms, which excel in optimizing complex problem domains, has been largely overlooked.

With it's ability to emulate natural selection processes and iteratively evolve solutions, genetic algorithms offer a unique approach to addressing the complexities inherent in FPL team selection. Therefore, a compelling research gap exists that our study aims to address by investigating the potential of genetic algorithms for FPL team management. By filling this gap, our research endeavors to develop innovative methodologies and decision-support systems that can significantly improve the efficiency of FPL team selection processes.

The identified research gap underscores the need for further exploration into the application of genetic algorithms in the context of FPL. By bridging this gap, our study aims to contribute novel insights and methodologies to advance the field of FPL optimization; ultimately empowering fantasy sports enthusiasts with more sophisticated and data-driven strategies for team management.

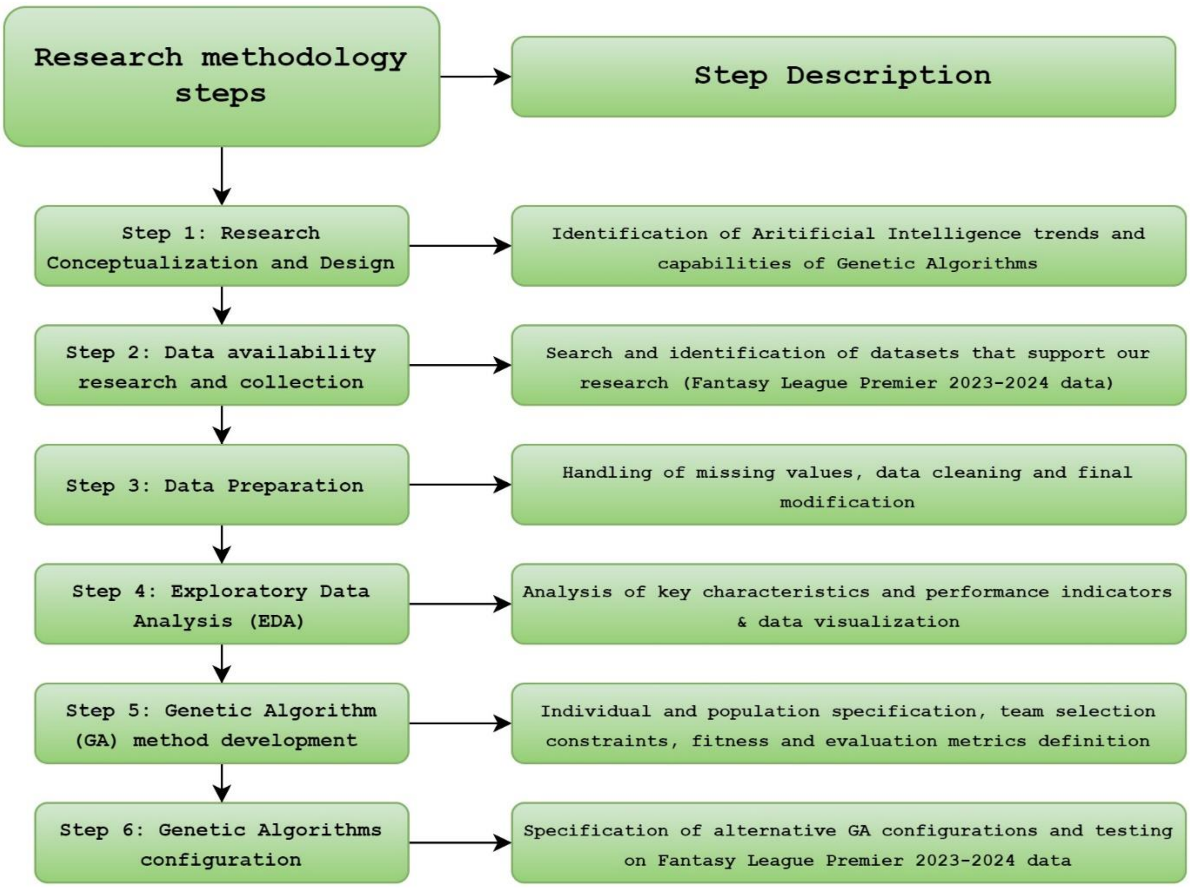
3. METHODOLOGY

This chapter outlines the methodology for developing and applying a genetic algorithm (GA) for optimizing budget management and team composition in Fantasy Premier League (FPL). The approach involves several key stages, including data collection and preprocessing, designing the genetic algorithm framework, implementation, and evaluation of the algorithm's performance.

3.1. Research Design

The research methodology for this study is designed to apply genetic algorithms (GA) to Fantasy Premier League (FPL) team management through six structured steps. The process begins with identifying trends in Artificial Intelligence and exploring the capabilities of GAs. Next, relevant datasets from the FPL 2023-2024 season are collected. The data is then meticulously prepared by cleaning and handling missing values. This is followed by Exploratory Data Analysis (EDA) to understand key characteristics and performance indicators. In the development phase, the GA method is crafted by specifying individuals and populations, defining constraints, and establishing evaluation metrics. Finally, various GA configurations are tested on the FPL data to develop innovative methodologies and decision-support systems, aimed at enhancing the effectiveness and efficiency of FPL team selection processes.

Figure 2 – Overview of Research Methodology



3.2. Retrieving and Managing Data

3.2.1. Data Collection

The data employed in this study was sourced from a Google Sheet that receives updated Fantasy Premier League (FPL) data and analytics every 12 hours, covering the recently concluded 2023/2024 season, which ended in May 2024. This data is made publicly accessible by (Wafir Manakad, 2023) on a platform called medium, allowing anyone to make a copy for their analysis, thereby eliminating the requirement of an expensive and time-consuming data collection process. This data was selected for its real-time updates, comprehensive coverage of FPL player statistics, and the availability of crucial indicators for analysis.

Dataset Name: FPL Data 2023/2024

Dataset Explanation: Medium page explaining dataset - Manakad, W

Dataset Availability: Access Fantasy Premier League Data (23/24) on Google Sheets

It was possible to access the dataset by downloading an offline xlsx copy and saving locally to machine. The dataset contains the following tabs:

- *Intro*
- *Data*
- *Transfer Picks*
- *Top 100 Managers*
- *Top 100 Template*
- *Transfers-IN*
- *Transfers-OUT*
- *Team Form*
- *Difficulty*
- *Price Increase*
- *Price Decrease*

The 'Data' tab was chosen among all available tabs in the dataset because it directly aligns with the purpose and theme of this study. It includes 861 rows and 36 rows of relevant player information, and performance metrics.

3.2.2. Data Features

Table 2 shows the columns in the "Raw FPL Data" tab and their meanings. This detailed breakdown provides a comprehensive understanding of the dataset and the various data points considered for optimizing team composition in the Fantasy Premier League (FPL):

Table 2 - Column Details of the FPL Dataset

Columns	Description
Player Name	The name of the player.
Team	The team the player is currently playing for.
Position	The position of the player (e.g., Goalkeeper, Defender, Midfielder, Forward).
Points	The total points accumulated by the player in the current season.
Goals	The total number of goals scored by the player in the current season.
Assists	The total number of assists made by the player in the current season.
Cost	The current cost of the player in the FPL.
Form	The player's current form, based on recent performances.
Total Transfers	The total number of transfers in and out for the player.
Minutes Played	The total minutes played by the player in the current season.
Clean Sheets	The number of clean sheets kept by the player (relevant for goalkeepers and defenders).
Yellow Cards	The number of yellow cards received by the player.
Red Cards	The number of red cards received by the player.
Influence	A measure of the player's influence on matches.

Creativity	A measure of the player's creativity in creating goal-scoring opportunities.
Threat	A measure of the player's goal threat.
ICT Index	A composite index that combines Influence, Creativity, and Threat.
Influence	A measure of the player's influence on matches.

3.2.3. Data Preparation

Before beginning with the analysis, it is important to prepare the data to fit with the defined objectives. This section explains the methods adopted to preprocess and clean the data retrieved from the Fantasy Premier League (FPL) dataset. The dataset is imported into a Python environment using the Pandas module, and a first review is performed to obtain insight into the data and identify any relevant concerns.

- **Handling Missing Values**: Missing values were discovered and treated by replacing with the mean and also the min where appropriate.
- **Data modification**: Certain columns were modified or calculated to increase their usefulness for the research. This also includes the generation of new features based on existing variables.
- **Feature Selection**: Traditionally, players are selected based on past and predicted performance tied to several features. Here, a correlation matrix was used during Exploratory Data Analysis (EDA) to analyse and identify relevant features such as player points per game, cost, influence, among other performance metrics.
- **Final Dataset Preparation**: After completing the preparation processes, the dataset was cleaned and transformed for use by encoding categorical variables such as player positions and teams into numerical values.

3.3. DATA ANALYSIS

A crucial step in the research methodology of this thesis has been the Exploratory Data Analysis (EDA) on the Fantasy League Premier 2023-2024 dataset. In this step of our approach our efforts were focused on understanding the dataset through the EDA, and to extract insights and information about the problem that is attempted to be solved by the thesis.

The Exploratory Data Analysis (EDA) was conducted in Google Colab (Python notebook tool) and follows a comprehensive and systematic approach to uncover insights into the features (i.e. fields, columns) of the dataset. While several features are included within this dataset, as they are referenced in Section 3.2.2 and Table 2, not all of them are required for the goal of this thesis. This EDA was a necessary step (of the research methodology) to fully understand the importance of the dataset features (i.e. fields, columns) and relationship to the objective of this study. For that reason, by utilizing several Python libraries, namely `pandas`, `numpy`, `matplotlib`, `seaborn`, and `plotly.express`, data manipulation, statistical analysis, and visualization are achieved within the Google Colab environment in order to derive the necessary descriptive analytics.

In more detail, the EDA starts with team-level exploration. The number of players per team (e.g. Nottingham Forest, Chelsea, Luton, etc.) is analyzed and visualized using bar charts. This visualization is achieved by grouping the data by 'Team' and counting the number of players, which is then displayed using the plotly library to produce an interactive bar chart. This allows for an intuitive understanding of team representation within the league and the quantities of players available.

