

MDSAA

Master Degree Program in
Data Science and Advanced Analytics

Optimizing Football Coaching Through Data-Driven Insights: A Power BI Dashboard Approach

Andreas Schneeweiss

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

**Optimizing Football Coaching Through Data-Driven Insights: A Power BI Dashboard
Approach**

by

Andreas Schneeweiss

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

Supervised by

Prof. Vitor Santos, Ph.D.

NOVA Information Management School

23rd of November, 2023

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.

Vienna, 23.11.2023

ACKNOWLEDGEMENTS

I would like to express my gratitude to my thesis advisor Professor Vitor Duarte dos Santos for his insightful feedback, and unending support throughout the thesis process.

Completing this project would have not have been possible without all the follow-up meetings and his guidance. I am also deeply thankful to my family and friends for their encouragement and patience during this academic journey.

ABSTRACT

Given the remarkable technological development over the years, football coaches are also recognizing the essential need for data-driven tools to improve their overall decision-making processes. This master thesis addresses this need by presenting a functional and intuitive data-driven football coaching dashboard. This tool is specifically designed to provide football coaches with analytical capabilities that can support them in their various tasks. The developed dashboard provides three different views of a national league assessment, a player performance evaluation and a detailed scouting platform. In addition, two advanced features are integrated: a Machine Learning model for predicting the market value of football players and a Recommendation System that suggests and identifies players with similar attributes. Moreover, this coaching tool can help football coaches optimizing their team performance and improving their budget allocation, player scouting and transfer negotiations. As football coaching is continuing to adapt to the digital age, this Power BI dashboard is proving to be a comprehensive toolkit for gaining a competitive advantage in this dynamic field and contribute to the ongoing development.

KEYWORDS

Business Intelligence; Sports Analytics; Machine Learning; Recommendation System

Sustainable Development Goals (SDG):



TABLE OF CONTENTS

1. Introduction.....	1
1.1. Context	1
1.2. Motivation	2
1.3. Objectives	2
2. Methodology	4
2.1. Design Science Research	4
2.2. Research Strategy	5
2.3. Chronogram of DSR methodology	6
3. Literature Review.....	8
3.1. Sports Analytics in Football	8
3.1.1. Concept.....	8
3.1.2. Machine Learning in Football	9
3.1.3. Opportunities and Challenges	9
3.1.4. Football Coaching Tools.....	10
3.2. Recommendation Systems	11
3.2.1. Definition	11
3.2.2. Filtering Techniques	11
3.2.2.1. Content-Based Filtering.....	12
3.2.2.2. Collaborative Filtering	13
3.2.2.3. Hybrid Filtering.....	13
3.2.2.4. Challenges.....	14
3.3. Business Intelligence	14
3.3.1. Concept.....	14
3.3.2. Data Warehouse	16
3.3.3. Data Warehouse Architectures.....	17
3.3.4. Data Visualization Tools.....	20
3.3.4.1. Microsoft Power BI.....	21
3.3.4.2. Salesforce Tableau	22
3.3.4.3. Qlik.....	22
4. Conceptual Model.....	23
4.1. Key Findings and Assumptions	23
4.2. Conceptual Model Proposal	24

4.2.1. Dataset Selection and Description	25
4.2.2. Feature Implementation.....	26
4.2.2.1. Machine Learning Feature.....	26
4.2.2.2. Recommendation System Feature.....	33
4.2.3. Power BI Model & Design.....	36
5. Conceptual Model Use Case – SL Benfica Lisbon	45
5.1. Landing Page.....	45
5.2. Portuguese League Assessment View.....	46
5.3. SL Benfica Player Performance Evaluation View	47
5.4. Scouting Platform View	48
6. Model Evaluation.....	50
7. Conclusions.....	52
7.1. Synthesis of the developed work.....	52
7.2. Limitations	52
7.3. Future Work.....	53
Bibliographical References.....	54

LIST OF FIGURES

Figure 2.1 – DSR Methodology (Peffer et al., 2007)	4
Figure 3.1 – Sports Analytics framework (Morgulev et al., 2018)	8
Figure 3.2 – Recommendation Systems Filtering Techniques (Ko et al., 2022)	12
Figure 3.3 – BI Framework (Watson & Wixom, 2007).....	15
Figure 3.4 – Inmon’s Data warehouse architecture (Yessad & Labiod, 2016).....	18
Figure 3.5 – Kimball’s Data warehouse architecture (Yessad & Labiod, 2016)	19
Figure 3.6 – Magic Quadrant for Analytics & BI Platforms (Gartner, 2023)	21
Figure 4.1 – Steps for Models Realization (Own Illustration)	24
Figure 4.2 – Machine Learning Steps (Own Illustration)	26
Figure 4.3 – Feature Selection technique with the best obtained results (Own Illustration)..	28
Figure 4.4 – Code Snippet XG Boost (Own Illustration)	31
Figure 4.5 – Best achieved metrics with XG Boost (Own Illustration)	32
Figure 4.6 – First part of the Code for Recommendation System (Own Illustration).....	34
Figure 4.7 – Second Part of Code for the Recommendation System (Own Illustration)	35
Figure 4.8 – Output of the Recommendation System feature (Own Illustration).....	36
Figure 4.9 – Dimension Table Club (Own Illustration)	38
Figure 4.10 – Dimension Table Player (Own Illustration).....	40
Figure 4.11 – Fact Table Fifa (Own Illustration)	42
Figure 4.12 – Star Schema for Power BI Model (Own Illustration).....	43
Figure 5.1 – Landing Page (Own Illustration).....	45
Figure 5.2 – Portuguese League Assessment View (Own Illustration).....	46
Figure 5.3 – Player Performance Evaluation View (Own Illustration).....	47
Figure 5.4 – Scouting Platform View (Own Illustration)	48

LIST OF TABLES

Table 2.1 – Chronogram (Own Illustration) 6

Table 4.1 – Metrics Overview for Model Evaluation (Own Illustration) 29

Table 4.2 – Metrics Overview for the applied Models (Own Illustration).....30

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ANOVA	Analysis of Variance
BI	Business Intelligence
CSV	Coma-Separated Values
DAX	Data Analysis Expressions
DSR	Design Science Research
ER	Entity Relationship
ETL	Extract, Transform and Load
FIFA	Fédération Internationale de Football Association
KNN	K-Nearest-Neighbor-Algorithm
KPI	Key Performance Indicators
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
OLAP	Online Analytical Processing
OTLP	Online Transaction Processing
SaaS	Software as a Service
USD	US Dollar
3NF	Third Normal Form

1. INTRODUCTION

1.1. CONTEXT

Football is known as one of the most popular and lucrative sports in the world, with clubs around the globe spending millions of dollars to acquire players who can help them win matches and trophies.

Over the last decades European football clubs have experienced a significant increase in transfer fees for the recruitment of new players (Richau et al., 2021). In 2022, English Premier League clubs spent a new record of 2.2 billion USD on transfer fees, exceeding for the first time ever the 2 billion USD mark (FIFA, 2022).

As a result, football clubs are under immense pressure to select and identify the right players in order to stay competitive, not only in terms of the financial aspect but also in sporting success within the football industry (Inan & Cavas, 2021).

Traditionally, football clubs have relied on human-based scouting methods for finding and evaluating prospective players. Nevertheless, these methods have their limitations, including high costs, the inability to effectively evaluate thousands of players, and the introduction of subjective bias. (Stanojević & Gyarmati, 2016)

However, Machine Learning algorithms can identify patterns and correlations that might be missed by coaches and scouts, by analyzing vast amounts of data (Herold et al., 2019). Therefore, Machine Learning models can contribute to traditional scouting methods and help football coaches to identify suitable players more accurately and efficiently.

Nowadays, technology is constantly advancing, and the collection of progressively larger data becomes feasible in a cost-effective and accessible manner. However, as the data volumes grow also the complexity is increasing. This development underlines the crucial requirement for effective tools such as visualization techniques for football coaches to handle the overall cognitive load when dealing with this data. By integrating such tools football coaches can focus on relevant information, leading to a better understanding of underlying facts. Despite this, there remains a lack of tools to facilitate the easy understanding of complex variables for coaches. Consequently, there is a rising demand for specialized tools that ideally offer the ability to effectively filter and present the data being analyzed in a way that improves the analytical capabilities of football coaches. (Benito-Santos et al., 2018)

Hence, the primary goal of this Master thesis is to conceptualize and develop an intuitive visualization tool in form of a Power BI dashboard tailored for the needs of football coaches. The designed tool focuses on the areas of i) a comprehensive league assessment, ii) an overall player performance evaluation, and iii) an insightful player scouting platform.

Additionally, a Recommendation System for identifying similar players and a Machine Learning model estimating players' market values is also incorporated, providing an extensive toolkit for football coaches.

1.2. MOTIVATION

Among the many decisions that football coaches are facing, one of the most important and challenging is the process of player selection in the current team as well as in the transfer market.

Creating as well as maintaining a competitive and successful football team requires a balance between numerous factors, such as the player's performance, experience, age, or the team's transfers policy (Poli et al., 2015). Therefore, evaluating and selecting the right players can have a big impact on the success, profitability and popularity of a football club.

Hence, there is a need to develop a systematic and data-driven tool that can improve and support the complex decision-making process, enabling football coaches to make well-informed choices that maximize their clubs' potential.

Moreover, the benefits of such a tool or dashboard are versatile. Football coaches can use the dashboard for instance to assess and compare the strengths and weaknesses of current players. Consequently, coaches can use different data-driven insights to identify key areas for improvement, tailor training programs, and optimize the team strategies.

Ultimately, clubs can find in a huge scouting database hidden talents that traditional scouting methods might miss. This potentially leads to an optimized transfer policy which can result in cost effective player signings.

1.3. OBJECTIVES

The core objective of this master thesis is to provide a meaningful tool that supports and optimizes the decision-making process of football coaches realized through a Power BI dashboard. In addition, the created dashboard aims to empower coaches with enhanced insights that would improve their strategic choices, ultimately achieving more effective results in managing their football team.

By providing this analytical dashboard, the thesis contributes to the subject of data-driven coaching and management practices within the area of football.

In order to achieve this goal, the following objectives are defined:

- Conduction of a systematic literature review covering concepts related to relevant topics like Sports Analytics and Machine Learning in Football, Recommendation Systems and Business Intelligence.
- Definition of a concrete methodology for setting up a framework for the dashboard.
- Selection of an appropriate dataset for the visualization dashboard.
- Preparation and Manipulation of the obtained data.
- Application of different Machine Learning algorithms to predict the market value of players.
- Evaluation of the different Machine Learning algorithms performances in that context.
- Application and Evaluation of Recommendation System Algorithms for finding similar players.
- Implementation of the different visualizations to the Power BI dashboard in order to create an extensive tool for football coaches.
- Presentation of project limitations and further recommendations for enhancement.

2. METHODOLOGY

In this chapter the methodology for the thesis will be defined and further elaborated, as well as graphically presented for a better overview. This will ensure a comprehensive understanding of the chosen research approach for the reader.

2.1. DESIGN SCIENCE RESEARCH

This master thesis is following the Design Science Research methodology (DSR). According to vom Brocke et al. (2020) “Design Science Research (DSR) is a problem-solving paradigm that seeks to enhance human knowledge via the creation of innovative artifacts. Simply stated, DSR seeks to enhance technology and science knowledge bases via the creation of innovative artifacts that solve problems and improve the environment in which they are instantiated.”

Therefore, for the objective of creating an innovative tool to improve the decision-making of football coaches, this methodology is suitable, as the Power BI dashboard characterizes the artifact in that context.

In literature there exists a huge amount of different methods of the DSR methodology with each of the authors presenting similar approaches (Haj-Bolouri, 2015).

However, this master thesis will follow the method of Peffers et al. (2007). This method consists of the six consecutive steps in an iterative process. Below you can find figure 2.1 that gives an overview of the model.

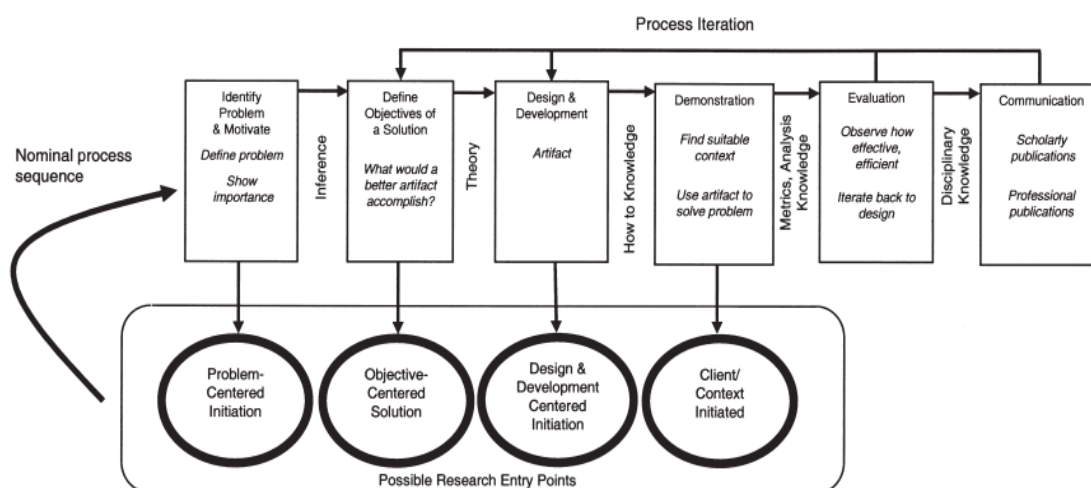


Figure 2.1 – DSR Methodology (Peffers et al., 2007)

The first step of the model is “Problem identification and Motivation”. In that step the specific research problem will be defined, and the significance of the solution will be demonstrated.

After that, the second phase, “Define Objectives of a Solution”, is about deriving the goals based on the problem definition in order to develop a solution.

Thirdly, the step of “Design & Development” is focusing on building the artifact including its desired functionality and its architecture.

The next step of “Demonstration” aims to point out the use of the artifact in solving the problem.

Then the “Evaluation” phase is concentrating on assessing the artifact's suitability in solving the defined problem by applying relevant metrics and analysis techniques.

Finally, the last step of the methodology “Communication” is centering on sharing the defined problem and its significance as well as the suitability of the designed artifact, to relevant stakeholders.