After that initial phase, a significant portion of the analysis is dedicated to examining key features such as Points per Game, Cost Today, Influence, and Chance of Playing Next. For instance, the average Points per Game per team is calculated and visualized to highlight performance disparities among teams. This is achieved by grouping the dataset by 'Team', calculating the mean Points per Game, and creating a sorted bar chart for the top 20 teams. Such visualizations have been crucial for identifying high-performing teams and understanding their contributions to the league's dynamics.

At the player level, the analysis focuses on identifying players or teams in exceptional form. The mean Form per team is calculated, and the results are visualized using a bar chart to depict teams with outstanding performance. Additionally, the top players based on their form are identified and visualized, providing insights into individual excellence within the league. In Figures 3 & 4, below one can notice the results extracted for the Form attribute/feature of the datasets.

Figure 3 - Exceptional Form per Team in the Fantasy Premier League dataset for 2023-2024 season

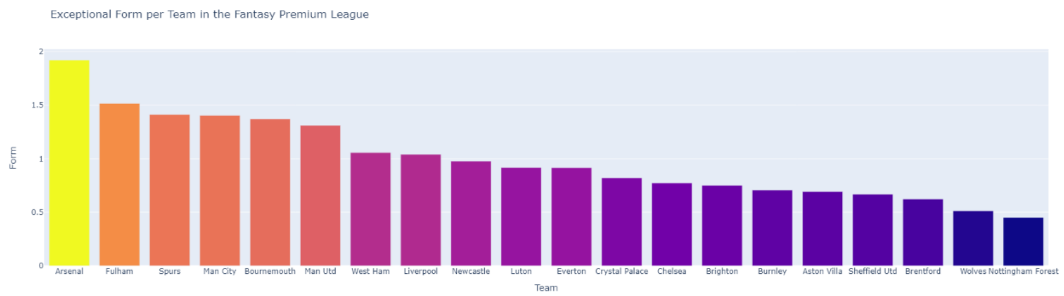
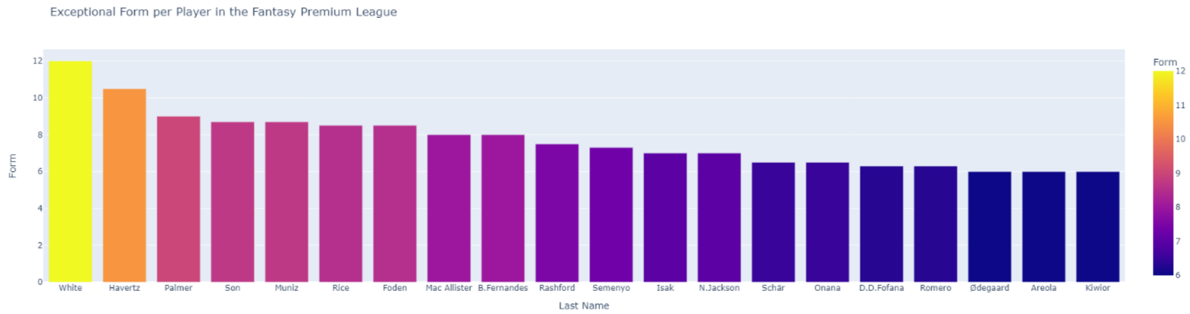


Figure 4 - Exceptional Form per Player Team in the Fantasy Premier League dataset for 2023-2024 season



The Google Colab notebook also explores trends in transfer patterns based on player or team performance, although it acknowledges limitations due to data availability. Fixture difficulty is analyzed to identify top players with favorable upcoming fixtures, which involves sorting the dataset by Difficulty Score and filtering players based on Points per Game and Minutes played. Further insights are gained by calculating the average playing minutes (excluding zero values) to provide additional context about player performance. The rest of the Figures plotted within this EDA can be found in the Appendix of this thesis or in the respective Python notebook files.

In summary, the EDA employs a structured approach to data exploration, utilizing descriptive statistics, univariate and bivariate analysis, and interactive visualizations. The analysis emphasizes critical features, identifies key trends, and uncovers patterns that are essential for informed decision-making. While these patterns are not very important to the main results of this thesis (as presented in Section 4), they can help us create the overall background for player characteristics and team composition which will be fundamental building block for developing the GA configurations later in this thesis.

3.4. Team Selection Constraints

We also explain the process of constructing an individual or team, highlighting the constraints necessary for developing effective and competitive teams.

- One key constraint is the limitation of not more than three players per real-life football team.
- The total cost of selected players must not exceed the predefined budget limit of 100 million
- the squad comprises of 15 individuals and must include a specified number of players for each position: two goalkeepers, five defenders, five midfielders, and three forwards.
- Each player is only selected once in the final team to ensure uniqueness

3.4.1. Further conditions for team selection

These constraints are not induced from the FPL settings but added to improve the optimization of the team selection

- Selecting players with a high likelihood of playing
- Selecting players who have played at least the average number of minutes in the dataset, which is 846 minutes. This requirement is essential to avoid selecting players with very few minutes played or those who have left the league for other clubs, thus ensuring the reliability and continuity of player performance

3.5. FITNESS FUNCTION

The fitness function plays a crucial role in genetic algorithms, as it significantly impacts the optimization results and the accurate representation of models, such as in the case of computing the electrical network settling behavior (Solopov & Samulchenkov, 2019). The fitness function in this research aims to maximize overall team performance while adhering to budget constraints and player selection rules, ensuring that the optimization process meets the specific requirements of the FPL. The ***calculate_fitness*** function does this by combining individual player contributions with penalties for exceeding the budget and violating team composition rules.

3.5.1. Player Performance Priority

The function starts with a lambda function, priority_lambda, which calculates a weighted sum of key performance indicators for each player. These indicators include:

- Points per game (40%)
- Current form (40%)
- ICT index (30%), which reflects the player's influence, creativity, and threat in matches.

This prioritization ensures that players who perform well individually are favored in the team selection process.

3.5.2. Budget Penalty

The function then calculates the total cost of the team and compares it to the available budget. If the team's cost exceeds the budget, a penalty is applied proportional to the excess amount. This penalty is subtracted from the fitness score to ensure budget constraints are respected.

3.5.3. Team Composition Constraints

Next, the function checks the distribution of players across different teams and positions. It ensures:

- No more than three players from the same team are selected.
- The team includes exactly two goalkeepers, five defenders, five midfielders, and three forwards.

Penalties are imposed for any violations of these constraints, calculated based on the excess number of players from any single team and deviations from the required number of players in each position.

3.5.4. Final Fitness Score

The final fitness score is adjusted by subtracting the penalties for budget and team composition violations from the initial performance-based fitness score. The Fitness function is defined as:

$$Fitness = \sum_{i-1}^{N}(0.4 \cdot PPG_i + 0.4 \cdot Form_i + 0.3 \cdot ICT_i) - max\left(0, \sum_{i-1}^{N} Cost_i - Budget\right) - \left(\sum_{team} max(0, Count_{team} - 3) + |GK_{Count} - 2| + |DEF_{Count} - 5| + |MID_{Count} - 5| + |FWD_{Count} - 3|\right)$$

Where $PPGi$ = Points Per Game for player i

$Formi$ = Current form for player i

$ICTi$ = ICT Index for player i

3.6. EVALUATION METRICS

In assessing the performance of the genetic algorithm (GA) for optimizing Fantasy Premier League (FPL) team selection, we focus on three key metrics: average fitness over generations, best fitness in each generation, and the number of valid teams produced. Most evaluations focus on fitness, which takes into account important aspects of a player's abilities, like their performance, form, and influence. This means that an algorithm with a higher fitness value for the best solution will include players who are performing better overall.

Average Fitness Over Generations

The average fitness of the population in each generation reflects the overall quality of solutions. It is calculated by averaging the fitness scores of all individuals in the population for each generation. This metric helps track the algorithm’s convergence and improvement over time.

Best Fitness in the Generation

The best fitness score in each generation indicates the highest fitness achieved by any individual in the population. This metric highlights the peak performance of the GA and its ability to find optimal team configurations.

Average points per game of best team

This metric represents the mean number of points scored per game by the team configuration that achieves the highest overall fitness score during the optimization process

Number of Valid Teams Produced

The number of valid teams produced in each generation measures how many solutions meet all constraints (e.g., budget, positions, player uniqueness). This metric ensures that the algorithm generates feasible and compliant solutions. We are not considering team performance because player performance is measured based on their individual contributions like goals, assists, clean sheets, and bonus points, while also deducting points for negative actions like cards and own goals. Unlike real football, a player's FPL points aren't tied to their team's performance; a player can still score high points even if their team loses, and vice versa.

3.7. PARAMETERS AND CONFIGURATIONS OF GENETIC ALGORITHMS

Here, the research concentrates on the parameters that are employed in a GA and methods of applying them to satisfy the given issue of FPL team selection. The GA will be implemented in Python, utilizing libraries such as NumPy for numerical computations and DEAP (Distributed Evolutionary Algorithms in Python) for genetic algorithm operations. The implementation involves the following parameters:

Table 3 - Parameters and Configurations of the Genetic Algorithm Used for Fantasy Premier League (FPL) Team Selection

Parameter	**Description**
Population Size	Determines the number of candidate solutions (individuals) in each generation.
Budget constraint	Allows the algorithm to consider financial constraints and optimize without.
Mutation Rate	Governs the probability of introducing random changes to individual solutions.
Crossover	Combines parts of two parent solutions to produce new offspring by mixing genetic material from high-performing individuals.
Elitism (Elite Size)	Number of top-performing individuals that are carried over unchanged to the next generation.
Generations	The number of iterations the GA will run to evolve the population.

4. RESULTS AND DISCUSSION

After going over our methodology in Section 3, we proceed to present the application of this methodology on a real-world historical dataset from the FPL season 2023-2024. In that regard, Section 4 is organized into three sub-sections, with the first focusing on the application scope and input datasets, the second focusing on the main comparative analysis, and the third focusing on the discussion of the results, as well as the connection to the overall topic of this thesis.

4.1. Application Scope and Input Datasets

The goal of the application within Section 4 of this thesis is to conduct a comparative approach between different configurations of GA approaches with different genetic operators. To achieve that, we compare ten different GA approaches on the same input datasets, each including various combinations of GA operators. They all attempt to solve the same problem: finding an optimal team synthesis for the FPL.

For this comparative analysis, we use the FPL historical dataset of player performance for the season 2023-2024, whose features are described in Section 3.2.2 and Table 2. For the construction of the ten different GA approaches, we combine the following GA operators:

Table 4 - Selection operators that are used in the GA configurations of the comparative analysis

Genetic operator	Description
Tournament selection	Tournament selection is a genetic algorithm technique where the top-performing individual from a randomly selected population subset is chosen for reproduction in the next generations.
Rank selection	Rank selection is a genetic algorithm technique in which individuals are ordered by their fitness levels, and the probability of selection is determined by their rank rather than their raw fitness scores, ensuring more uniform selection pressure.
Roulette wheel selection	Roulette wheel selection is a genetic algorithm technique where individuals are chosen for reproduction based on their fitness proportion (stochastic), giving those with higher fitness a better chance of being selected.

Table 5 - Crossover operators that are used in the GA configurations of the comparative analysis

Genetic operator	Description
Single point crossover	Single-point crossover is a genetic algorithm technique where two parent individuals swap parts of their chromosomes at a randomly chosen point, resulting in offspring with genetic material from both parents.
Double point crossover	Double-point crossover is when two parents swap parts of their chromosomes between two randomly chosen points, producing offspring that contain genetic material from both parents in those segments.
Uniform crossover	Uniform crossover is a genetic algorithm technique where genes from two parents are swapped randomly at each position (with equal probability), creating offspring with a mix of genes from both parents.

Table 6 - Mutation operators that are used in GA configurations of the comparative analysis

Genetic operator	Description
Swap mutation	Swap mutation is a genetic algorithm technique in which two genes within an individual's chromosome are randomly chosen and exchanged. This introduces variation while maintaining the overall structure of the chromosome.
Inversion mutation	Inversion mutation is when an individual's chromosome segment is selected and reversed, introducing variation while keeping the chromosome's length unchanged.
Scramble mutation	Scramble mutation is applied when we want a segment of an individual's chromosome to be randomly mixed up, adding variation to the solution while keeping the segment's length and genes the same.

In Tables 4 to 6, we present the GA operators that have been selected based on our survey of the respective literature on GA approaches. These operators have been implemented in Python 3 and Jupyter Notebooks, by utilizing several Python standard libraries (math, random, copy, etc.) and well-established external libraries such as pandas, numpy, plotly, and matplotlib for plotting. Based on these, in Table 7, we include the 10 GA configurations (i.e. approaches) tested in the comparative analysis of Section 4 of this thesis.

Table 7 - Configurations of genetic operators that comprise the 10 GA approaches tested in this comparative analysis

Genetic algorithm approach ID (#)	**Selection operator used**	**Crossover operator used**	**Mutation operator used**
GA #0	Tournament selection	Single point crossover	Swap mutation
GA #1	Rank selection	Uniform crossover	Inversion mutation
GA #2	Roulette wheel select.	Double-point crossover.	Scramble mutation
GA #3	Tournament selection	Double-point crossover.	Scramble mutation
GA #4	Tournament selection	Double-point crossover.	Swap mutation
GA #5	Tournament selection	Uniform crossover	Swap mutation
GA #6	Tournament selection	Uniform crossover	Scramble mutation
GA #7	Roulette wheel select.	Uniform crossover	Scramble mutation
GA #8	Roulette wheel select.	Single point crossover	Scramble mutation
GA #9	Rank selection	Single point crossover	Scramble mutation

While all the approaches in Table 7 consider different modelling of the genetic evolution process, they can utilize the same modelling concept for the solution of the GA, where an individual of the population is a team of players, with each player contributing to the attributes of the team (points, financial cost, form, overall fitness). Each team is also subject to constraints during the evolution (optimization) process, such as budgetary constraints and team composition by player type. Based on this shared modelling of the FPL teams, we can run the experiments based on a core implementation of Python classes and methods. Some of the core Python classes implemented and supported this comparative analysis are described in Table 8.

Table 8 - Python implementation details with classes and methods used (shared across all genetic operators)

Python classes	
players_data	List of player objects with attributes like minutes, position, cost, etc.
budget	The maximum budget allowed for the team.
population_size	The number of individuals (teams) in the population that is subject to evolution.
elite_size	The number of top individuals preserved across generations.
mutation_rate	The probability of an individual undergoing mutation (of any type)
tournament_size	The number of individuals competing in tournament selection (if tournament selection is used)
sigma_share	A parameter used (as a threshold) in fitness sharing to maintain diversity.
penalty_budget	This Budget penalty is applied based on the difference between the team cost and the available budget.
team_composition_penalty	This penalty is calculated based on the violation of the player positions and the desired team composition.
Python methods	
create_individual	This method creates a team with two goalkeepers, five defenders, five midfielders, and three forwards.
calculate_fitness	This method evaluates an individual's fitness based on player performance metrics (points per game, form, ICT index) and

	penalises teams that exceed the budget or violate team composition constraints.
evolve	This method runs the GA for a specified number of generations.
is_valid_team	This method ensures that a team does not exceed the budget and complies with position and team constraints.
sharing_function and *similarity* methods	The sharing function, defined in *'sharing_function'*, reduces the fitness of similar individuals. If the similarity (determined by the '*similarity'* method) between individuals is less than a specified threshold (*'sigma_share'*), the sharing function returns a value less than 1, reducing the effective fitness. If the similarity is above this threshold, the sharing function returns 0, indicating no sharing effect.
calculate_shared_fitness	This method calculates the shared fitness of an individual by dividing its fitness by the sum of the sharing function values for all individuals in the population. This adjustment ensures that individuals in densely populated areas of the solution space have lower shared fitness compared to isolated individuals.

A few implementations details will be shared at this point about the implementation of some of these Python methods. In Figure 5, we present the create_individual method which creates the teams (i.e. candidate solutions to the problem), by making sure that there are no duplicates, and a correct team synthesis is achieved (2 goalkeepers, 5 defenders, 5 midfielders, 3 forwards).

Figure 5 - The create_individual methods that reassures the teams (candidate solutions) are created with the proper synthesis of player positions

```python
def create_individual(self):
    unique_players = random.sample(self.players_data, len(self.players_data))
    goalkeepers = [player for player in unique_players if player.position == 'GK']
    defenders = [player for player in unique_players if player.position == 'DEF']
    midfielders = [player for player in unique_players if player.position == 'MID']
    forwards = [player for player in unique_players if player.position == 'FWD']

    while len(goalkeepers) < 2 or len(defenders) < 5 or len(midfielders) < 5 or len(forwards) < 3:
        unique_players = random.sample(self.players_data, len(self.players_data))
        goalkeepers = [player for player in unique_players if player.position == 'GK']
        defenders = [player for player in unique_players if player.position == 'DEF']
        midfielders = [player for player in unique_players if player.position == 'MID']
        forwards = [player for player in unique_players if player.position == 'FWD']

    team = goalkeepers[:2] + defenders[:5] + midfielders[:5] + forwards[:3]
    team = self.remove_duplicates(team)
    return team
```

Figure 6 - The calculate_fitness Python method that calculates the score for each solution at each GA generation/iteration

```python
def calculate_fitness(self, individual):
    priority_lambda = lambda player: (0.4 * player.points_per_game) + \
                                     (0.4 * player.form) + \
                                     (0.3 * player.ICT_index)
    fitness = sum(priority_lambda(player) for player in individual)

    team_cost = sum(player.cost_today for player in individual)
    penalty_budget = max(0, team_cost - self.budget)

    teams_count = {player.team: 0 for player in individual}
    teams_count_pos = {'GK': 0, 'DEF': 0, 'MID': 0, 'FWD': 0}

    for player in individual:
        teams_count[player.team] += 1
        teams_count_pos[player.position] += 1

    penalty_team_selection = sum(max(0, count - 3) for count in teams_count.values())
    penalty_team_selection += abs(teams_count_pos['GK'] - 2)
    penalty_team_selection += abs(teams_count_pos['DEF'] - 5)
    penalty_team_selection += abs(teams_count_pos['MID'] - 5)
    penalty_team_selection += abs(teams_count_pos['FWD'] - 3)

    fitness -= penalty_budget + penalty_team_selection

    return fitness
```

Respectively, in Figure 6, one can notice the calculate_fitness Python method that assigns a score to our GA individuals (teams) at each round of the algorithm's execution. The method assigns a penalty budget as well as team selection penalty, as explained above.