2.2. RESEARCH STRATEGY

The following section outlines the six steps that will be applied in order to achieve the thesis goal considering the methodology described before.

The first step will be to point out the concrete research problem by assessing the needs of a football coach nowadays, based on that, the usage for an interactive visualization tool and the integration of different features will be identified.

Secondly, the objectives of the artifact will be defined through focusing on optimizing the decision-making process of football coaches to maximize their team's potential.

The next step will consist of conducting a literature review in order to establish the theoretical foundation for this master thesis. This review will cover analyzing the main concepts of Business Intelligence, as well as the identified features of a Recommendation System and Machine Learning in football. After that the conceptual design of a Power BI dashboard will be defined and then developed through the insights gained from the literature review. Finally, the visualizations in the dashboard with different metrics and KPIs will be displayed using the chosen dataset.

Next, the fourth and fifth step are about demonstrating and testing the Power BI dashboard to evaluate its usefulness and functionality for improving and optimizing the coaching success. The evaluation will be based on different defined metrics as well as a qualitative and user-centric assessment.

The final step of this methodology is “Communication”. That step will be achieved through publishing the master thesis in the public repository of the Nova University Lisboa.

Each of those six steps represent a significant contribution for realizing the dashboard's potential and enhancing data-driven decision-making processes for football coaches.

2.3. CHRONOGRAM OF DSR METHODOLOGY

In this chapter a detailed chronogram for the applied methodology is provided associated with a timeline. This well-structured chronogram helped to plan and organize the research activities and ensured a sufficient time allocation to each of the defined steps as well as guaranteeing an effective time management.

The Table 2.1 below shows the chronogram according to the steps of the DSR methodology with the deliverables D1: First Draft of Dashboard, D2: Full Functional Dashboard and D3: Final Thesis Report.

Table 2.1 – Chronogram (Own Illustration)

Steps of DSR	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov
1.Problem Identification and Motivation	x	x							
2.Definitions of the Objectives		x	x						
3.Design & Development			x	x	x	x			
4.Demonstration						x	x	x	
5.Evaluation								x	x
6.Communication									x

Deliverables			D1: First Draft Dashboard			D2: Full Functional Dashboard			D3: Final Thesis Report
---------------------	--	--	---------------------------	--	--	-------------------------------	--	--	-------------------------

It's worth noting that, following the principles of the DSR methodology, an iterative process was applied in this thesis. However, this chronogram is giving an overview of the principal milestones achieved through the defined deliverables in a certain timeline.

3. LITERATURE REVIEW

In this chapter the literature review aims to summarize the relevant research for three different fields that are of importance for developing this master thesis.

First of all, there will be a comprehensive overview of sports analytics in football including the use of Machine Learning in football, the associated challenges, and opportunities, as well as football coaching tools. Next, a literature review about Recommendation Systems is being conducted covering the three main filtering types of Content-Based, Collaborative and Hybrid Filtering. Ultimately, the concept of Business Intelligence is being reviewed including Data warehouses, the different architectures, and the leading Data Visualization tools.

3.1. SPORTS ANALYTICS IN FOOTBALL

3.1.1. Concept

Sports Analytics is described as the systematic management of structured historical data, the deployment of predictive analytic models leveraging this data, and the utilization of information systems, in order to empower decision-makers, enabling them to guide their organizations towards achieving a competitive advantage on the field of play (Alamar & Mehrotra, 2011).

Morgulev et al. (2018) further expanded the sports analytics concept by defining a theoretical sports analytics framework, which is illustrated in Figure 3.1 below.

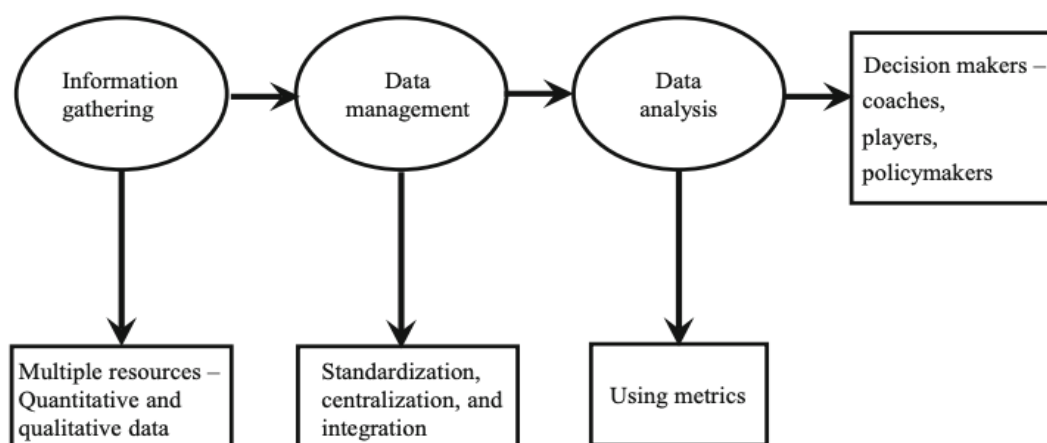


Figure 3.1 – Sports Analytics framework (Morgulev et al., 2018)

This framework shows the theoretical workflow for sport analytics. The initial step is focusing here on gathering information from multiple resources, followed by managing the data through standardization, centralization, and integration. Ultimately, the last step is analyzing the data with defined metrics in order to present it to the decision makers, like coaches or players.

3.1.2. Machine Learning in Football

In the area of sports analytics, Machine Learning stands as one of the fundamental approaches (Kapadia et al., 2019).

Through the ongoing advancements of the last decades in technology it is possible to collect a vast amount of data from athletes in different sports. Various devices such as wearable technology, video analysis systems, tracking systems, and questionnaires are now widely used to record and collect data in sports. Given this huge amount of data, the demand for complex models for analyzing the data are becoming essential. Therefore, the use of Machine Learning algorithms in sports science is on the rise. (Rossi et al., 2021)

With the increasing computational power, the better access to diverse data sources, and the wide applicability of Machine Learning algorithms, sport organizations have recognized new opportunities and strategies (Chmait & Westerbeek, 2021).

However, the football industry has realized the potential of Machine Learning relatively late compared to other sports, due to football's dynamic and uncontrollable nature, making data collection and analysis difficult (Tuyls et al., 2021).

Since then, several studies in the area of Machine Learning in football have been published, focusing on topics like injury prediction, physical performance prediction, club performance, and match attendance or match results prediction and betting (Nassis et al., 2023).

Nevertheless, the data's potential for supporting managerial decisions is still ignored. Especially, data-driven methods for estimating market value of football players have not yet gained acceptance in professional football. (Müller et al., 2017)

3.1.3. Opportunities and Challenges

One big future opportunity for Machine Learning applications in football is the field of explainable AI. As most of the studies about Machine Learning in football are performed by computer scientists, they are focusing on more complex models. As a consequence, those algorithms, particularly neural networks, can result in a black box, meaning that the output or

result is obtained without a clear understanding of which factors were most influential. This lack of transparency can hinder the interpretability of Machine Learning models, making it difficult to identify the most important factors. For instance, a passing model in football can be very precise in quantifying passes, but coaches cannot identify which are the most important factors that lead to a good pass, like pass velocity or position. Therefore, explainable AI models can be beneficial as they are visualizing and communicating the results in ways appropriate to the field of sports science and sports practice. Furthermore, another opportunity for Machine Learning in sports can be to tackle the lack of a clear learning curve. The current research is more focused on developing new and innovative approaches, rather than incorporating existing approaches to build better and improved models for practical use in football. (Herold et al., 2019)

However, there are also some challenges associated with Machine Learning in sports and especially football. As mentioned before, the main difficulties in using Machine Learning in football are the characteristics and rules of the sport. First of all, football has comparable to other sports a large number of active players and the huge size of the pitch makes it even more challenging to analyze than sports like tennis or basketball. (Tuyls et al., 2021)

In addition, Machine Learning approaches in sports are facing challenges due to confidential data about players, teams and commercial data that are not published to avoid providing other parties with competitive information. For the future, privacy regulations and data ownership issues must be addressed to ensure ethical use and sharing of sports data to improve Machine Learning algorithms and its adoption to sports. (Chmait & Westerbeek, 2021)

Additionally, to that point, Hammes et al. (2022) are also emphasizing the importance of networking and knowledge-sharing between Machine Learning and sports practitioners, in order to improve the development of Machine Learning applications in sport.

3.1.4. Football Coaching Tools

Nowadays, many different football clubs are spending a large amount of resources analyzing their game performances, by using analytical tools or manual methods (Halvorsen et al., 2013).

However, football analytics companies have only relatively recently begun to collect big data like player tracking or pose information (Tuyls et al., 2021).

Moreover, the existing football analysis tools are providing very basic statistical insights into the match and the performance of the players. The typical information provided includes metrics like player distances covered, top and average player speeds, pass counts or shot

attempts. In order to uncover deeper statistical insights and gain a better understanding of the data, there is a strong demand for more advanced and sophisticated analytical tools in that field. (Gudmundsson & Wolle, 2012)

Especially as already stated before, there is a need for having effective visualization tools for football coaches to manage the cognitive load of dealing with complex data. By integrating visualization tools, it would help football coaches focusing on important data, supporting an overall better understanding. However, in that area is still a lack of tools that can simplify complex data through visualizations. (Benito-Santos et al., 2018)

3.2. RECOMMENDATION SYSTEMS

This chapter focuses on the concept of Recommendation Systems and the three main filtering techniques. Moreover, the challenges regarding Recommendation Systems are also discussed.

3.2.1. Definition

According to Sharma & Gera (2013) Recommendation Systems are “a subclass of information filtering system that seek to predict ‘rating’ or ‘preference’ that a user would give to an item (such as music, books or movies) or social element (e.g. people or group) they had not yet considered, using a model built from the characteristics of an item (content based approaches) or the user’s social environment (collaborative filtering approaches).

3.2.2. Filtering Techniques

The existing recommendation techniques can be categorized in to three major categories (Sharma & Gera, 2013):

- Collaborative Filtering
- Content Based Filtering
- Hybrid Recommendations

Figure 3.2 below displays the process for each of the three classified filtering techniques that will be described in detail in the next chapter.

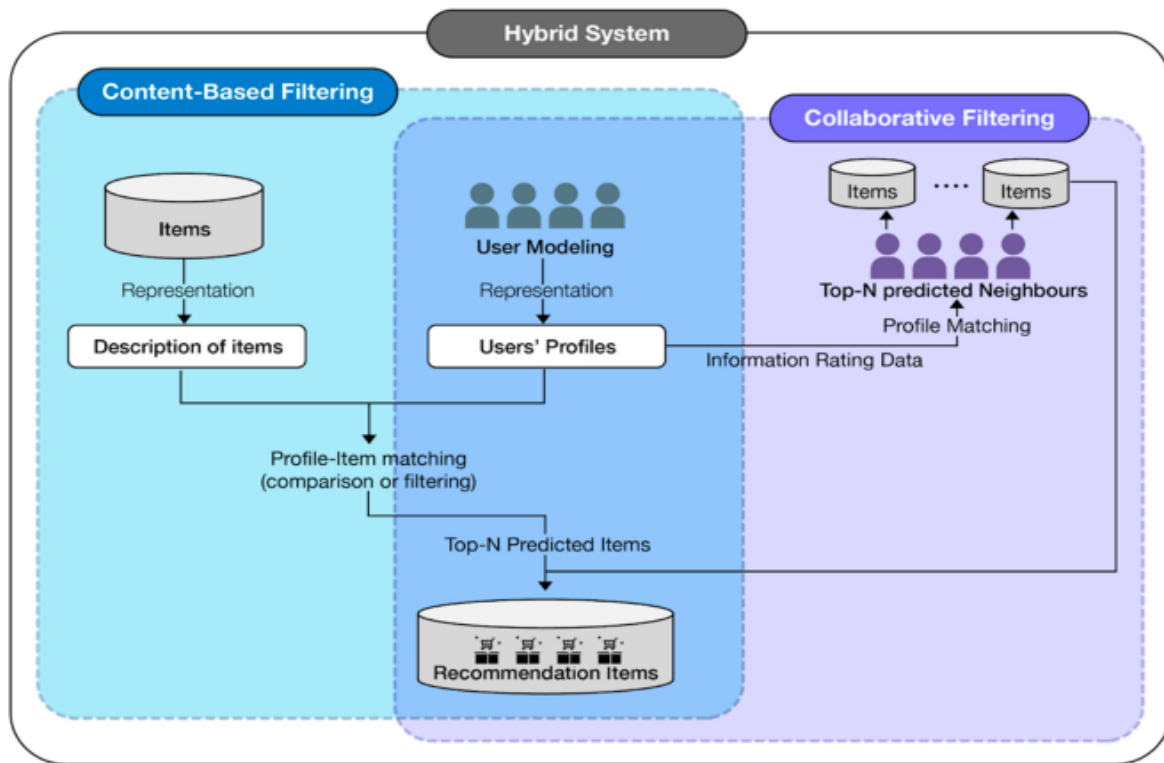


Figure 3.2 – Recommendation Systems Filtering Techniques (Ko et al., 2022)

3.2.2.1. Content-Based Filtering

The first main filtering method is called Content-Based Filtering. This technique focuses on suggesting similar items based on information of items previously chosen by the user. (Ko et al., 2022)

Therefore, Content-based filtering tries to recommend items based on similarity measurements (Thorat et al., 2013).

Furthermore, Content-based filtering techniques do not rely on the profile of other users as they are not impacting the recommendation. Even if the user profile is changing, Content-based filtering techniques still have the potential to rapidly adopt its recommendations. However, the main disadvantage of this method is the need for a comprehensive understanding and detailed description of the item features in the user's profile. (Isinkaye et al. 2015)

This particular filtering technique is for instance used for recommending web pages, TV programs or news articles (Sharma & Gera, 2013).