Finally, it is worth mentioning that the execution of a GA algorithm itself is defined by selecting several parameters that dictate the exploration versus exploitation dynamic. In genetic algorithms, exploitation leverages the best solutions to find optimal results, while exploration searches diverse solutions to avoid local optimal solutions. Balancing both is crucial to prevent premature convergence and ensure an efficient optimization process. To achieve this balance in our approaches, we have selected the following parameters for our comparative analysis: the initial population size was 100 individuals (teams). The elite size has been selected to be 10. The mutation probability has been chosen to be 5%, and the overall generations (i.e. iterations) of the GA have been chosen to be 100. With these concepts now being described, we proceed to Section 4.2, where the results of the comparative analysis are presented

4.2. COMPARATIVE ANALYSIS OF GENETIC ALGORITHM CONFIGURATIONS

To assess the performance of each GA approach with their respective genetic operator configurations, we have defined a set of optimization criteria for evaluating the suitability of each GA approach. Please note here that these criteria differ from the fitness criteria for the individuals (i.e. teams). Still, they are connected since the GA configuration criteria emerge from the finesses of individuals (teams) calculated at each evolution stage (iteration of the algorithm). The GA configurations assessment criteria are calculated based on the resulting final population of the GA (end of iterations) and are chosen as follows:

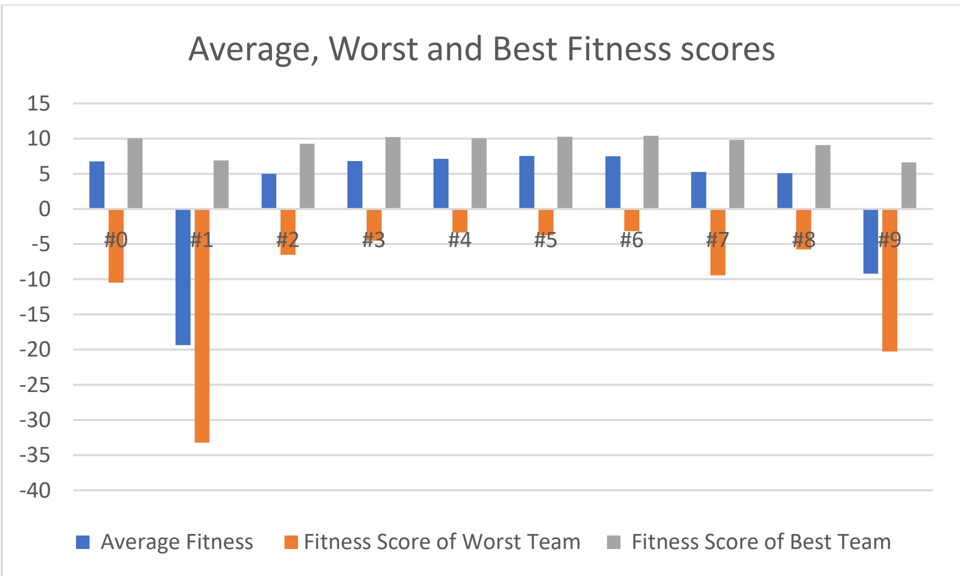
- Average Fitness: The mean fitness score of all individuals (teams) in the population.
- Best Fitness: The highest fitness score achieved by any individual in the population.
- Worst Fitness: The lowest fitness score among the individuals in the population.
- Average Points/Game of Best Team: The average points per game of the players in the best-performing team.
- Cost of Best Team: The total cost of the best-performing team.
- Average Form of Best Team: The average form metric of the players in the best-performing team.
- Valid Teams Produced by GA: The number of valid teams (meeting all constraints) produced by the genetic algorithm.
- Fitness Score of Best Team: The fitness score of the best-performing team.
- Average ICT Index of Best Team: The average ICT Index of the players in the best-performing team.

While these criteria are considered key, the most important one is the fitness score of the best team because we believe it is the end goal based on the GA's modelling approach, as explained in Section 3. Given these considerations, in Table 9 and Figure 7, we present the results of the comparative analysis between the GA configurations.

Table 9 - Results of the comparative analysis of the 10 GA configurations

GA #ID	Average Fitness	Fitness Score of Worst Team	Fitness Score of Best Team	Average Points per Game of Best Team	Cost of Best Team	Average Form of Best Team	Valid Teams Produced by GA	Average ICT Index of Best Team
#0	6.75	-10.4725	10.061	0.6703	92.1	0.643	99	0.4846
#1	-19.37	-33.217	6.9023	0.5416	90.7	0.331	100	0.3696
#2	4.967	-6.5034	9.2756	0.634	100.0	0.5296	100	0.509
#3	6.806	-4.535	10.228	0.709	96.0	0.616	100	0.5047
#4	7.137	-3.3401	10.039	0.699	99.2	0.58	100	0.5255
#5	7.556	-3.7587	10.281	0.693	98.8	0.611	100	0.544
#6	7.49	-3.1676	10.404	0.704	98.8	0.629	96	0.5329
#7	5.24	-9.4517	9.7898	0.685	99.8	0.541	100	0.5395
#8	5.06	-5.7727	9.0973	0.660	96.7	0.474	100	0.5086
#9	-9.23	-20.28	6.6048	0.498	80.6	0.358	100	0.32596

Figure 7 - Average, Worst and Best fitness across all teams and the GA configurations (#0 to #9)



Based on Table 9 and Figure 7, we can see considerable differentiation between the performance of our GA configurations, highlighting our study's necessity and importance. Several key observations can be made: Firstly, Configuration 6 (#6) achieved the best fitness score of approximately 10.4. This configuration also shows high average fitness (7.49) and a good balance in other metrics, indicating its robustness. Configuration 5 also shows a high average fitness score of 7.55, along with Configuration 6. This suggests that these configurations consistently produce high-quality solutions across the population. Configurations 1 and 9 had the lowest worst fitness scores of -33.217 and -20.28, respectively. Both of these GA approaches used Rank selection, and their low scores indicate that rank selection combined with these specific crossover and mutation methods can lead to poor-performing individuals in the population. The average points per game of the best team is highest for Configuration 3 at 0.709. This configuration also maintains a reasonable team cost of 96. The cost of the best team is within the budget constraints for all configurations, with the maximum cost observed being 100 for several configurations. Configuration 0 has a high average form of the best team at 0.643, indicating the efficient selection of in-form players. Finally, regarding the Valid Teams Produced metric, most configurations produced 100 valid teams, except for Configuration 6, which produced 96 valid teams. This slight reduction might be due to the strict constraints or specific characteristics of the scramble mutation method in this configuration.

An additional comment can be added regarding specific genetic operators and their effect horizontally across multiple configurations. Tournament Selection consistently performs well across different crossover and mutation methods. It combines effectively with uniform crossover and scramble mutation, leading to high average and best fitness scores. Rank Selection tends to underperform compared to other selection strategies, especially when combined with single-point crossover and scramble mutation. This indicates that rank selection might be less effective in maintaining genetic diversity and guiding the population towards optimal solutions. Roulette Wheel Selection shows moderate performance but outperforms tournament selection in most metrics.

As it emerges from this comparative analysis, Configuration 6 (#6) has provided the best result, effectively balancing exploration and exploitation through its selection, crossover, and mutation mechanisms. The fitness function and validity checks seemed strong enough to ensure the evolved teams were high-quality and compliant with constraints. This thorough approach to method experimentation and parameter tuning helps achieve robust and optimal solutions for team selection under budget constraints. From our analysis within the Jupyter Python notebooks, more details can be provided for Configuration 6 and the algorithm's behavior throughout the evolution process.

Figure 8 - Converge behavior of the GA Configuration 6

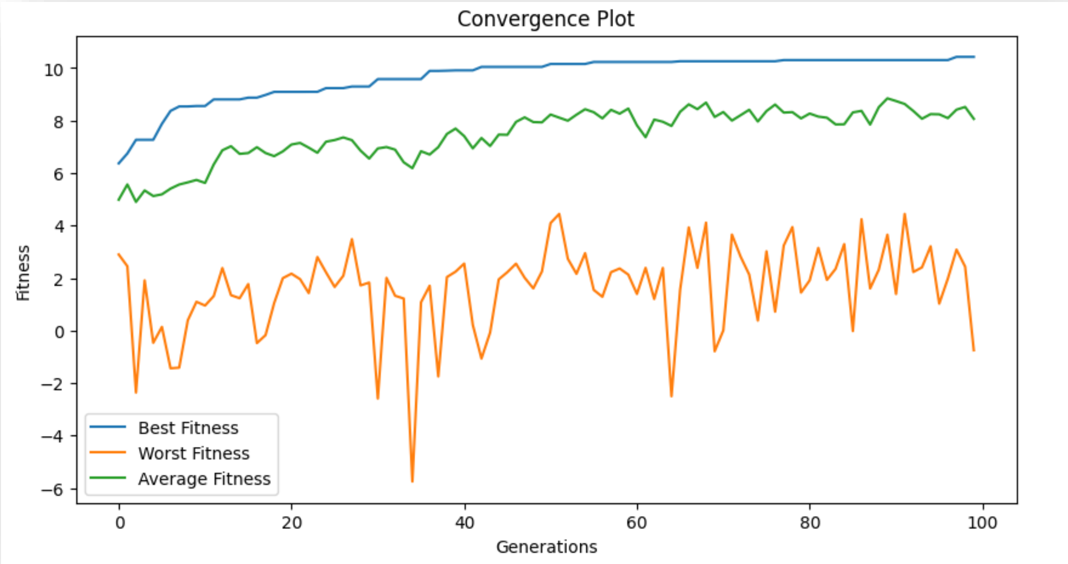
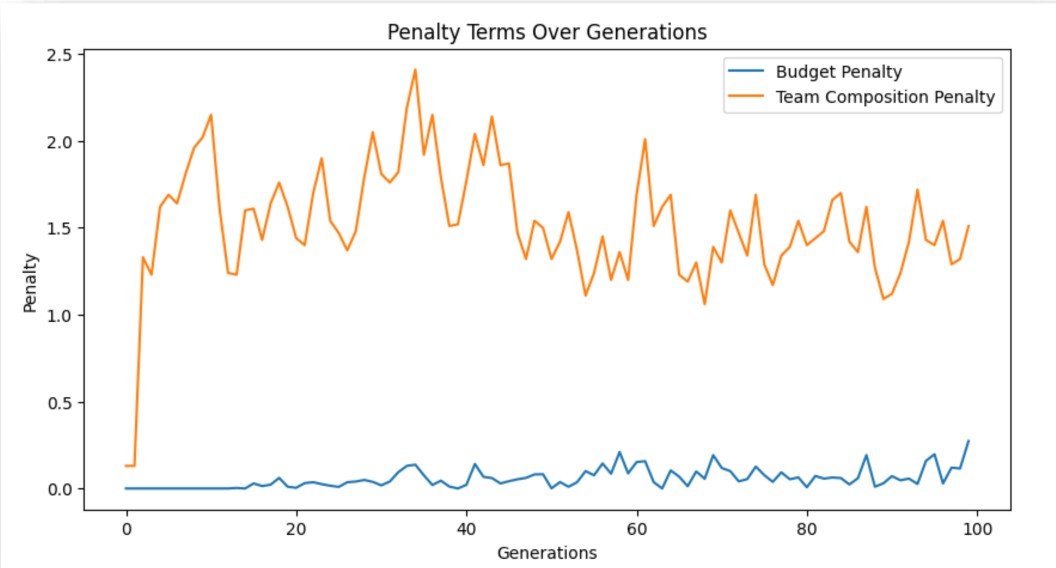


Figure 9 - Penalty evolution over the generation of the GA algorithm for GA configuration #6



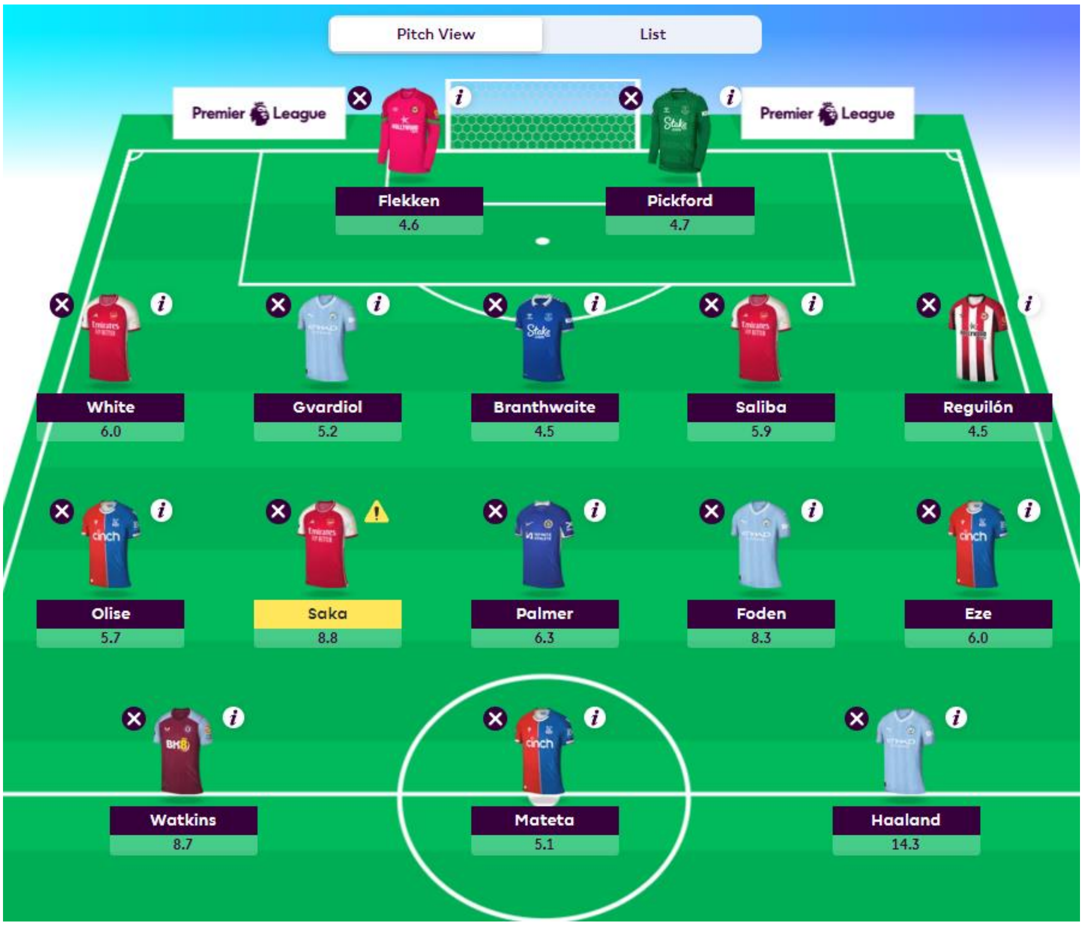
In Figure 8, the convergence behavior of GA configuration 6 is given, with three metrics being displayed: i) the best fitness (blue line), ii) the average fitness (green line), and iii) the worst fitness (orange line). Some observations are that by approximately generation 50, the best fitness value stabilizes and converges near 10, indicating that the algorithm has found and is maintaining a high-quality solution early. After generation 20, the average fitness stabilizes around a value slightly above 5. This suggests that, on average, the individuals in the population maintain a decent fitness level, but there is some variation among them. Finally, the worst fitness value is quite volatile, indicating a significant variation in the fitness of individuals within the population. This volatility decreases slightly over time but remains present, showing that some individuals still perform poorly even as the population improves. In Figure 9, we also provide the evolution of penalties for our algorithm across its iterations. Regarding the budget penalty, it is consistently low, suggesting that the algorithm is very effective at keeping team costs within the budget limit. The team composition penalty varies greatly, indicating that the algorithm struggles to ensure the correct number of players in each position. Nonetheless, there are periods when the team composition penalty increases, possibly due to the stochastic nature of genetic algorithms and the specific crossover and mutation operations used. With these two Figures presented in Section 4, we hopefully provide an overall picture, with the graphs highlighting the challenges faced by the genetic algorithm in deriving optimal solutions while at the same time maintaining budget and team composition constraints. As last reference to our results, the team or optimal solution consisting of 15 players, as calculated by the best GA Configuration (#GA 6) is given below in Table 10

Table 10 - The players selected as the optimal team in the FPL by the best performing GA (Configuration #6)

First name	Last name	Team	Position	Cost	Points per game	Form	ICT Index
Mark	Flekken	Brentford	GK	4.70	0.46	0.40	0.25
Jordan	Pickford	Everton	GK	4.80	0.57	0.73	0.26
Sergio	Reguilón	Brentford	DEF	4.50	0.42	0.43	0.24
Joško	Gvardiol	Man City	DEF	5.10	0.62	1.00	0.33
Jarrad	Branthwaite	Everton	DEF	4.50	0.50	0.69	0.27
William	Saliba	Arsenal	DEF	5.90	0.61	0.56	0.26
Benjamin	White	Arsenal	DEF	6.10	0.69	0.73	0.41
Bukayo	Saka	Arsenal	MID	8.90	0.90	0.62	1.00

Michael	Olise	Crystal Palace	MID	5.70	0.92	0.80	0.43
Phil	Foden	Man City	MID	8.50	0.88	0.60	0.87
Cole	Palmer	Chelsea	MID	6.30	1.00	0.47	0.85
Eberechi	Eze	Crystal Palace	MID	6.00	0.62	0.47	0.56
Jean-Philippe	Mateta	Crystal Palace	FWD	5.00	0.60	0.85	0.46
Ollie	Watkins	Aston Villa	FWD	8.90	0.88	0.43	0.80
Erling	Haaland	Man City	FWD	14.20	1.00	0.76	0.80

Figure 10 – FPL graphic representation of optimal team selected by #GA6



4.3. DISCUSSION OF RESULTS

Several genetic algorithm configurations were evaluated to determine the most effective approach for optimizing team selection in the Fantasy Premier League (FPL). The primary goal was to maximize the overall performance of the selected teams while adhering to constraints such as budget limits and player positional requirements. This study tested various configurations of genetic operators, including selection strategies, crossover methods, and mutation techniques.

The key findings are that Tournament Selection emerged as the most effective selection strategy. This method consistently selected higher-quality individuals for reproduction, which led to more efficient convergence towards optimal solutions. Tournament selection's balance between exploration (diversity) and exploitation (focus on high-performing individuals) was crucial in achieving high fitness scores. Uniform Crossover was identified as the most effective crossover method. Uniform crossover facilitated greater solution space exploration by allowing genes to be exchanged with equal probability at each position. This increased the likelihood of discovering highly fit individuals, leading to better-performing teams. Finally, Scramble Mutation was the most effective mutation technique. It introduced variability into the population by randomly shuffling a chromosome segment, preventing premature convergence and maintaining genetic diversity. This ensured that new genetic material was continually introduced, enhancing the algorithm's ability to explore a wider range of potential solutions.

Configuration 6 (tournament selection, uniform crossover, scramble mutation) achieved the highest performance across several metrics: Best Fitness Score: 10.404363, Average Fitness: 7.495548, Average Points/Game of Best Team: 0.704630, Fitness Score of Best Team: 10.404363 and Average ICT Index of Best Team: 0.532978. This configuration consistently produced high-quality solutions, demonstrating its robustness in optimizing FPL team selection, as can be noticed in Figures 8 and 9.

The problem of finding the best team in FPL involves selecting players that collectively maximize the team's performance while staying within budgetary and positional constraints. The metrics used to evaluate performance include points per game, form, ICT index, and overall fitness scores. By using a genetic algorithm with tournament selection, uniform crossover, and scramble mutation, digital managers (i.e. players of the game) can optimize their teams to maximize points while adhering to game constraints. Based on our comparative analysis, we can claim that the GA optimization approach, and more specifically Configuration 6, offers a competitive advantage, providing insights into the most effective player combinations and enabling managers to make data-driven decisions. This advantage is interesting and of considerable importance for various reasons, with the main one being that such a GA approach can be used as part of Artificial Intelligence (AI) system that plays the game effectively in real-time with the user.

5. CONCLUSIONS

The application of genetic algorithms (GA) to optimize Fantasy Premier League (FPL) team selection demonstrated significant potential in forming competitive teams within budget constraints. The analysis revealed that certain combinations of selection, crossover, and mutation techniques, particularly those involving tournament selection, consistently produced high-performing teams. Among the various configurations tested, the combination of tournament selection, uniform crossover, and scramble mutation emerged as the most effective, achieving the highest fitness scores and robust performance metrics. This approach not only optimized player selection but also efficiently managed the budget and team composition constraints, showcasing the versatility and power of genetic algorithms in navigating complex decision-making environments.

Throughout the study, the GA's ability to iteratively improve solutions was evident in the convergence plots. The steady rise in best fitness scores and the stabilization of average fitness values over generations highlighted the algorithm's capacity to identify increasingly optimal team configurations. Additionally, the penalty terms for budget and team composition constraints demonstrated the GA's effectiveness in balancing multiple objectives, ensuring that the final team selections adhered to the predefined limits. These results underscore the GA's practical applicability in real-world scenarios where balancing cost and performance is crucial.