3.2.2.2. Collaborative Filtering

The second main filtering method is the Collaborative Filtering. This technique generates personalized recommendations of items by analyzing patterns of user ratings or usage, such as purchases, without requiring exogenous information about either items or users. (Koren et al. 2021)

The fact that Collaborative Filtering only relies on users' ratings on items has its advantages and disadvantages. On the positive side, Collaborative Filtering doesn't require the intensive analysis of content-rich items. Nevertheless, this advantage can be a weak point as Collaborative Filtering can provide unexpected recommendations in situations where items appeared suitable to a user but they don't align with the actual user profile. (Do et al., 2010)

There are some areas where Collaborative Filtering outperforms Content-Based Filtering. In particular in domains where there is limited content associated with items or where content analysis is challenging for computer systems, such as opinions and ideals. (Isinkaye et al. 2015)

3.2.2.3. Hybrid Filtering

Both of the above described filtering techniques have their limitations. The Content-Based Filtering model relies on user item metadata, while Collaborative Filtering relies on user item rating data. (Ko et al., 2022)

To address these limitations and in order to improve the recommendation performance a Hybrid recommendation model was introduced (Basilico & Hofmann, 2004).

In general, there exists several hybrid filtering techniques for combining Content-Based Filtering and Collaborative Filtering, depending on the specific domain and the characteristics of the data. These different hybrid methods include (Sherma & Gera, 2013):

- Collaborative Filtering and Content-Based techniques are implemented separately, and their predictions are combined.
- Integrating Content-Based elements into the Collaborative Filtering approach.
- Incorporating Collaborative elements into the Content-Based Filtering approach.
- Building a comprehensive unified model that incorporates both Content-Based and Collaborative Filtering characteristics.

Many hybrid approaches primarily build upon Collaborative Filtering, but Content-Based techniques play a crucial role in maintaining user profiles and such profiles are used to identify users with similar preferences (Sherma & Gera, 2013).

3.2.2.4. Challenges

There are some known challenges and hurdles in those filtering techniques. First of all, one of the major problems in Recommendation Systems is data sparsity. The main reason for that issue is that the majority of users are not rating items and therefore the available ratings are usually sparse. Collaborative Filtering is especially affected by that problem as it relies on the rating matrix for recommendations. Another common challenge is the cold start problem that occurs when new users or items just enter the system, resulting in a lack of sufficient information for recommendations. In this case, Collaborative Filtering cannot make useful recommendations due to the lack of relevant data. Content-Based Filtering methods on the other hand, are still able to provide recommendations for new items, as they do not rely on previous user ratings. (Sherma & Gera, 2013)

3.3. BUSINESS INTELLIGENCE

In this chapter the Concept of Business Intelligence as well as an overview about the Data warehouse, the different Architectures and the three leading Data Visualization tools are being discussed and presented.

3.3.1. Concept

The rapidly changing business landscape, marked by factors such as globalization or technological innovation, has forced companies to reassess their business strategies. Within this competitive environment, Business Intelligence plays a crucial role in facilitating the decision-making process to enhance competitiveness and establishing an effective link between business strategies and IT. (Khan & Quadri, 2012)

Business Intelligence can be described as the transformation of data into information and subsequently into knowledge. The gained knowledge is typically about customer requirements, customer decision-making processes, the competition, industry conditions, as well as general economic, technological, and cultural trends. (Golfarelli et al., 2004)

The primary goal of BI is to support in managing the extensive flow of business information both within and outside the organization. This is achieved by initially identifying the

information and subsequently processing it into concise and useful managerial knowledge and intelligence. (Rouhani et al., 2012)

Hence, Business Intelligence includes various tasks like (Khan & Quadri, 2012):

- Generating forecasts by analyzing historical data, past and present performance, and estimation of future trends.
- Conducting "What if" analysis to assess the impacts of changes and alternative scenarios.
- Enabling ad hoc data access to answer specific, non-routine questions.
- Providing strategic insights.

The technology for Business Intelligence is continuously evolving and improving to address more and more complex business challenges. Among the most applied Business Intelligence enabling technologies are data warehousing, on-line analytical processing (OLAP), and data mining. Those Business Intelligence technologies are aiming to empower individuals to make optimized business decisions by delivering accurate, up-to-date, and relevant information precisely when it is needed in order to gain sustainable competitive advantage. (Khan & Quadri, 2012)

According to Watson & Wixom (2007) "Business Intelligence is a process that includes two primary activities of "Getting data in" and "Getting data out" ".

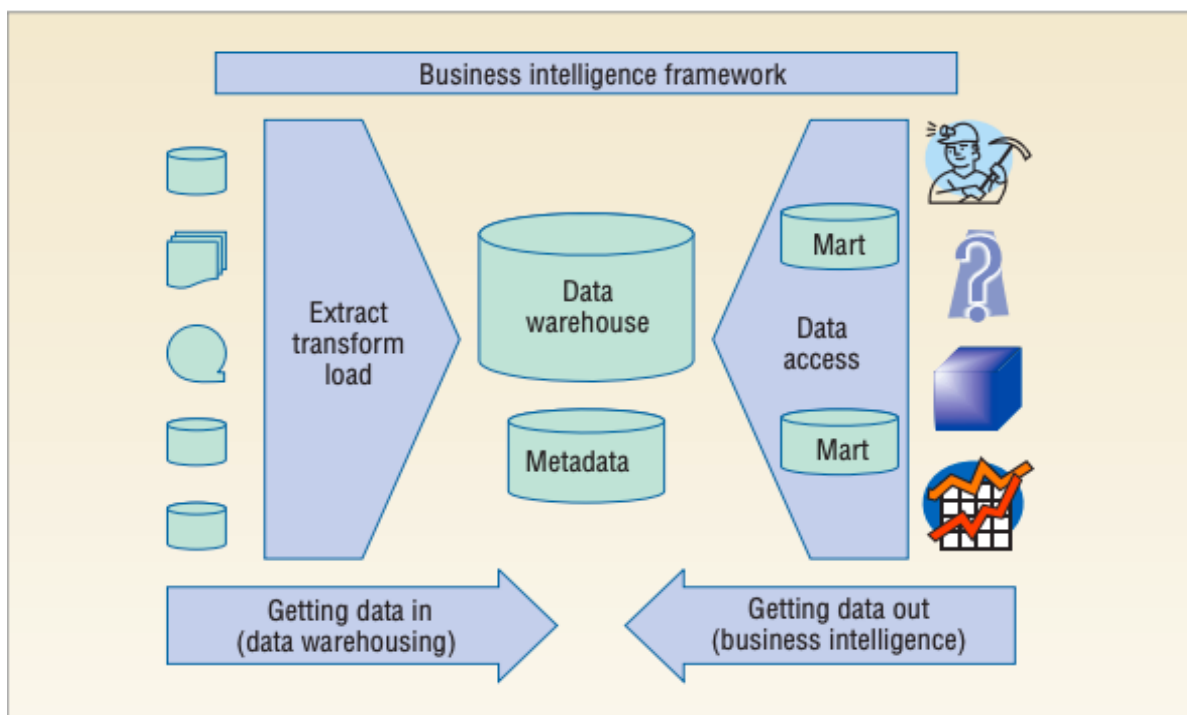


Figure 3.3 – BI Framework (Watson & Wixom, 2007)

As figure 3.3 above illustrates with the framework in place, the process unfolds as follows:

- Information is extracted from the numerous information systems within the organization, from an external data provider, or from a business partner.
- This information undergoes the Extract Transform and Load process (ETL) and is subsequently stored within the Data warehouse.
- Depending on the architecture, Data warehouses may feed dependent data marts with narrower scopes. Metadata is essential in data warehousing due to the complexity of data migration, as it is describing technical and business aspects and is providing transparency throughout the data flow process.
- Business Intelligence then takes charge, collecting the data and converting it into valuable information and then knowledge. This is achieved by accessing data from the Data warehouse to perform enterprise reporting, OLAP, querying, and predictive analytics.

3.3.2. Data Warehouse

The framework's first activity of "Getting data in", is known as data warehousing, which involves transferring data from a collection of source systems into an integrated Data warehouse. This process is the most challenging aspect of Business Intelligence, demanding about 80 percent of the time and effort while generating more than 50 percent of the unforeseen project costs. These challenges arise from multiple factors, including poor data quality in the source systems, disagreement about data ownership, and legacy technology. (Watson & Wixom, 2007)

In that context it is crucial to differentiate a Data warehouse from data warehousing. According to Inmon (2005) "a Data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of managements decision-making process."

Data warehouses are typically kept separate from the organization's operational databases. Several reasons underline this segregation. Data warehouse supports online analytical processing (OLAP), which has different functional and performance requirements compared to the online transaction processing (OLTP) applications typically supported by operational databases. Operational databases handle repetitive day to day tasks like order entry, and consist of short, atomic, isolated transactions, while Data warehouses, which store consolidated historical data, are significantly larger and designed for query-intensive, ad hoc analysis for the decision support. Therefore, Data warehouses prioritizing, query throughput and response times over transaction throughput. In addition, Data warehouses are typically

modeled multidimensionally, in order to facilitate complex analyses and visualization. (Chaudhuri & Dayal, 1997)

3.3.3. Data Warehouse Architectures

The two most important approaches for Data warehouse architectures are the dimensional and the normalized approach. The normalized approach associated with Bill Inmon, is focusing on modeling the Data warehouse by using an E-R (Entity Relationship) model or normalized model and is often referred to as the 3NF model (Third Normal Form). (Shobirin et al., 2017)

Inmon's approach consists of all information systems and their databases throughout a given organization. Inmon also proposes that a physical separation exists between a Data warehouse and a Data mart. There are four different levels in the overall database environment of the organization (Inmon, 2005):

- the operational level
- the atomic or the Data warehouse level
- the departmental (or the Data mart level)
- the individual level

These different levels of data represent the foundation of a larger architecture called the corporate information factory. Whereas the last three levels concern the Data warehouse and the first level, the operational level supports day to day operations and transactional processes. The data is then transformed and loaded into the atomic Data warehouse by applying the ETL process of Extracting, Loading and Transforming. (Inmon, 2005)

To summarize, Bill Inmon suggests a top-down development approach which involves adapting traditional relational database tools to the development requirements of an enterprise-wide Data warehouse. This enterprise-wide data store then serves as the foundation of individual departmental databases that are developed to improve most decision support needs. (Breslin, 2004)

An overview of this approach is displayed below in Figure 3.4.

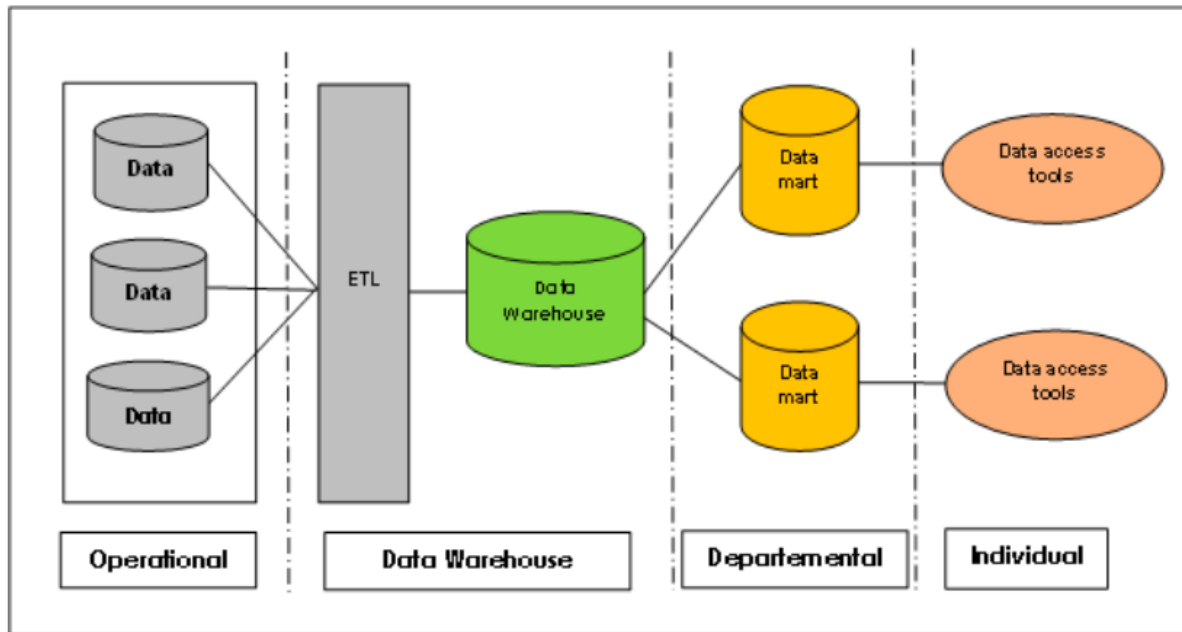


Figure 3.4 – Inmon's Data warehouse architecture (Yessad & Labiod, 2016)

In contrast, the approach of Kimball states to model the Data warehouse using a dimensional Model in specific a Star schema (Shobirin et al., 2017).

This approach relies on dimensional modeling, a data modeling technique unique to data warehousing and follows a bottom-up methodology. Instead of forming a single enterprise-wide database, Kimball suggests creating one database (or data mart) for each major business process. The sum of those Data marts constructs the organization's Data warehouse. Moreover, enterprise-wide cohesion is ensured through using a data bus standard innovation. (Breslin, 2004)

Figure 3.5 below illustrates the overall architecture of Kimball's approach.