5.1. LIMITATIONS

However, the study also revealed some limitations. The variability in performance across different configurations indicated that the GA's success is highly dependent on the chosen parameters. For instance, rank selection methods generally yielded lower fitness scores compared to tournament selection, highlighting the importance of careful parameter tuning.

Additionally, while the GA produced valid and competitive teams, the reliance on historical player data means that unforeseen factors such as injuries or transfers could impact the actual performance of the selected teams. This inherent unpredictability in sports adds a layer of complexity to the optimization process, which the GA must continuously adapt to.

Moreover, the computational cost associated with running multiple generations of the GA can be significant, particularly when optimizing large populations. This resource-intensive nature of GAs might pose challenges for real-time or frequent team updates. Future research could explore hybrid approaches, combining GAs with other optimization techniques to enhance efficiency and adaptability.

5.2. FUTURE WORK

Building upon the promising results obtained from the application of Genetic Algorithms (GAs) in optimizing team selection for the Fantasy Premier League (FPL), several avenues for future research and enhancement are suggested. These potential directions aim to refine the GA framework, expand its applicability, and explore new methodologies to further improve performance and decision-making in fantasy sports management.

Firstly, future work could focus on incorporating additional performance metrics and constraints into the GA model. While this study considered key metrics such as points per game, cost, influence, and threat, other factors like player injury history, team fixtures, and opponent strength could provide a more comprehensive evaluation of player performance and potential. By integrating these additional variables, the GA could offer even more nuanced and strategic team selections.

Secondly, exploring advanced genetic operators and hybrid approaches could enhance the GA's optimization capabilities. For instance, combining GAs with other machine learning techniques, such as neural networks or reinforcement learning, could create a hybrid model that leverages the strengths of both approaches. This integration could improve the algorithm's ability to adapt to the dynamic nature of fantasy sports and provide more accurate predictions and robust team compositions.

Additionally, the scalability of the GA framework could be investigated by applying it to other fantasy sports leagues and competitions. This would test the generalizability and flexibility of the proposed methodology, allowing for the identification of any sport-specific adjustments needed to optimize performance. Extending the research to different sports would also provide a broader validation of the GA's effectiveness in various competitive environments.

Another promising area for future work involves the real-time application of GAs in fantasy sports. Developing a dynamic GA model that updates team selections based on live data and real-time player performance could offer participants a significant competitive advantage. This real-time optimization would require the integration of live data feeds and efficient computational techniques to ensure timely and accurate updates.

User interaction and customization options within the GA framework could also be explored to enhance the user experience. Allowing users to input their preferences and strategies, such as risk tolerance or favorite players, could tailor the GA's recommendations to individual needs and preferences. This personalization could increase user engagement and satisfaction, making the GA tool more appealing to a broader audience of fantasy sports enthusiasts.

In conclusion, the future work outlined above presents exciting opportunities to enhance the application of GAs in fantasy sports management. By incorporating additional performance metrics, exploring hybrid approaches, extending the framework to other sports, enabling real-time optimization, offering user customization, and addressing ethical considerations, future research can build on the current study's findings to further improve strategic decision-making and optimize team performance in the dynamic world of fantasy sports.

BIBLIOGRAPHICAL REFERENCES

All About FPL. (2021, August). *FPL Now Vs Then | How Has FPL Evolved Since The Early Years*. All About FPL. https://allaboutfpl.com/2021/08/fpl-now-vs-then-how-has-fpl-evolved-since-the-early-years/

Al-Milli, N. (2010). Hybrid Genetic Algorithms with Great Deluge For Course Timetabling. *International Journal of Computer Science and Network Security*, *10*. https://www.researchgate.net/publication/266863507_Hybrid_Genetic_Algorithms_with_Great_Deluge_For_Course_Timetabling

Al-Milli, N. (2014). Hybrid Genetic Algorithm with great deluge to solve constrained optimization problems. *Journal of Theoretical and Applied Information Technology*, *59*, 385–389.

Bäck, T. (1996). *Evolutionary Algorithms in Theory and Practice*. Oxford University Press. https://doi.org/10.1093/oso/9780195099713.001.0001

Bangdiwala, M., Choudhari, R., Hegde, A., & Salunke, A. (2022). *Using ML Models to Predict Points in Fantasy Premier League*. https://doi.org/10.1109/ASIANCON55314.2022.9909447

Batista, M., Moreno-Pérez, J., & Moreno-Vega, J. (2009). Algoritmos Genéticos. Una visión práctica. *Números, ISSN 0212-3096, Nº. 71, 2009 (Ejemplar Dedicado a: Darwin)*.

Beasley, D., Bull, D. R., & Martin, R. R. (1993). *An overview of genetic algorithms: Part 1*. https://api.semanticscholar.org/CorpusID:15123759

Bhatt, S., Chen, K., Shalin, V. L., Sheth, A. P., & Minnery, B. (2019). Who Should Be the Captain This Week?Leveraging Inferred Diversity-Enhanced Crowd Wisdom for a Fantasy Premier League Captain Prediction. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 103–113. https://doi.org/10.1609/icwsm.v13i01.3213

Burton, R., Hall, K., & Paul, R. (2013). The Historical Development and Marketing of Fantasy Sports Leagues. *The Journal of SPORT*, *2*(2), 185–215. https://doi.org/10.21038/sprt.2013.0225

Conti, M. (2010). *The Fantasy Focus: A content analysis of sports magazines examining the rise in popularity of fantasy games* [West Virginia University Libraries]. https://doi.org/10.33915/etd.796

Costa, L. M., Souza, A. C. C., & Souza, F. C. M. (2019). An Approach for Team Composition in League of Legends using Genetic Algorithm. *2019 18th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)*, 52–61. https://doi.org/10.1109/SBGames.2019.00018

Deb, K. (1999). An introduction to genetic algorithms. *Sadhana*, *24*(4–5), 293–315. https://doi.org/10.1007/BF02823145

Dhanday, S., & Ranjan, S. (2024). *Fantasy Sports Meets Reality: An Analytical Approach to Craft Probable Winning Dream11 Team Using IPL Data Insights*.

Forrest, S. (1996). Genetic algorithms. *ACM Computing Surveys*, *28*(1), 77–80. https://doi.org/10.1145/234313.234350

geeksforgeeks. (2021, April 21). *Steady State Genetic Algorithm (SSGA)*. Geeksforgeeks. https://www.geeksforgeeks.org/steady-state-genetic-algorithm-ssga/

Gupta, A. (2019). *Time Series Modeling for Dream Team in Fantasy Premier League*. https://arxiv.org/abs/1909.12938

Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems*. The MIT Press. https://doi.org/10.7551/mitpress/1090.001.0001

Hussain, A., Arshad, S., & Hassan, A. (2024). RunsGuard Framework: Context Aware Cricket Game Strategy for Field Placement and Score Containment. *Applied Sciences*, *14*(6), 2500. https://doi.org/10.3390/app14062500

Imbrie, P. K., Agarwal, J., & Raju, G. (2020). Genetic Algorithm Optimization of Teams for Heterogeneity. *2020 IEEE Frontiers in Education Conference (FIE)*, 1–5. https://doi.org/10.1109/FIE44824.2020.9274243

Janga Reddy, M., & Kumar, D. N. (2012). Computational algorithms inspired by biological processes and evolution. *Current Science*, *103*, 370–380.

Kolb, A. Y., & Kolb, D. A. (2010). Learning to play, playing to learn. *Journal of Organizational Change Management*, *23*(1), 26–50. https://doi.org/10.1108/09534811011017199

Koseler, K., & Stephan, M. (2017). Machine Learning Applications in Baseball: A Systematic Literature Review. *Applied Artificial Intelligence*, *31*(9–10), 745–763. https://doi.org/10.1080/08839514.2018.1442991

Kotrba, V. (2020). Heuristics in fantasy sports: is it profitable to strategize based on favourite of the match? *Mind & Society*, *19*(1), 195–206. https://doi.org/10.1007/s11299-020-00231-7

Lam, H. K. W., & Yin, Y. F. (2001). Genetic algorithm-based approach for transportation optimization problems. In *The practical handbook of genetic algorithms : applications* (pp. 235–273). Chapman & Hall/CRC.

López-Pintado, D., & Moreno-Ternero, J. D. (2011). *On the optimal management of teams under budget constraints* (Working Papers, Issue 11.11). Universidad Pablo de Olavide, Department of Economics. https://doi.org/DOI:

MATEJ ŠUĽAN. (2023, June 15). *When did FPL start? History of Fantasy Football*. Fantasy Football Reports. https://www.fantasyfootballreports.com/history-of-fantasy-football/

Mirjalili, S. (2018). Genetic Algorithm. *Handbook of Machine Learning*. https://api.semanticscholar.org/CorpusID:17088312

Muniz, M., & Flamand, T. (2023). Sports analytics for balanced team-building decisions. *Journal of the Operational Research Society*, *74*(8), 1892–1909. https://doi.org/10.1080/01605682.2022.2118634

O'Brien, J. D., Gleeson, J. P., & O'Sullivan, D. J. P. (2021). Identification of skill in an online game: The case of Fantasy Premier League. *PLOS ONE*, *16*(3), e0246698. https://doi.org/10.1371/journal.pone.0246698

Pantuso, G. (2017). The Football Team Composition Problem: a Stochastic Programming approach. *Journal of Quantitative Analysis in Sports*, *13*(3), 113–129. https://doi.org/10.1515/jqas-2017-0030

Papageorgiou, G., Sarlis, V., & Tjortjis, C. (2024). An innovative method for accurate NBA player performance forecasting and line-up optimization in daily fantasy sports. *International Journal of Data Science and Analytics*. https://doi.org/10.1007/s41060-024-00523-y

Pontes, K. V., Embiruçu, M., & Maciel, R. (2016). Optimization of a Large Scale Industrial Reactor Towards Tailor Made Polymers Using Genetic Algorithm. *International Journal of Chemical Reactor Engineering*, *14*(1), 259–267. https://doi.org/10.1515/ijcre-2015-0041

Pretorius, A., & Parry, D. A. (2016). Human Decision Making and Artificial Intelligence. *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists*, 1–10. https://doi.org/10.1145/2987491.2987493

Rajesh, V., Arjun, P., Jagtap, K. R., M, S. C., & Prakash, J. (2022). Player Recommendation System for Fantasy Premier League using Machine Learning. *2022 19th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 1–6. https://doi.org/10.1109/JCSSE54890.2022.9836260

Ramdas, D. (2022). *Using Convolution Neural Networks to Predict the Performance of Footballers in the Fantasy Premier League*. https://doi.org/10.13140/RG.2.2.10010.72645/2

Rezende, M. C. A. F., Costa, C. B. B., Costa, A. C., Maciel, M. R. W., & Filho, R. M. (2008). Optimization of a large scale industrial reactor by genetic algorithms. *Chemical Engineering Science*, *63*(2), 330–341. https://doi.org/10.1016/j.ces.2007.09.001

Solopov, R. V., & Samulchenkov, A. S. (2019). Development of Fitness Function of Electrical Network Evolutionary Model for Computation of Settling Behavior by Method of Genetic Algorithm. *2019 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*, 1–5. https://doi.org/10.1109/ICIEAM.2019.8742968

Spiros Valouxis. (2023). *Machine Learning in Fantasy Premier League* [Master's Thesis, National Technical University of Athens]. https://doi.org/https://doi.org/10.26240/heal.ntua.25723

Vidgen, B., Chung, Y.-L., Johansson, P., Kirk, H. R., Williams, A., Hale, S. A., Margetts, H. Z., Röttger, P., & Sprejer, L. (2023). Tracking Abuse on Twitter Against Football Players in the 2021 – 22 Premier League Season. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4403913

Wafir Manakad. (2023, September 5). *Access Fantasy Premier League Data (23/24) on Google Sheets*. Medium. https://medium.com/@wafirmanakad/access-fantasy-premier-league-data-20-21-on-google-sheets-ee2cefce07cd

Wang, H., Li, J., Song, Y., Huang, J., Li, J., & Chen, Y. (2022). An Improved Genetic Algorithm for Team Formation Problem. *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, 774–781. https://doi.org/10.1109/SSCI51031.2022.10022143

Yin, Y. (2000). Genetic-Algorithms-Based Approach for Bilevel Programming Models. *Journal of Transportation Engineering*, *126*(2), 115–120. https://doi.org/10.1061/(ASCE)0733-947X(2000)126:2(115)

APPENDIX A

In this section, you'll find the results of the Exploratory Data Analysis (EDA) performed on the FPL dataset. It helped to understand the basic structure of our data, identify patterns, and uncover any anomalies or interesting trends.

Figure 11 – Counts of players per team in the premier league

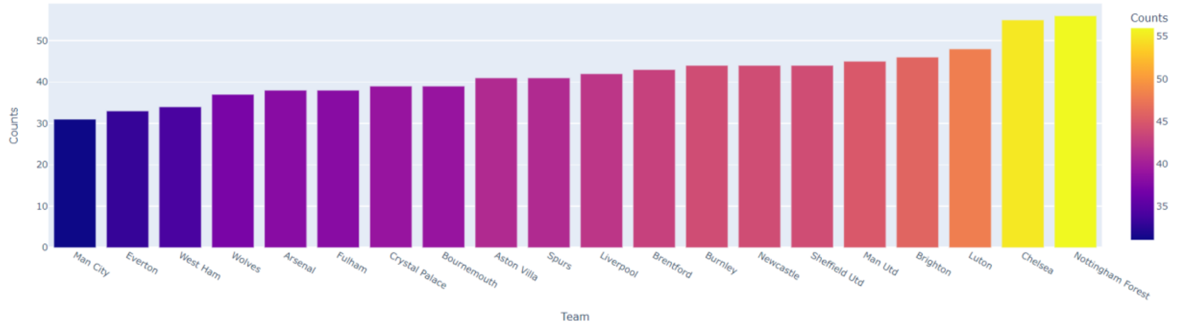


Figure 12 - Average points per game achieved by each team in the Fantasy Premier League

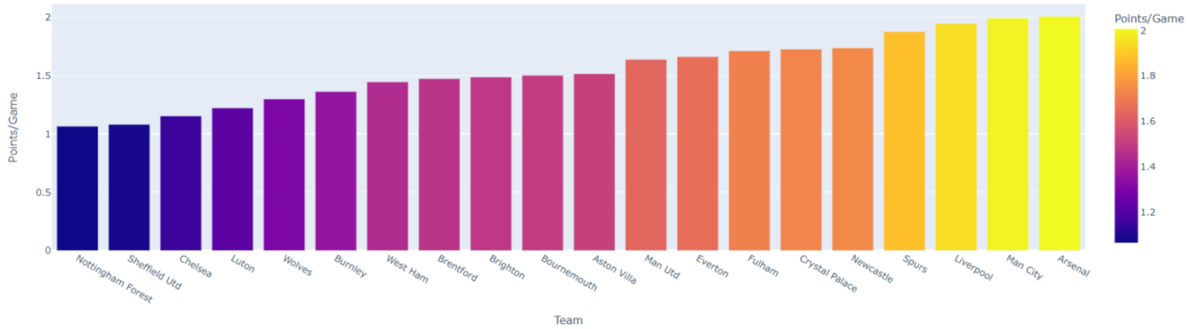


Figure 13 - Average price of players from each team in the Fantasy Premier League.

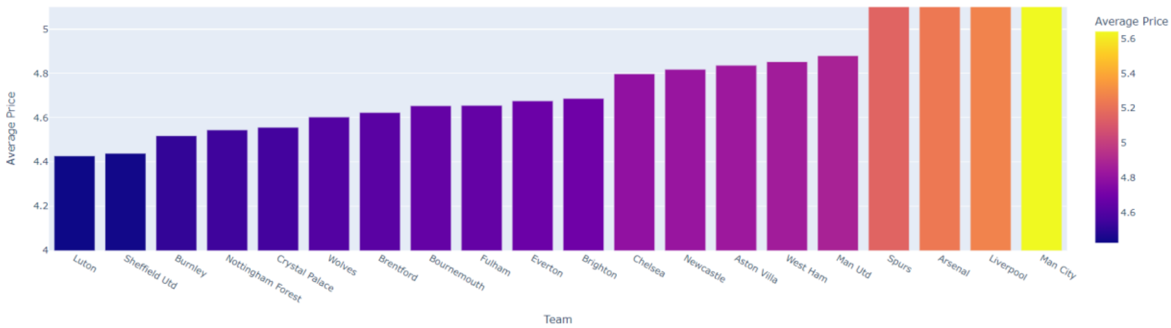


Figure 14 - Influence scores of players from each team in the Fantasy Premuer League

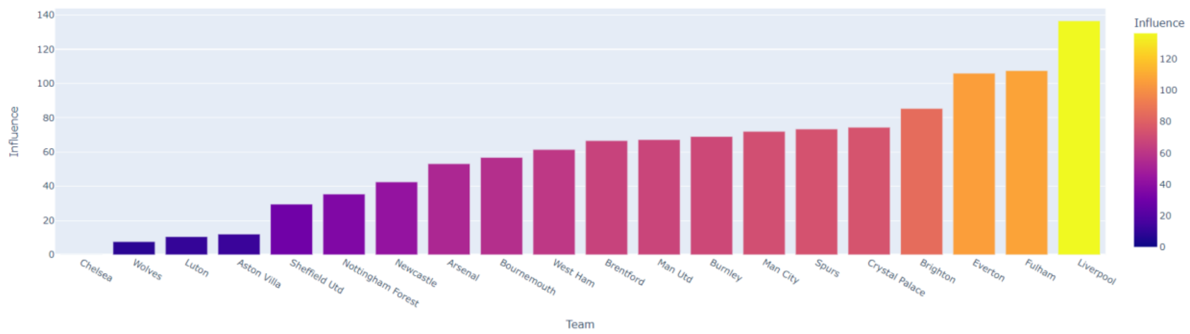


Figure 15 – Distribution of points per game in the Fantasy Premier League.

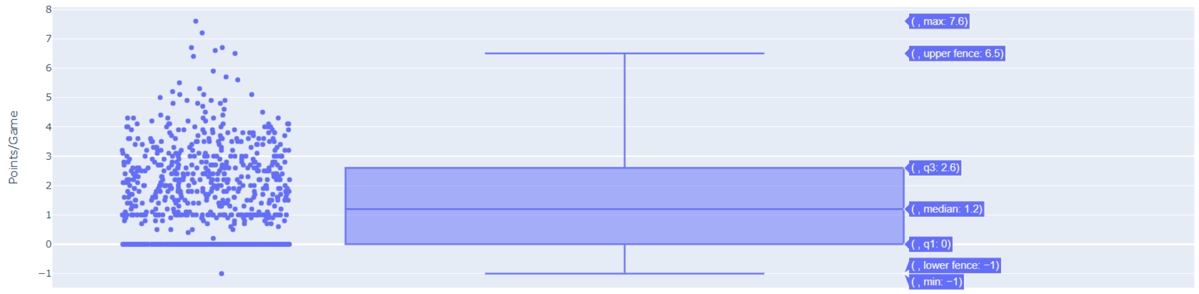


Figure 16 - Distribution of current player costs in the Fantasy Premier League.

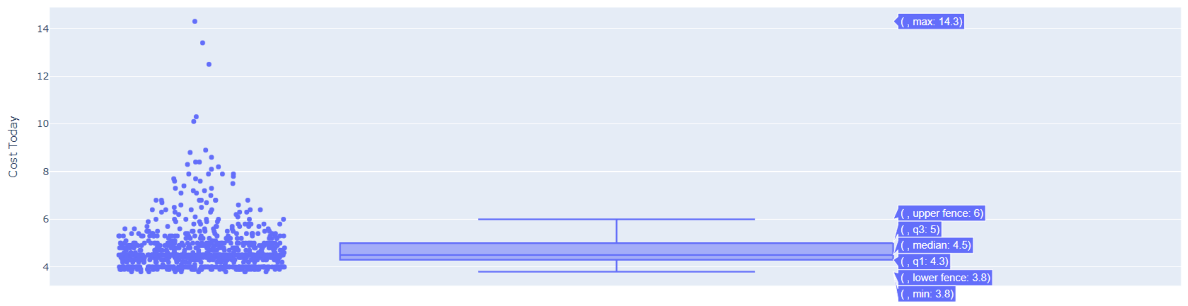


Figure 17 - Distribution of points per game across different player positions in the Fantasy Premier League.