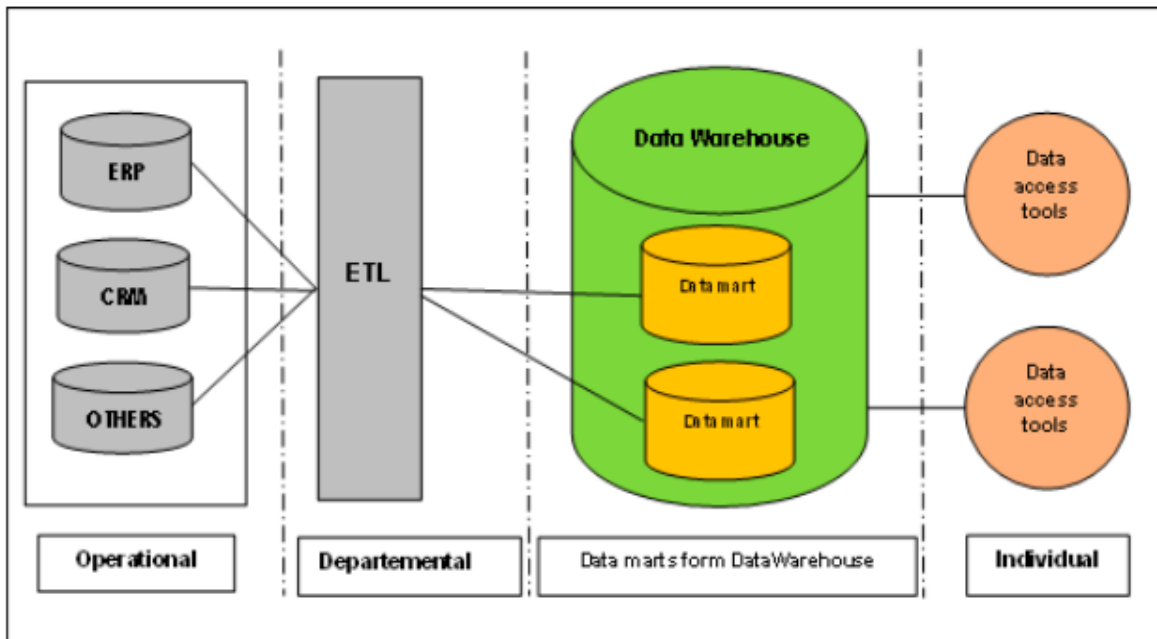


Figure 3.5 – Kimball's Data warehouse architecture (Yessad & Labiod, 2016)

Kimball's approach includes four key decisions during the design of the dimensional model:

1. Select the business process.
2. Declare the grain.
3. Identify the dimensions.
4. Identify the facts.

Fact tables in that context are the essence of the Data warehouse and express many to many relationships. Facts represent a business measure and are typically described as continuously valued attributes, which helps distinguish them from dimension attributes. The most valuable facts are those that are numeric and additive. In addition, all fact tables have two or more foreign keys that are linked to the dimension tables' primary keys. (Kimball & Ross, 2013)

Dimensions tables, according to Kimball & Ross (2013), "contain the textual context associated with a business process measurement event. They describe the "who, what, where, when, how, and why" associated with the event."

Because of the similarity of dimensional models to a star-like structure, Kimball & Ross (2013) define them as Star schemas and recommend that detailed and atomic information should be loaded into Star schemas.

In contrast, when dimension tables are being normalized, it results in what is known as the Snowflake schema. Kimball & Ross (2013) state that “Snowflaking is a legal extension of the dimensional model, however, we encourage you to resist the urge to snowflake given the two primary design drivers: ease of use and performance.”

Star schemas have the advantage of reducing the number of tables in a database as well as the number of relationships between them and hence the number of joins required in user queries (Moody & Kortink, 2000). This is one of the reasons why the Star schema is typically considered to be the most efficient Data warehouse architecture (Martyn, 2004).

Inmon’s and Kimball’s architecture have certain similarities, like the treatment of time-stamped data. Even though there are differences in the ways in which each of the two models tackle this challenge, the two models are more alike than not in modeling the time attribute. Likewise, both approaches address the complexity of transforming operational data in a similar way. This transformation process, named ETL (Extract, Transform and Load), represents one of the most labor-intensive aspects of the Data warehouse. However, there are essential differences between the model of Kimball and Inmon. Specifically in the areas of development methodologies, data modeling and the overall Data warehouse architecture. (Breslin, 2004)

3.3.4. Data Visualization Tools

A fundamental role for increasing the competitive advantage of organizations is the autonomy to visualize and analyze existing data and the knowledge derived from it. In this context, Data Visualization tools serve as great aids in decision-making, by leveraging Self-Service Business Intelligence techniques. These tools are designed to enhance agility across different business departments, empowering users to generate their own reports and access data without depending on specialists. However, the right choice of tools must be based on a deep analysis of each company's unique requirements and needs. (Lousa et al., 2019)

By analyzing Gartner Magic Quadrant for Analytics and Business Intelligence Platforms 2023 (Figure 3.6), we can define the current leading tools as:

- Microsoft Power BI
- Salesforce Tableau
- Qlik



Figure 3.6 – Magic Quadrant for Analytics & BI Platforms (Gartner, 2023)

3.3.4.1. Microsoft Power BI

Microsoft is the Leader in this Magic Quadrant with its tool Power BI. With the comprehensive and visionary product roadmap it reaches a massive market reach through its Microsoft Office channel (Gartner, 2023).

Microsoft has developed Power BI from a collection of Excel add-ons into a new software and service package (Becker & Gould, 2019).

Power BI is offering data preparation, visual-based data exploration, interactive dashboards and augmented analytics. Power BI can be accessed as a Software as a Service (SaaS) solution running in the Azure cloud or as an on-premises option in Power BI Report Server. Moreover, Power BI Desktop, a stand-alone and free personal analysis tool, is required for power users who are authoring complex data mashups involving on-premises data sources. (Gartner, 2023)

The fundamental layer of Power BI is Power Query also known as Get and Transform, which provides functions for accessing tabular data in a diverse variety of formats, ranging from large-scale databases to basic text files. Furthermore, Power Query is introduced through a graphical user interface window that appears when launched from Excel or Power BI Desktop

and is essentially serving as an interface for a scripting language known as "M." After the data have been shaped and configured by Power Query, these tables can be loaded into the Power BI visualization layer. Within a specific set of visualizations, multiple tables can be employed and linked together in a "data model" comparable to those found in traditional relational database systems. The Power BI report contains one or more pages where you can organize visuals by a drag-and-drop process. In addition, standard descriptive statistics like sum totals, counts, minimums, or maximums are as measures for visualizations available. For more sophisticated calculations, Power BI offers Data Analysis Expressions (DAX), which shares similarities with Excel's formulas and is adapted for tabular data and dynamic filtering. (Becker & Gould, 2019)

3.3.4.2. Salesforce Tableau

The second tool is Tableau from Salesforce and is characterized as a Data Visualization and Business Intelligence software that supports the decision-making process of organizations, by providing effective visualization dashboards. Moreover, Tableau is the fastest growing Data Visualization tool in the Business Intelligence Industry. (Balaji et al., 2021)

For these reasons, Tableau is the second leader in this Magic Quadrant. The tool provides a visual-driven exploration experience enabling business users to access, prepare, analyze, and communicate their findings in the data. (Gartner, 2023)

3.3.4.3. Qlik

The third leader in the Magic Quadrant is Qlik. It provides a strong product vision for ML- and AI driven augmentation but compared to its main competitors, Tableau and Power BI it has a lower market momentum. Qlik's engine enables different skilled users to combine data as well as exploring information without the constraints of query-based tools and also adds Artificial Intelligence and Machine Learning functionalities. (Gartner, 2023)

4. CONCEPTUAL MODEL

The following chapter is separated into two distinct sections describing the development process for the data-driven dashboard tailored for the needs of football coaches.

The first part is focusing on the fundamental assumptions the dashboard will be based on, gained through the key findings from the previous conducted literature review. The second section presents in detail the conceptual model steps of the proposed dashboard, including the dataset description, feature implementations, Power BI Modeling and design, in accordance with the third phase of the Design Science Research methodology.

4.1. KEY FINDINGS AND ASSUMPTIONS

Based on the main findings of the extensive literature review that was conducted in the chapter before, several assumptions can be derived that must be considered for developing the intended dashboard:

1. There is a strong demand for more advanced and sophisticated football coaching tools, as the existing tools are only covering basic statistical insights about match and performance.
2. There is especially a lack for visualization tools for football coaches in order to manage the cognitive load of complex data and enhancing their analytical abilities.
3. The potential of data for supporting football managerial decisions is still ignored.
4. Data-driven approaches for estimating market value of football players have not yet gained acceptance in professional football.
5. For the usage of Machine Learning, Explainable AI is recommended in order to understand the most influential factors and improve the interpretability of the models.
6. Machine Learning approaches in sports need to address the lack of a clear learning curve by incorporating existing approaches rather than developing completely new models.
7. Collaborative Filtering only relies on users' ratings on items whereas Content-Based Filtering tries to recommend items based on similarity measurements.
8. There are two competing Data warehousing architectures Inmon approach and the Kimball approach.

9. Star schema of Kimball's approach is considered as the most efficient Data warehouse architectures as it reduces the number of tables and relationships, reducing the need for complex joins in user queries.
10. The key for improving competitive advantage of organizations is the ability to independently visualize and analyze existing data and the knowledge derived from it by using Data Visualization tools.
11. Microsoft's Power BI is the leading Data Visualization Tool for Business Intelligence.

4.2. CONCEPTUAL MODEL PROPOSAL

In the following section, the steps will be described that need to be applied in order to realize the conceptual model for the data-driven football coaching dashboard. Overall, there were three steps identified, in order to create the final dashboard.

The first step will be covering the dataset selection as well as describing the data that is being used for developing the dashboard. Next the feature implementation step will present the two developed features for the dashboard: the Machine Learning prediction for the player's market value and the Recommendation System for similar players. The last step will be demonstrating the proposed dimensional Power BI model and will also focus on the detailed design of the dashboard.

Below Figure 4.1 provides an illustration that represents an overview of all three consecutive steps.



Figure 4.1 – Steps for Models Realization (Own Illustration)

4.2.1. Dataset Selection and Description

In this master thesis, data of the FIFA football video game was chosen, which is already commonly utilized in the literature. For each year new datasets based on the video game are available.

The EA Sports FIFA video game series is offering very detailed information about football players, which covers physical, mental, and technical skills. Since the year of 2014, researchers as well as clubs have used the FIFA video game as an alternative source for data. (Al-Asadi & Tasdemir, 2022)

This FIFA data has especially proven valuable in Machine Learning tasks for solving real problems, with very accurate results achieved (Shin & Gasparyan, 2014).

EA Sports has established a wide and extensive network of real-life scouts. These scouts are attending matches and reviewing game footage to assess the skill scores for each player in the game. Additionally, the skill scores can change with each new version of the FIFA game series. For instance, if a young player demonstrates improvement in shooting or dribbling during a real-life season, his score will be increased in the next edition of the FIFA game. Whereas if an older player experiences a decline in speed or stamina, his change will also be reflected in their score. Consequently, a score for each of these various criteria is defined to every player, on a scale up to 100. (Al-Asadi & Tasdemir, 2022)

In this master thesis, the FIFA 21 dataset for the dashboard visualization was used, provided on Kaggle. The FIFA 21 data contained 18.944 male players with 110 different attributes. The data was being extracted as a CSV File.

For the visualization part the FIFA 21 dataset was further processed, to guarantee a high data quality. A unique player id was added for each player and the following columns were dropped as they had a high percentage of missing values: "club_loaned_from" - 96%, "nation_team_id"- 94%, "nation_position" - 94%, "nation_jersey_number" - 94%, "player_tags" - 93%, "player_traits" - 56%, "goalkeeping_speed" - 89% and "nation_logo_url" - 94%.

The insertion of the unique player id was particularly important as there were some players with the same names. Moreover, just the first mentioned position of each player was kept and mapped to either being a Goalkeeper, Defender, Midfielder or Forward. Ultimately, unnecessary columns were deleted like "sofifa Id" or abbreviation columns without a meaning.

The resulted outcome was the Prepared FIFA Data CSV File that was later used as basis data for the visualization dashboard.

Ultimately, the FIFA 20 dataset was applied as a train file for the Machine Learning prediction of the players market value, which is also provided on Kaggle. This dataset was also a CSV File and contained 18,483 male players with the same 110 different attributes.

4.2.2. Feature Implementation

As previously stated in the literature review, most of the existing coaching tools are just based on simple statistical insights about matches and player's performance. Based on those findings, a gap was identified, resulting in a need for more sophisticated football coaching tools. Therefore, two advanced features were implemented to the dashboard by applying Machine Learning models and a Recommendation System in a Jupyter Notebook environment, in order to provide a more advanced toolkit for football coaches.

4.2.2.1. Machine Learning Feature

Given the information gained from the literature review, the potential of data for supporting football managerial decisions is still ignored, especially for estimating the market value of football players. The feature of a Machine Learning prediction for the market value of each football player was incorporated aiming to close this gap.

This feature not only allows coaches to assess the monetary value in the context of transfers but also empowers them to evaluate if players might be undervalued or overvalued in their market value.

The prediction task was carried out by using the FIFA 20 dataset as the train file and the FIFA 21 dataset accordingly as a test file for evaluating the achieved results. In accordance with the findings in the literature review the models that were tested and finally used, were interpretable, in order to understand the most important factors influencing the player's market value.

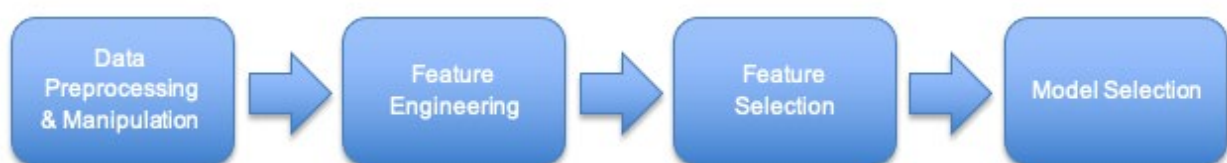


Figure 4.2 – Machine Learning Steps (Own Illustration)

Above Figure 4.2 gives an overview for each Machine Learning Step that was applied for predicting the market value of the players.

The initial step was the Data Preprocessing and Data Manipulation. First of all, the FIFA 20 dataset was loaded as train file and the FIFA 21 dataset as a test file, in both datasets a unique player id for each player was created and it was ensured that only players with a market value were loaded. After that an Exploratory Analysis was conducted to get an overview of the data. Next, the train as well as the test files were splitted into X_train, X_test and y_train, y_test for preventing any data leakage. Then the X_train and X_test were processed further by resetting the index to the player id of the football player and by dropping columns that are not important for the prediction like "player_url", "club_team_id" etc.

The reset of the index was important to precisely map each player with a unique prediction because there were players with the exact same name. Moreover, the following features were being dropped as they were having more than 50% of missing values "club_loaned_from - 94%", "nation_position - 95%", "nation_jersey_number - 95%", "player_traits - 58% missing values". After that a function was defined that fill up the missing values for X_train and X_test by filling the numeric columns with mean and categorical columns with mode.

The second step of the Machine Learning prediction was the Feature Engineering. A new feature "BMI" for X_train and X_test was introduced by dividing the weight feature by the squared height feature. Furthermore, the Date of Birth was separated for each player in Birth Year, Birth Month and Birth Day for better processing power. In addition, the first mentioned position for each player was kept and the other discarded. Then X_train as well as X_test were separated into categorical and numerical variables. After that, the numerical variables of X_train and X_test were scaled with the Robust Scaler, which handles outliers and reduces the discrepancies between the features of data like high bias, when values differ from each other on a broad scale. Furthermore, the categorical variables in both files were encoded using the One Hot Encoder. This encoding was necessary because Machine Learning primarily relies on mathematical equations and typically cannot handle categorical values.

After that a Feature Selection was conducted. The most important features for this prediction task were evaluated and tested with the feature selection techniques of ANOVA, Backward Elimination, feature importance of Extra Tree Classifier and ADA Boost Classifier. The most accurate feature subset was obtained through feature importance by applying the Extra Tree Classifier algorithm. Figure 4.3 below shows the Code Snippet with the feature importance of the Extra Tree Classifier.

Feature Selection Extra Tree Classifier

```
#feature selection technique with extra tree classifier
model = ExtraTreesClassifier()
model.fit(X_train, y_train)
print(model.feature_importances_)

#sorted by feature importance bigger than 0.01
feat_importances = pd.Series(model.feature_importances_, index=X_train.columns)
feat_importances = feat_importances[feat_importances>0.01].sort_values(ascending=False)
feat_importances
```

overall	0.030905
release_clause_eur	0.025967
potential	0.025573
age	0.019113
birthyear	0.017817
movement_reactions	0.016498
skill_ball_control	0.016348
wage_eur	0.015562
attacking_short_passing	0.015466
skill_dribbling	0.015310
mentality_composure	0.015302
mentality_interceptions	0.015242
defending_marking_awareness	0.015220
defending_standing_tackle	0.015177
mentality_positioning	0.015019
power_shot_power	0.015006
mentality_vision	0.014933
defending_sliding_tackle	0.014824
birthday	0.014766

Figure 4.3 – Feature Selection technique with the best obtained results (Own Illustration)

The last step of the market value prediction was the model selection. For a transparent comparison of the different models and the final selection of the best performing model some metrics were defined. Table 4.1 represents the five different metrics that were applied for evaluating the Machine Learning models.

Table 4.1 – Metrics Overview for Model Evaluation (Own Illustration)

Mean Absolute Error	Mean Absolute percentage Error	Root Mean Squared Error
$MAE = \frac{1}{m} \sum_{i=1}^m X_i - Y_i $	$MAPE = \frac{1}{m} \sum_{i=1}^m \left \frac{Y_i - X_i}{Y_i} \right $	$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (X_i - Y_i)^2}$
Mean Squared Error	R-squared	
$MSE = \frac{1}{m} \sum_{i=1}^m (X_i - Y_i)^2$	$R^2 = 1 - \frac{\sum_{i=1}^m (X_i - Y_i)^2}{\sum_{i=1}^m (\bar{Y} - Y_i)^2}$	

After defining the evaluation metrics, different algorithms were tested and compared based on the metrics. Below Table 4.2 is giving an overview of the algorithms that were applied and evaluated with the corresponding metrics.

Table 4.2 – Metrics Overview for the applied Models (Own Illustration)

Algorithm	Mean Absolute Error	MAPE	Mean Squared Error	Root Mean Squared Error	R-squared
Random Forest Classifier	897569.1452397498	0.35%	9834057677660.77	3135930.11	0.835
Decision Tree Classifier	887804.6185919709	0.34%	9832897920617.951	3135745.19	0.836
Logistic Regression	1108072.8069706527	0.39%	14290703992088.523	3780304.74	0.761
Lasso Regression	968644.2992415084	0.69%	9804765534910.049	3131256.22	0.836
Random Forrest Regressor	1162025.4808692685	0.76%	10710762856765.223	3272730.18	0.821
KNN Regressor	978538.4518192024	0.37%	12535988042529.465	3540619.72	0.790
Support Vector Machine	1135894.737071714	0.47%	18244508434034.492	4271359.08	0.695
XG Boost	849425.9008053195	0.33%	8615923187823.2	2935289.28	0.856

The best results were obtained with XG Boost Regressor by applying with the best feature subset of Extra Tree Feature Selection of the following features:

'overall', 'release_clause_eur', 'potential', 'age', 'birthyear', 'skill_ball_control', 'defending', 'movement_reactions', 'dribbling', 'attacking_short_passing', 'defending_marking_awareness', 'skill_dribbling', 'wage_eur', 'mentality_composure', 'defending_standing_tackle', 'mentality_interceptions', 'shooting', 'defending_sliding_tackle', 'passing', 'attacking_heading_accuracy', 'attacking_finishing', 'mentality_positioning', 'mentality_vision', 'mentality_aggression', 'birthday', 'power_shot_power', 'power_long_shots', 'skill_long_passing', 'pace', 'attacking_crossing', 'power_stamina', 'physic', 'skill_curve', 'attacking_volleys', 'movement_sprint_speed', 'movement_acceleration', 'power_jumping', 'birthmonth', 'skill_fk_accuracy',

'mentality_penalties', 'movement_agility', 'BMI', 'power_strength', 'club_jersey_number', 'movement_balance', 'height_cm', 'weight_kg', 'club_joined_day', 'club_contract_valid_until', 'club_joined_year', 'club_joined_month' and 'weak_foot'.

Figure 4.4 is illustrating the corresponding code of the applied XG Boost Regressor algorithm in Jupyter Notebook.

```
XGBOOST

rfModel= xgb.XGBRegressor()
# Fit the boost on the training data
rfModel.fit(X_feat_train, y_train)

# Make predictions on the test data
y_pred = rfModel.predict(X_feat_test)

# Calculate the difference between actual and predicted values
difference = y_test - y_pred

# Create a dataframe to compare predicted and actual values
comparison_df = pd.DataFrame({'Predicted Value': y_pred,
                              'Actual Value': y_test,
                              'Difference': difference})

# Set the index of the dataframe as the matching index of X_feat_train
comparison_df.index = X_feat_test.index

# Calculate evaluation metrics
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)
r2 = r2_score(y_test, y_pred)
mape = mean_absolute_percentage_error(y_test, y_pred)

print(comparison_df)
print("Mean Absolute Error:", mae)
print(f"MAPE: {mape:.2f}%")
print("Mean Squared Error:", mse)
print("Root Mean Squared Error:", rmse)
print("R-squared:", r2)
comparison_df.sort_values(by = "Difference")
```

Figure 4.4 – Code Snippet XG Boost (Own Illustration)

Overall XG Boost achieved the most accurate results and outperformed the other algorithms with the following metrics (Figure 4.5):

- Mean Absolute Error: 849425.9008053195
- MAPE: 0.33%
- Mean Squared Error: 8615923187823.2
- Root Mean Squared Error: 2935289.285202261
- R-squared: 0.8563123106395081

	Predicted Value	Actual Value	Difference
player_id			
1	7.577634e+07	103500000	2.772366e+07
2	4.300464e+07	63000000	1.999536e+07
3	7.681856e+07	111000000	3.418144e+07
4	1.016402e+08	132000000	3.035976e+07
5	8.105002e+07	129000000	4.794998e+07
...	...	...	...
18940	4.178431e+04	70000	2.821569e+04
18941	6.177750e+04	70000	8.222500e+03
18942	6.166396e+04	45000	-1.666396e+04
18943	7.786858e+04	130000	5.213142e+04
18944	7.480196e+04	130000	5.519804e+04
[18707 rows x 3 columns]			
Mean Absolute Error: 849425.9008053195			
MAPE: 0.33%			
Mean Squared Error: 8615923187823.2			
Root Mean Squared Error: 2935289.285202261			
R-squared: 0.8563123106395081			

Figure 4.5 – Best achieved metrics with XG Boost (Own Illustration)

As Figure 4.5 above shows, the outcome of this market value feature was the unique player Id column with the predicted market value, the actual market value, and the difference of both market values.

4.2.2.2. Recommendation System Feature

The second feature that was implemented to the football coaching dashboard is a Recommendation System for similar players. Based on the findings in the literature review, a Content-Based Filtering Technique was integrated, as it recommends football players, based on similarity measurements rather than only relying on past user ratings or preferences about football players.

This feature of a Recommendation System can support football coaches in their scouting activities. It involves the suggestion of similar players, in case if the desired player is for example not available or not fitting in the financial framework of the club.

For realizing the Recommendation System, the same steps of Data Preprocessing, Feature Engineering and Feature Selection as in the Machine Learning feature were applied. Also, important to mention is once more the reset of the index to a unique player ID, as in the data there were players with the exact same name. After these initial steps the following tasks were performed in the Jupyter Notebook with Python:

1. The prepared data was converted into a matrix. Then a Nearest Neighbors search object “knn_search” was initialized with a cosine similarity metric and the brute-force algorithm. This object was used to find the nearest neighbors in the matrix.
2. The code created an empty dictionary called “rec_dict” to store the recommendations for each player.
3. The next step was to iterate through each player in the matrix and find the four nearest neighbors (including the player itself, resulting in three) based on the cosine similarity. Then these neighbors and their similarity scores were stored in the “rec_dict” dictionary.
4. The fourth step was to create several empty lists (“rec_players_1”, “rec_players_2”, “rec_players_3”, “similarity_scores_1”, “similarity_scores_2”, “similarity_scores_3”) in order to store the recommended players and their similarity scores. (Figure 4.6)

KNN Recommendation with Cosine Similarity

```
matrix = csr_matrix(df1.values)
knn_search = NearestNeighbors(metric='cosine', algorithm='brute')
knn_search.fit(matrix)

rec_dict = {} # Create a dictionary to store recommendations for each player

for player_idx, player_row in enumerate(matrix):
    distances, indices = knn_search.kneighbors(player_row.reshape(1, -1), n_neighbors=4) # Get 4 neighbors to include

    player = df1.index[player_idx]
    recommendations = [] # Store recommendations for the current player

    for elem in range(1, len(distances.flatten())):
        rec_player = df1.index[indices.flatten()[elem]]
        similarity_score = 1 - distances.flatten()[elem]
        recommendations.append((rec_player, similarity_score))

    rec_dict[player] = recommendations

# Create lists for new columns
rec_players_1 = [] # First recommended player
rec_players_2 = [] # Second recommended player
rec_players_3 = [] # Third recommended player
similarity_scores_1 = [] # Similarity score for first recommended player
similarity_scores_2 = [] # Similarity score for second recommended player
similarity_scores_3 = [] # Similarity score for third recommended player
```

Figure 4.6 – First part of the Code for Recommendation System (Own Illustration)

- Following that, these lists are filled with the top three recommended players and their similarity scores for each player. The recommendations are sorted by the similarity score in descending order.
- If there are not enough recommendations for a player (e.g. less than three), the corresponding list elements are set to “None”.
- Finally, new columns were added to the original DataFrame, including “RecPlayer_1”, “SimilarityScore_1”, “RecPlayer_2”, “SimilarityScore_2”, “RecPlayer_3”, and “SimilarityScore_3”, which contained the recommended players index and their similarity scores. (Figure 4.7)

```

# Populate new columns with recommended players and similarity scores
for player in df1.index:
    recommendations = rec_dict.get(player, [])
    recommendations.sort(key=lambda x: x[1], reverse=True) # Sort recommendations by similarity score

    if recommendations:
        rec_players_1.append(recommendations[0][0]) # Get the first recommended player
        similarity_scores_1.append(recommendations[0][1]) # Get the similarity score for first recommended player

        if len(recommendations) > 1:
            rec_players_2.append(recommendations[1][0]) # Get the second recommended player
            similarity_scores_2.append(recommendations[1][1]) # Get the similarity score for second recommended player

            if len(recommendations) > 2:
                rec_players_3.append(recommendations[2][0]) # Get the third recommended player
                similarity_scores_3.append(recommendations[2][1]) # Get the similarity score for third recommended player
        else:
            rec_players_1.append(None)
            similarity_scores_1.append(None)
            rec_players_2.append(None)
            similarity_scores_2.append(None)
            rec_players_3.append(None)
            similarity_scores_3.append(None)

# Add new columns to df1
df1['RecPlayer_1'] = rec_players_1
df1['SimilarityScore_1'] = similarity_scores_1
df1['RecPlayer_2'] = rec_players_2
df1['SimilarityScore_2'] = similarity_scores_2
df1['RecPlayer_3'] = rec_players_3
df1['SimilarityScore_3'] = similarity_scores_3

# Display the updated df1 DataFrame
df1

```

Figure 4.7 – Second Part of Code for the Recommendation System (Own Illustration)

To conclude, the Recommendation System essentially performed nearest neighbor-based recommendations for each player id in the DataFrame based on cosine similarity and stored the results for the most similar recommended player with their similarity scores. Below Figure 4.8 represents for a better understanding the output of this implemented feature.

```
df1 = df1.iloc[:, -6:]
df1
```

	RecPlayer_1	SimilarityScore_1	RecPlayer_2	SimilarityScore_2	RecPlayer_3	SimilarityScore_3
player_id						
1	114	0.991042	15	0.990425	55	0.988981
2	39	0.993560	13	0.992793	120	0.992643
3	9	0.998404	10	0.998026	23	0.997787
4	9	0.998644	10	0.998384	24	0.997389
5	24	0.995323	46	0.995285	28	0.994594
...	...	...	...	...	...	...
18940	18919	0.942718	18873	0.942220	18938	0.928316
18941	18876	0.909773	18920	0.909392	18917	0.872174
18942	18939	0.858364	18918	0.841842	18806	0.835585
18943	18587	0.921187	18812	0.918813	18611	0.905104
18944	18537	0.879490	18674	0.878474	18805	0.871805

18944 rows x 6 columns

Figure 4.8 – Output of the Recommendation System feature (Own Illustration)

4.2.3. Power BI Model & Design

After successfully setting up the two features a visualization tool needed to be defined first. The choice for the tool was Microsoft Power BI, which has been identified as the leading Data Visualization tool in the literature review.

Following the selection of Power BI, the next step was to establish an architecture in order to create a conceptual model. In accordance with the results obtained in the literature review, the Star schema of Kimball's dimensional modeling was implemented as it represents the most efficient Data warehouse architecture. This type of schema has several benefits like low complexity and improving the performance by reducing the time of queries.

The Star schema was established through the following tables that were loaded into the Power BI dashboard and defining Facts and Dimensions:

1. **Data Table:** The first step was to load the Prepared FIFA Data CSV file to Power BI to create a Raw Data Table. In there some minor steps were applied like removing unnecessary columns and changing some data types.

2. **Prediction Table:** The second step that was applied was the feature integration of the market value prediction. The prediction CSV file outcome was loaded into Power BI and unnecessary columns were removed again and the data type was changed to correctly display the market value in Euros, which then resulted in the prediction table.
3. **Attribute Table:** After that, an Attribute table was created in order to visualize a radar chart for the Player Performance Evaluation View based on player attributes. The basis was the Prepared FIFA Data CSV file and then irrelevant columns were removed, and the other required columns were renamed to more logical attribute names (Strength, Sprint Speed, Jumping, Rating, Potential and Dribbling Skills). Following that, the attributes were unpivoted into one single column, resulting in the six attributes for each player id with the corresponding score. After that the player id data type was changed to Integer 64.
4. **Dim Club Table:** Then the Dimension Club Table was created by also loading the Prepared FIFA Data CSV file to Power BI and removing unnecessary columns that are not related to the clubs or are not providing information about the club. Then the club IDs duplicates were removed to guarantee that there are just unique clubs. Lastly, the club team ID was sorted starting from 1 and the remaining blank were removed. Below Figure 4.9 gives an overview of the output and the described steps that were applied in Power BI for the first Dimension Table Club.

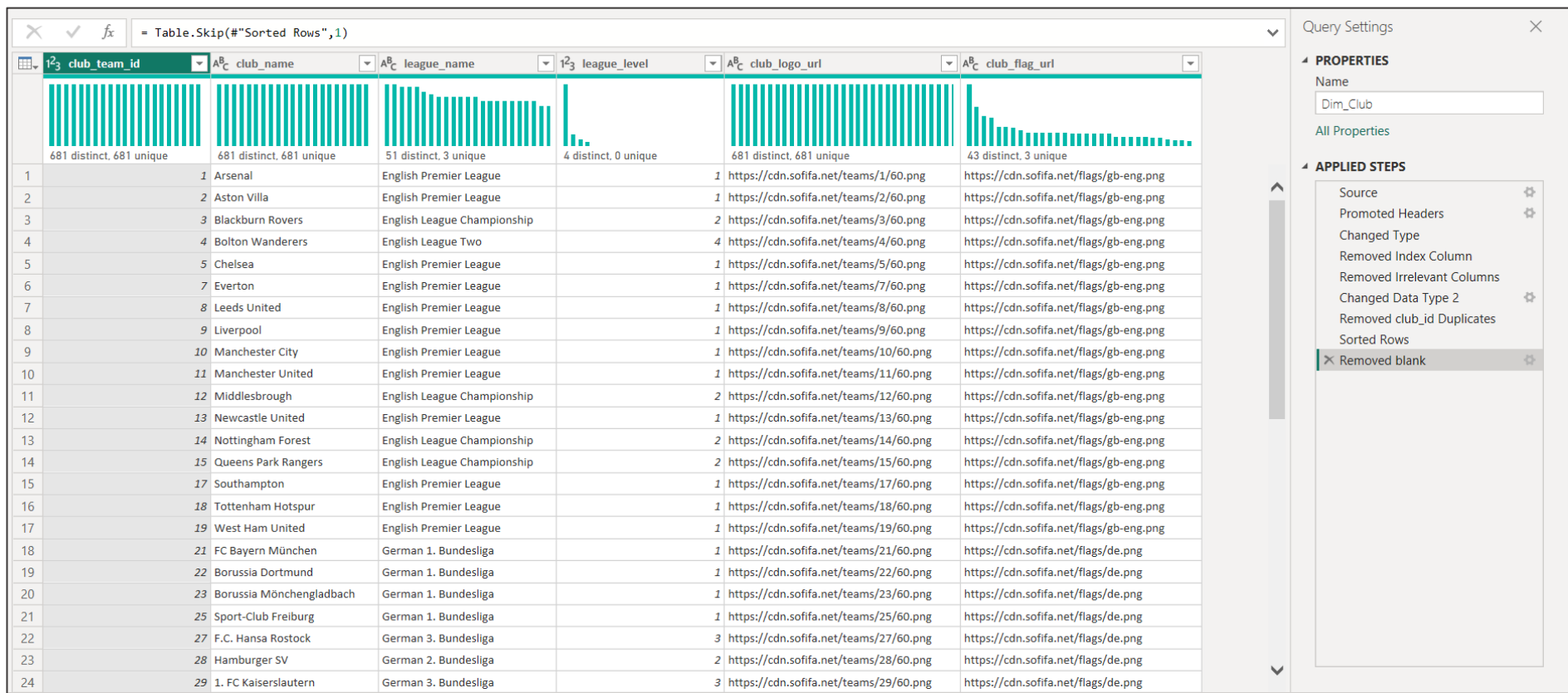


Figure 4.9 – Dimension Table Club (Own Illustration)

5. **Dim Player Table:** The next step was to create the second Dimension Table about the Players. Again, the Prepared FIFA Data CSV file was used for loading it to Power BI and then all relevant categorical columns about players were kept and the other columns were removed. That resulted in a table with all columns related to player characteristics. For a better understanding Figure 4.10 is showing the applied steps and the outcome of the Dim Players Table.

Table.TransformColumnTypes(#"Removed Columns2",{{"player_id", Int64.Type}})						
1000 distinct, 1000 unique	1000 distinct, 1000 unique	991 distinct, 982 unique	1000 distinct, 1000 unique	15 distinct, 0 unique	65 distinct, 19 unique	
1	https://sofifa.com/player/158023/lionel-messi/210002	L. Messi	Lionel Andrés Messi Cuccittini	RW		
2	https://sofifa.com/player/20801/c-ronaldo-dos-santos-aveiro/210002	Cristiano Ronaldo	Cristiano Ronaldo dos Santos Aveiro	ST		
3	https://sofifa.com/player/188545/robert-lewandowski/210002	R. Lewandowski	Robert Lewandowski	ST		
4	https://sofifa.com/player/190871/neymar-da-silva-santos-jr/210002	Neymar Jr	Neymar da Silva Santos Júnior	LW		
5	https://sofifa.com/player/192985/kevin-de-bruyne/210002	K. De Bruyne	Kevin De Bruyne	CAM		
6	https://sofifa.com/player/200389/jan-oblak/210002	J. Oblak	Jan Oblak	GK		
7	https://sofifa.com/player/192448/marc-andre-ter-stegen/210002	M. ter Stegen	Marc-André ter Stegen	GK		
8	https://sofifa.com/player/203376/virgil-van-dijk/210002	V. van Dijk	Virgil van Dijk	CB		
9	https://sofifa.com/player/208722/sadio-mane/210002	S. Mané	Sadio Mané	LW		
10	https://sofifa.com/player/209331/mohamed-salah/210002	M. Salah	Mohamed Salah Ghaly	RW		
11	https://sofifa.com/player/212831/alisson-ramses-becker/210002	Alisson	Alisson Ramsés Becker	GK		
12	https://sofifa.com/player/231747/kylian-mbappe/210002	K. Mbappé	Kylian Mbappé Lottin	ST		
13	https://sofifa.com/player/153079/sergio-aguero/210002	S. Agüero	Sergio Leonel Agüero del Castillo	ST		
14	https://sofifa.com/player/155862/sergio-ramos-garcia/210002	Sergio Ramos	Sergio Ramos García	CB		
15	https://sofifa.com/player/165153/karim-benzema/210002	K. Benzema	Karim Benzema	CF		
16	https://sofifa.com/player/167495/manuel-neuer/210002	M. Neuer	Manuel Peter Neuer	GK		
17	https://sofifa.com/player/192119/thibaut-courtois/210002	T. Courtois	Thibaut Courtois	GK		
18	https://sofifa.com/player/200145/carlos-henrique-venancio-casimiro/...	Casemiro	Carlos Henrique Venancio Casimiro	CDM		
19	https://sofifa.com/player/162835/samir-handanovic/210002	S. Handanović	Samir Handanović	GK		
20	https://sofifa.com/player/182521/toni-kroos/210002	T. Kroos	Toni Kroos	CM		
21	https://sofifa.com/player/183277/eden-hazard/210002	E. Hazard	Eden Hazard	LW		
22	https://sofifa.com/player/201024/kalidou-koulibaly/210002	K. Koulibaly	Kalidou Koulibaly	CB		
23	https://sofifa.com/player/202126/harry-kane/210002	H. Kane	Harry Kane	ST		

Query Settings

PROPERTIES

Name
Dim_Player

All Properties

APPLIED STEPS

- Source
- Promoted Headers
- Changed Type
- Removed Columns
- Reordered Columns
- Removed Columns1
- Changed Type with Locale
- Changed Type1
- Removed Columns2
- Changed Type2

Figure 4.10 – Dimension Table Player (Own Illustration)

6. **Dim Calendar Table:** The sixth step was to create the Dimension Table Calendar by using the DAX formula “CALENDARAUTO”. This table was important in the model for possible time measurements for visualizations.
7. **Fact Fifa Table:** The last step was to set up the Fact Table “Fifa”. The output of the Recommendation System feature was the basis for the Fact Table and was loaded as a CSV file into Power BI. Then the recommendation player and the score columns were merged together and unpivoted in the next step. After that, again unnecessary columns were removed and the unpivoted columns were splitted, that resulted in a table that had for each unique player ID three recommended players and similarity scores. Following that, the raw data table was merged on the player id and some data types were changed. The next step was to merge the Fact table with the Attribute table and the Prediction table on player id. Moreover, the Recommended Player Id of the Fact table was merged with the player id of the Dimension Player table in order to provide the name of the recommended players. After that, the player id of the Fact table was sorted, and a unique primary key was added through an index column. Figure 4.11 is illustrating the final output of the Fact Table and also shows the applied steps.

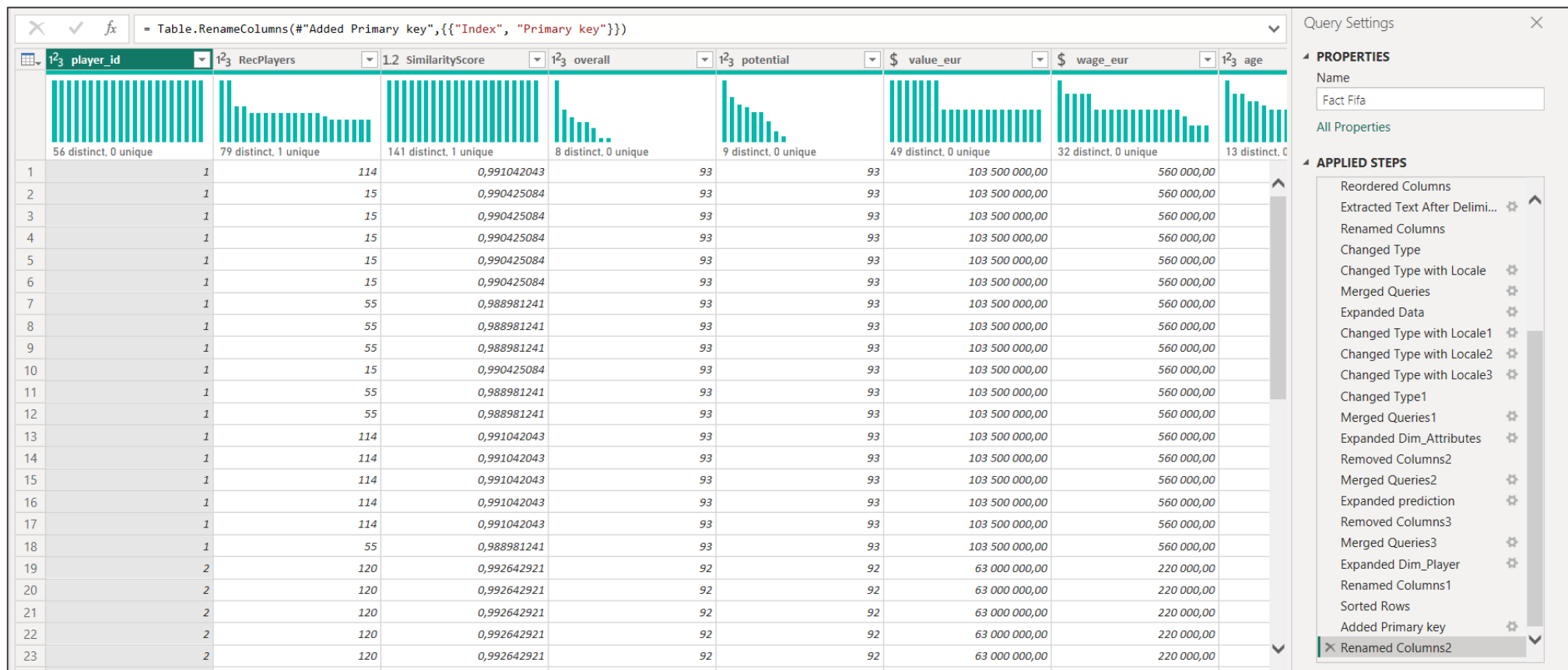


Figure 4.11 – Fact Table Fifa (Own Illustration)

By applying all of the steps described above a star schema design was implemented that consisted of the three described Dimensions: Calendar_Dimension; Player_Dimension; Club_Dimension and as mentioned one resulting Fact Table Fact_Fifa. The relationships were established as a many to one originating from the Fact Fifa Table. (Figure 4.12)

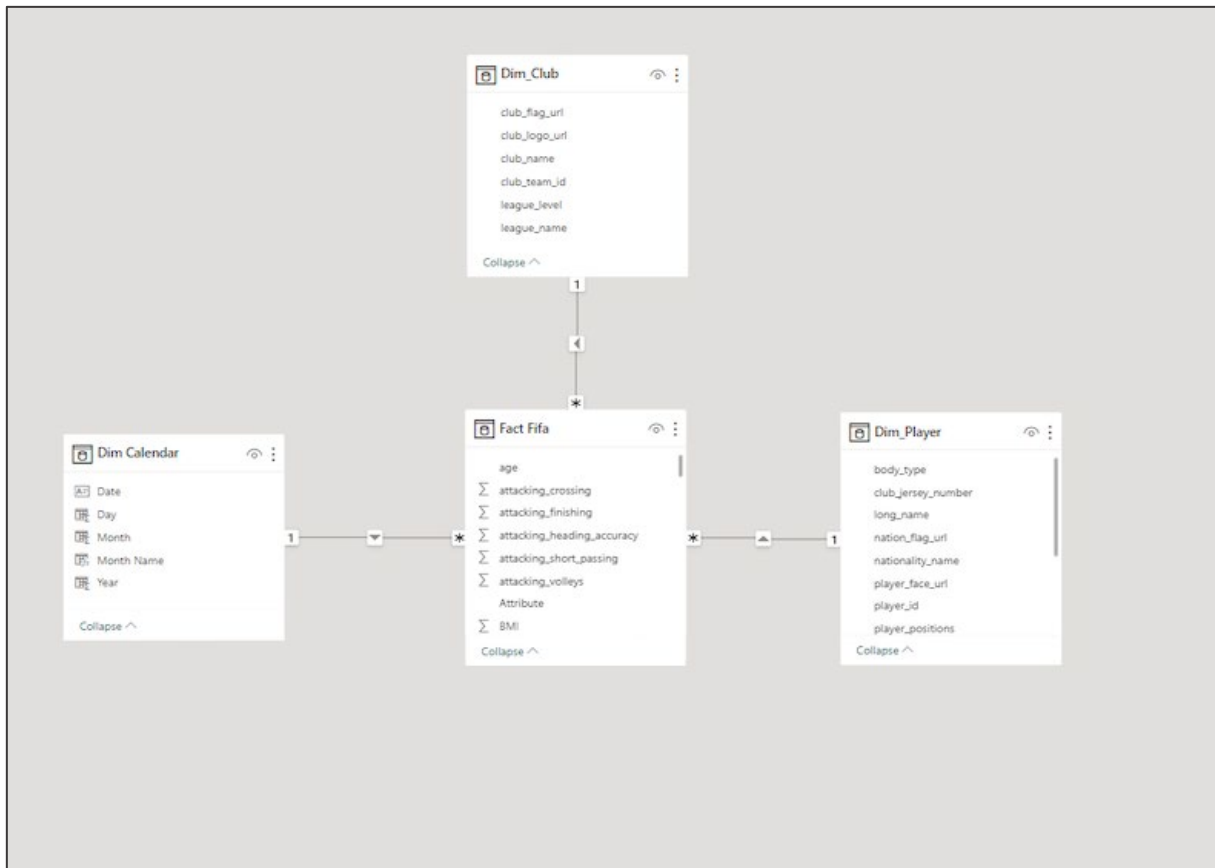


Figure 4.12 – Star Schema for Power BI Model (Own Illustration)

After setting up the conceptual model architecture, the next step was to define the conditions for the design of the dashboard.

To ensure a great user experience for the football coaches, the Power BI dashboard aimed to represent a systematic data-driven view that can improve and support the complex decision-making process of football coaches.

The requirement for the design of the dashboard was to facilitate intuitive data interpretation for the football coaches. Therefore, the dashboard layout was carefully structured into different aspects so that football coaches could access the key metrics without any effort. Moreover, the selection of the visualizations was based on the effectiveness for expressing the most crucial insights within the given context. These visualizations included for example radar charts for an easy comparison of different player characteristics, or a tachometer for

getting an easy overview of different metrics if players are under or overperforming. Another important design condition was the usage of different color to enhance the readability and quick interpretation of information.

Additionally, the dashboard interface had to empower football coaches with the ability to filter different data and make player comparisons, enabling them to make data-driven decisions effectively.

Moreover, to improve the navigation in the dashboard, a Landing Page was established, and the following three areas were identified for distinct views in the dashboard:

- **Comprehensive league assessment view:** This view provided an overview of league related data and measures.
- **Overall player performance evaluation view:** This view provided coaches a detailed assessment of their individual players.
- **Insightful player scouting platform view:** The last view offered in-depth insights into player scouting, with the two integrated features of recommended similar players and market value prediction.

5. CONCEPTUAL MODEL USE CASE – SL BENFICA LISBON

In this chapter a detailed explanation of the realized Power BI dashboard will be presented that was based on the specific perspective and use case of the SL Benfica Lisbon head coach.

This includes an in-depth analysis of each of the three identified dashboard areas, offering a more insightful understanding of its functionality and relevance within the context of coaching at SL Benfica.

First the initial landing page of the dashboard will be displayed, then the Portuguese league assessment view will be described, followed by the Player performance view of the Benfica Players, and last the integrated Player scouting platform view will be examined.

5.1. LANDING PAGE

Starting with the initial Landing Page, Figure 5.1 shows the design of the dashboard from the specific perspective of the Benfica coach. The coach can navigate from that first starting point to the three main views while this page also serves as the structural and logical guide for the overall dashboard.



Figure 5.1 – Landing Page (Own Illustration)

5.2. PORTUGUESE LEAGUE ASSESSMENT VIEW

After navigating from the landing page to the first view of the coaching dashboard the Portuguese league perspective is being displayed. (Figure 5.2) This comprehensive view provides the Benfica coach with valuable information about the national league, serving as a fundamental model for assessing and benchmarking Benfica Lisbon against the overall league performance areas.

Moreover, within this view, the Benfica coach gains deep insights into the Portuguese league, extending beyond the average age and rating of its players, it also displays financial perspectives, such as the average salary per week, average market value and the average release clause of the league's players. These financial indicators empower the coach to evaluate whether his team is excelling or lagging in these areas on a national view.

Furthermore, for a better comparison, the view also shows the average market value and average rating per club and also presents a breakdown of player positions in the Portuguese league, helping the coach to evaluate position distribution and identifying potential gaps in the squad.

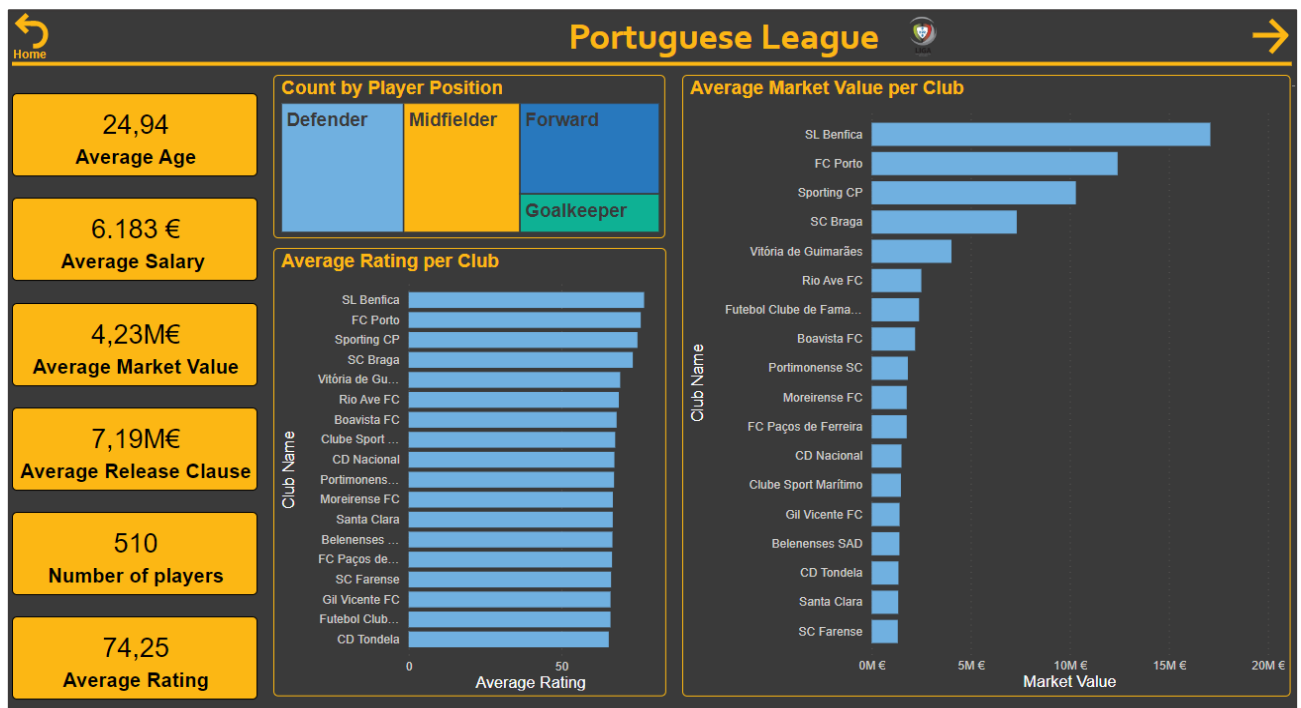


Figure 5.2 – Portuguese League Assessment View (Own Illustration)

All of those insights help the coach to get a detailed overview of his team's performance and resources in comparison to their national rivals, improving the strategic decision-making.

5.3. SL BENFICA PLAYER PERFORMANCE EVALUATION VIEW

The second view of the developed coaching dashboard focuses on displaying the characteristics and key performance indicators of the Benfica Lisbon players (Figure 5.3).

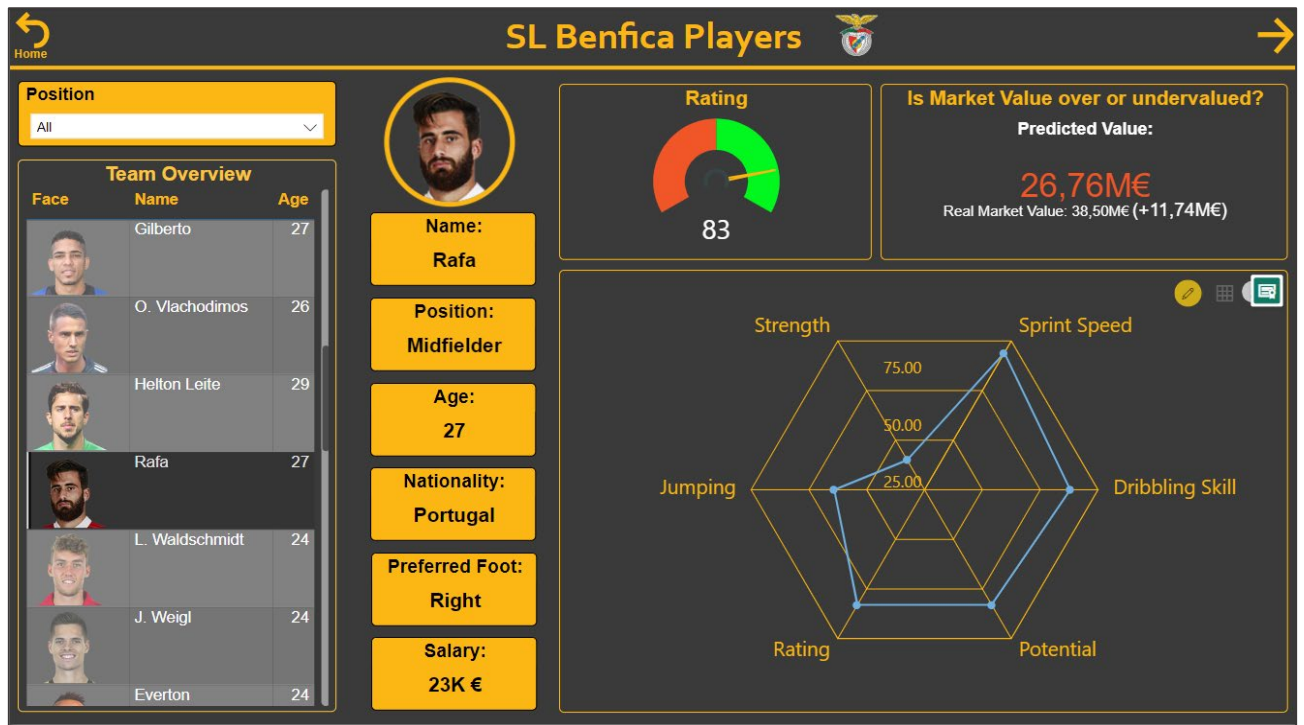


Figure 5.3 – Player Performance Evaluation View (Own Illustration)

In that view the Benfica coach can first select the desired positions in the filter that he wants to analyze in detail. Then a team overview table with Benfica players with that position will appear and the coach can select a player that he wants to assess in depth.

Additionally, the key characteristics of the selected player such as Name, Position, Age, Nationality, Preferred Foot and Salary will be presented to the coach. Furthermore, the Benfica Coach can evaluate the proposed rating for the selected player and analyze a Radar Chart that consists of the key factors Jumping, Strength, Sprint Speed, Dribbling Skill, Potential and Rating. This chart is a great illustration for the coach to assess how the different characteristics of a player are distributed and identify distinct player types. The scale for the values is reaching from 0-100, where 100 is the maximum value to achieve.

In order to support the Benfica coach for evaluating the market value of his players the Machine Learning feature for market value prediction is integrated in this view. This feature helps and assists the coach in determining whether a player's market value is over or undervalued, complementing his personal evaluation. This feature becomes particularly an

advantage when another club makes an offer for a Benfica player or if the coach considers selling the player. If the predicted market value is lower than the real market value it indicates that the player is likely overvalued and the predicted value is displayed in red color and suggests that in case of an offer the player can be sold. In contrast, if the predicted value exceeds the real market value, it is presented in green, suggesting that the player may be undervalued.

5.4. SCOUTING PLATFORM VIEW

The final component of the coaching dashboard is the Scouting platform. It provides the coach with the possibility to search for specific criteria of players among a vast pool of more than 18.000 different players across the world (Figure 5.4).

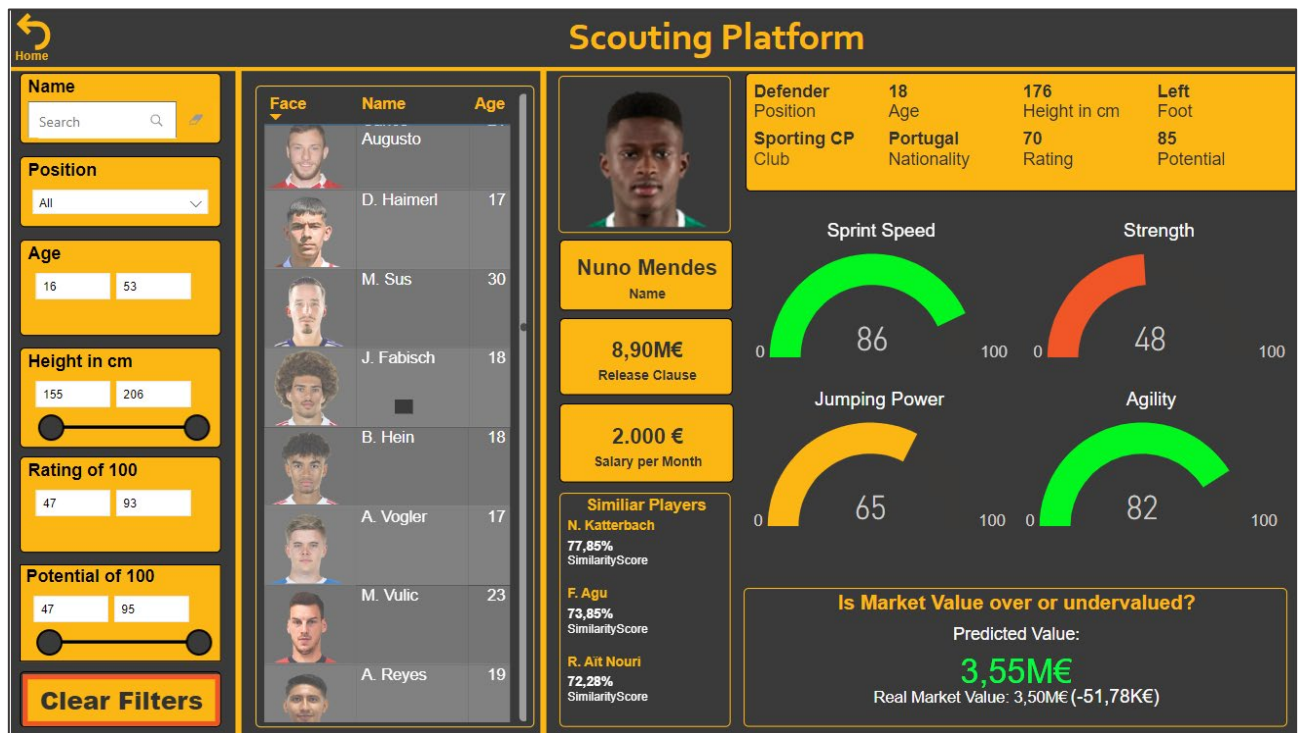


Figure 5.4 – Scouting Platform View (Own Illustration)

The view consists of different filters on the left side that the coach can specify and search for players such as a specific name, the position, the age, the height in cm and the player's rating and potential. Based on the filter settings a table is being displayed that shows all the players matching the filter selection. Then the coach can select one player and the key characteristics for this player will be displayed such as his Name, and financial variables like release clause and salary per month. Those variables are particularly important for the financial frame of the club in case the coach considers buying one player.

Moreover, key characteristics such as Position, Age, Height, Preferred Foot, Club, Nationality, Potential and Ranking are shown in the scouting platform to get a better overview of the selected player.

In addition, the player's physical variables of Sprint Speed, Strength, Jumping Power, and Agility are visually represented by a Gauge Chart. This helps to identify if a player physically aligns with the coach's preferred playing system and style. Again the scale is starting from value 0 to 100, as the maximum possible value that can be reached.

On top of that the recommendation feature was implemented to support the coach if the desired player is not available for a transfer or his financial demands are not aligning with the club's budget. This feature enables the coach to discover similar players have not been initially considered and compare them to each other.

The Recommendation System is based on cosine similarity and recommends similar players who share similarities in e.g. physical, technical or tactical characteristics. For each player its top three most similar players are being displayed with a score for the coach in that view.

Ultimately, the market value feature is also integrated in the Scouting platform view as it is valuable for assessing the market value of desired players, especially for less known players. The green color of the predicted market value indicates that the predicted market value was higher than the defined market value, which suggests that the player is, according to the prediction, undervalued and therefore an opportunity can arise for the football coach to buy the player. The opposite scenario is represented by the red color.

In summary, this view of a scouting platform enables the coach to search in detail for players using filter criteria, obtaining an in-depth analysis covering physical, financial and player characteristics. Additionally, the view is offering valuable market value insights through the integrated feature, helping to assess over and undervalued players while also providing recommendations for discovering similar players.

6. MODEL EVALUATION

This chapter will cover the evaluation of the developed coaching dashboard. Considering the Design science research (DSR) methodology, all of the steps were being applied and the overall objective of creating a Power BI dashboard tailored for the needs of football coaches was achieved successfully.

In detail the dashboard covered the three different views of a comprehensive Portuguese league assessment, an overall player performance evaluation and a player scouting platform. In order to assess the dashboard more precisely the views of the dashboard were created with the perspective of a Benfica Lisbon coach. These views are providing a Benfica coach an extensive overview for the Portuguese league as well as an analysis and comparison of his own players. Moreover, the Scouting platform enables the coach to search more than 18.000 different players based on filters that he can set.

In addition, the integrated market value feature provided a great tool for supporting the assessment of over or undervalued players, reaching a MAPE (Mean Absolute Percentage Error) of 0.33%, which indicates a high-level accuracy between the predicted market values and the actual market values of the players. Furthermore, the achieved R-squared value of 85.6% demonstrates a precise alignment between the predicted and actual market values of the players, reinforcing its effectiveness in player valuation.

Moreover, the feature that was implemented into the dashboard of recommending similar players also achieved precise results. Multiple qualitative evaluation assessments were conducted through testing a random test sample of players. For this type of evaluation, the Recommendation System consistently provided accurate and relevant recommendations of similar players.

Through implementing these features, the gap of missing advanced analytical tools that was prior identified in the literature review was closed.

In summary the successful realization of the desired dashboard was achieved through the integration of the following key components:

- ✓ The conduction of an extensive and detailed literature review that build based on the key findings and assumptions the foundation for the dashboard model.
- ✓ The selection of an extensive, high-quality and realistic FIFA dataset.
- ✓ The Integration of the two features to provide a tool with more advanced analytical possibilities for football coaches.
- ✓ The precise and accurate outcome that was achieved through the integrated features.
- ✓ The establishment of the Star schema as the Power BI architecture for faster query performance.
- ✓ The overall design requirements for creating an intuitive and interactive dashboard.

- √ The selection of the right visualizations for displaying different metrics for the football coaches.
- √ The overall structure in the dashboard for an effortless navigation between the three identified views for the coaches.

7. CONCLUSIONS

The last chapter will summarize the master thesis project and give an overview about the limitations of this master thesis and outline potential areas for future research and development.

7.1. SYNTHESIS OF THE DEVELOPED WORK

The developed Power BI dashboard for optimizing the decision-making of football coaches was following the defined design science research methodology. An initial literature review was the fundamental for gaining in-depth knowledge regarding the topics of Sports Analytics and Machine Learning in football, Recommendation Systems and Business Intelligence. Based on the main findings of this conducted literature review, assumptions were derived in order to build the conceptual model of the Power BI coaching dashboard. The final dashboard consisted of three different separate views in the perspective of the Benfica Lisbon coach including a comprehensive league assessment view, a detailed player performance view, and an insightful scouting platform view.

In addition to that two advanced features were integrated, namely a market value prediction feature and a Recommendation System for similar players, those features extended the dashboard into a versatile gadget for the Benfica Coach. In accordance with the design science research methodology the objective of creating an artifact of a visual coaching tool that would provide a data driven solution for optimizing the decision-making processes of football coaches was hereby, fully achieved.

7.2. LIMITATIONS

One of the limitations that was already identified in literature review is that the overall potential of data for supporting managerial decisions in football is still often ignored. Additionally, the lack of acceptance of data-driven approaches for estimating the market value in football raises the question regarding the acceptance of this developed coaching dashboard in the football community.

Moreover, while Power BI is the leading tool for Data Visualization, it is one of the limitations in the context of football coaching due to the disregard of such technologies within clubs. As a consequence, coaches and staff in football clubs are not familiar with the tool and therefore, require a comprehensive user training to bridge this knowledge gap.

Addressing this limitation is a crucial step towards ensuring that the dashboard becomes a valuable tool and can be effectively used in the area of football coaching and data-driven decision-making.

7.3. FUTURE WORK

The main focus for future work must be to enhance and speed up the integration and widespread acceptance of data analytic approaches within the football community.

In addition, another area of immense potential is the development of dashboards that provide real-time data updates. Such dashboards would allow coaches to make immediate decisions during matches. Therefore, the focus can be on setting up a future framework for integrating live data streams into the dashboard.

Moreover, the developed dashboard should be evaluated and validated by football experts, so they can express critical feedback and suggest future ideas for developing football coaching dashboards. By supporting this collaborative engagement between data analysts and football experts, the objective towards a data-driven future in football can be enhanced further.

BIBLIOGRAPHICAL REFERENCES

- Al-Asadi, M. A., & Tasdemir, S. (2022). Predict the value of football players using FIFA video game data and machine learning techniques. *IEEE Access*, 10, 22631-22645
- Alamar, B. & Mehrotra, V. (2011). Beyond 'Moneyball': Rapidly evolving world of sports analytics, Part I. Institute for Operations Research and the Management Sciences (INFORMS). <https://doi.org/10.1287/lytx.2011.05.05>
- Balaji, N., Pai, B. K., Bhat, B., & Praveen, B. (2021). Data visualization in Splunk and Tableau: a case study demonstration. In *Journal of Physics: Conference Series* (Vol. 1767, No. 1, p. 012008). IOP Publishing
- Becker, L. T., & Gould, E. M. (2019). Microsoft Power BI: Extending Excel to manipulate, analyze, and visualize diverse data. *Serials Review*, 45(3), 184-188
- Benito-Santos, A., Therón, R., Losada, A. G., Sampaio, J., & Lago-Peñas, C. (2018). Data-Driven Visual Performance Analysis in Soccer: an exploratory Prototype. *Frontiers in Psychology*, 9. <https://doi.org/10.3389/fpsyg.2018.02416>
- Breslin, M. (2004). Data warehousing battle of the giants. *Business intelligence journal*, 7, 6-20
- Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and OLAP technology. *ACM Sigmod record*, 26(1), 65-74
- Chmait, N., & Westerbeek, H. (2021). Artificial intelligence and Machine learning in sport Research: An Introduction for Non-data scientists. *Frontiers in Sports and Active Living*, 3. <https://doi.org/10.3389/fspor.2021.682287>
- Do, M. P. T., Nguyen, D. V., & Nguyen, L. (2010, August). Model-based approach for collaborative filtering. In *6th International conference on information technology for education* (pp. 217-228)
- FIFA (2022), Global Transfer Report 2022
- Gartner (2023), Magic Quadrant for Analytics and Business Intelligence Platforms
- Gudmundsson, J., & Wolle, T. (2012, November). Football analysis using spatio-temporal tools. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems* (pp. 566-569)
- Golfarelli, M., Rizzi, S., & Cella, I. (2004, November). Beyond data warehousing: what's next in business intelligence?. In *Proceedings of the 7th ACM international workshop on Data warehousing and OLAP* (pp. 1-6)

- Halvorsen, P., Sægrov, S., Mortensen, A., Kristensen, D. K., Eichhorn, A., Stenhaug, M., Dahl S., Stensland H., Gaddam V., Griwodz C. & Johansen, D. (2013, February). Bagadus: an integrated system for arena sports analytics: a soccer case study. In Proceedings of the 4th ACM Multimedia Systems Conference (pp. 48-59)
- Hammes F., Hagg A., Asteroth A. & Link D. (2022). Artificial Intelligence in Elite Sports—A Narrative Review of Success Stories and Challenges. *Front. Sports Act. Living* 4:861466. doi: 10.3389/fspor.2022.861466
- Haj-Bolouri, A. (2015). The notion of users in design science research. In 38th Information Systems Research Seminar in Scandinavia (IRIS 38), Oulu, Finland, 9-12 August 2015 (pp. 1-15)
- Herold, M., Goes, F. R., Nopp, S., Bauer, P., Thompson, C., & Meyer, T. (2019). Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science & Coaching*, 14(6), 798–817. <https://doi.org/10.1177/1747954119879350>
- Inan, T. & Cavas, L. (2021) Estimation of Market Values of Football Players through Artificial Neural Network: A Model Study from the Turkish Super League, *Applied Artificial Intelligence*, 35:13, 1022-1042, DOI: 10.1080/08839514.2021.1966884
- Inmon, W. H. (2005), 4th edition, Building the data warehouse. John Wiley & Sons
- Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian informatics journal*, 16(3), 261-273
- Kapadia, K., Abdel-Jaber, H., Thabtah, F., & Hadi, W. (2019). Sport analytics for cricket game results using machine learning: An experimental study. *Applied Computing and Informatics*, 18(3/4), 256–266. <https://doi.org/10.1016/j.aci.2019.11.006>
- Khan, R. A., & Quadri, S. M. (2012). Business intelligence: an integrated approach. *Business Intelligence Journal*, 5(1), 64-70
- Kimball, R., & Ross, M. (2013). The data warehouse toolkit: the definitive guide to dimensional modeling. John Wiley & Sons
- Koren, Y., Rendle, S., & Bell, R. (2021). Advances in collaborative filtering. *Recommender systems handbook*, 91-142
- Ko, H., Lee, S., Park, Y., & Choi, A. (2022). A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1), 141
- Lousa, A., Pedrosa, I., & Bernardino, J. (2019). Avaliação e Análise de Ferramentas Business Intelligence para Visualização de Dados Evaluation and Analysis of Business Intelligence Data Visualization Tools. *Ieeexplore. Ieee. Org*, June, 19-22

- Martyn, T. (2004). Reconsidering multi-dimensional schemas. *ACM Sigmod Record*, 33(1), 83-88
- Moody, D. L., & Kortink, M. A. (2000, June). From enterprise models to dimensional models: a methodology for data warehouse and data mart design. In *DMDW* (p. 5)
- Morgulev, E., Azar, O. H., & Lidor, R. (2018). Sports analytics and the big-data era. *International Journal of Data Science and Analytics*, 5(4