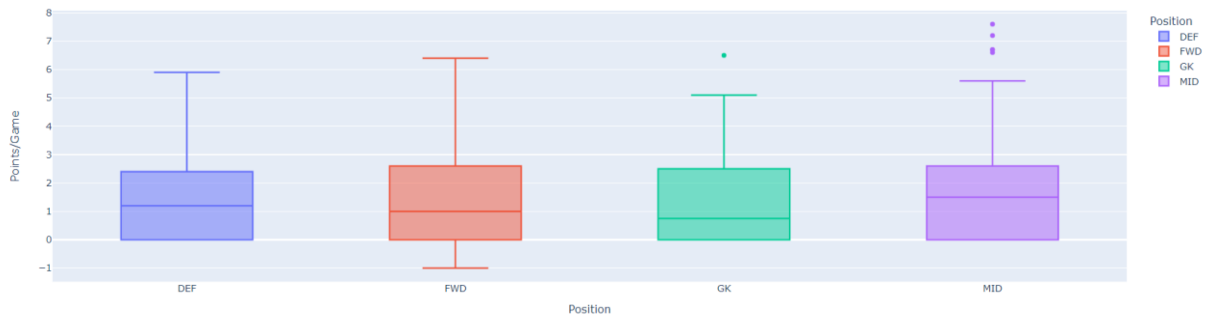


Figure 18 - This figure comprises four density plots, each representing different key metrics across player positions (FWD - Forward, DEF - Defender, MID - Midfielder, GK - Goalkeeper) in the Fantasy Premier League.

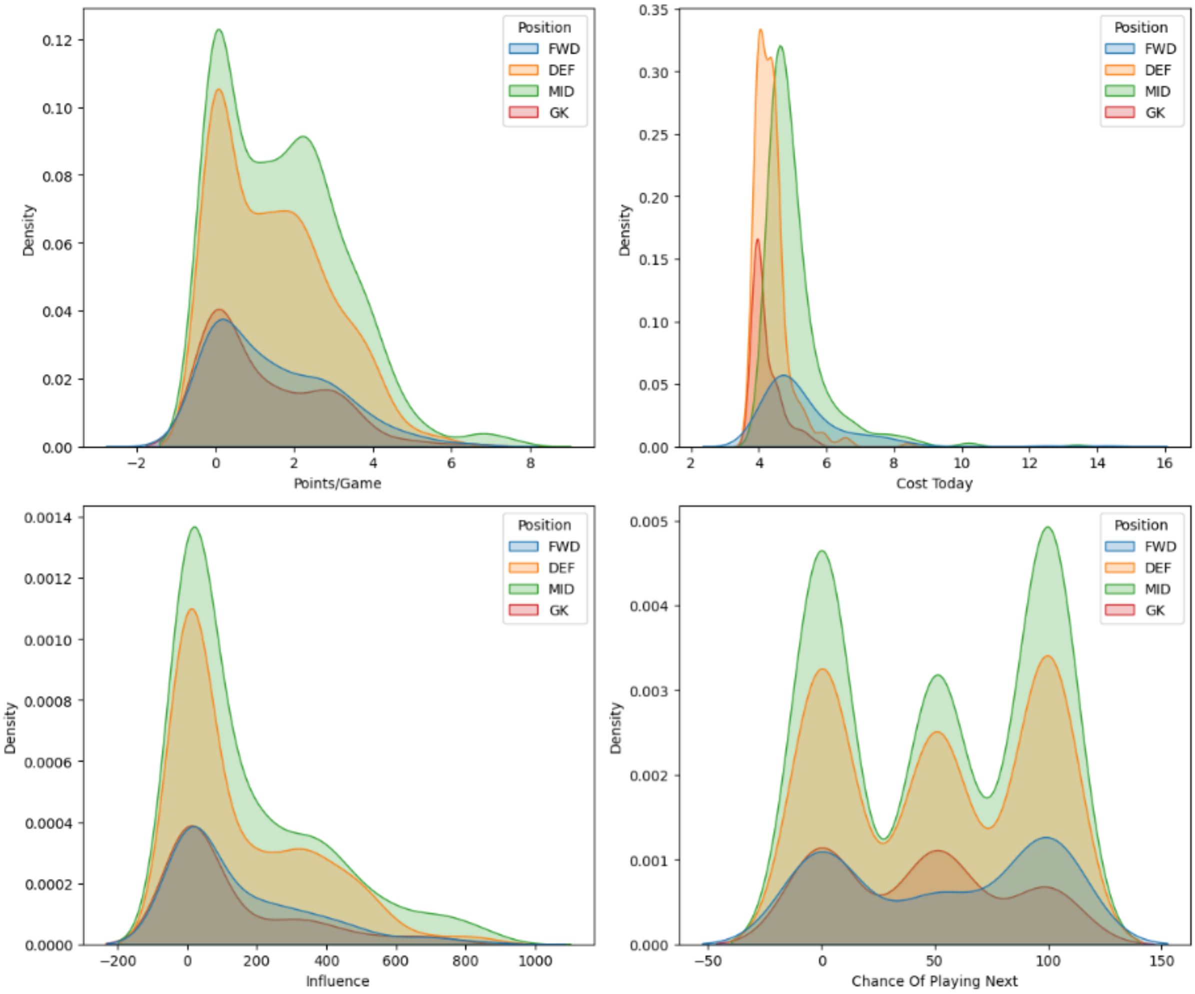


Figure 19 - This bar graph displays the top players offering the best value for money per points/game earned in the Fantasy Premier League (FPL)

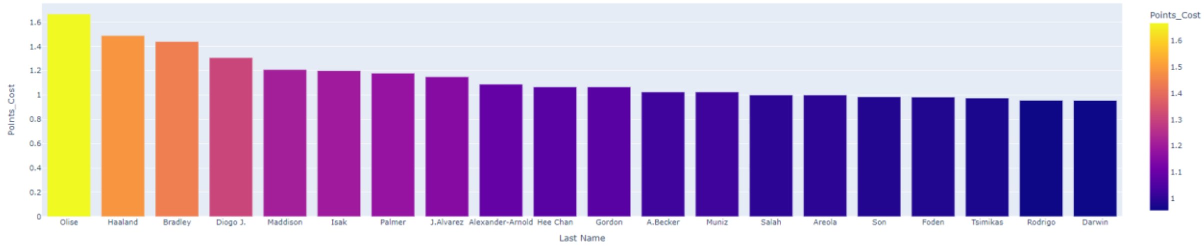


Figure 20 - Top undervalued players in the Fantasy Premier League (FPL) based on their form relative to their cost

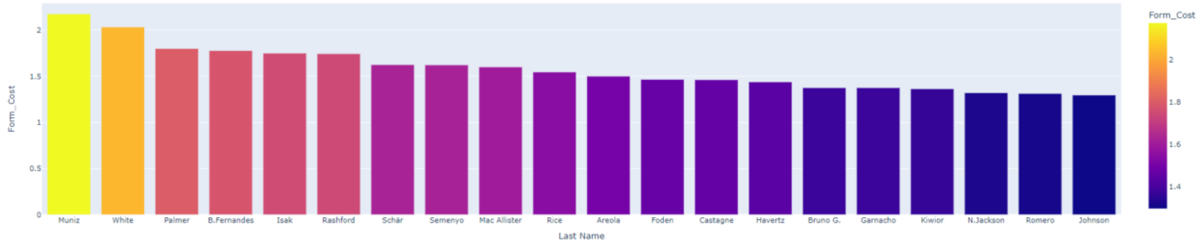


Figure 21 - Average form of teams in the Fantasy Premier League

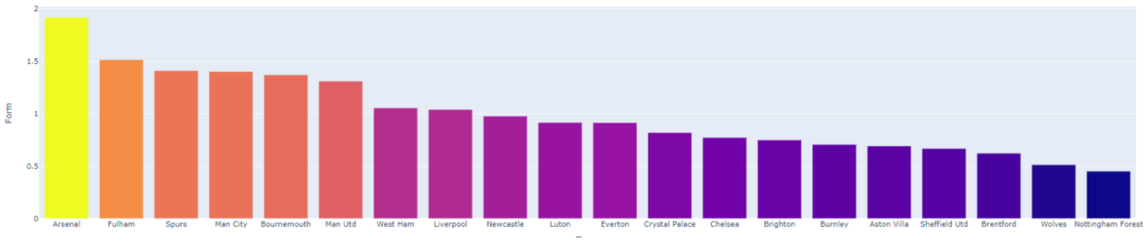
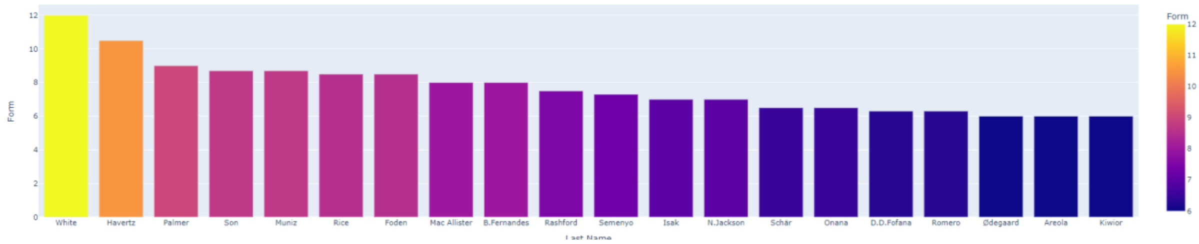


Figure 22 - This bar graph highlights the players with the highest form in the Fantasy Premier League



NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa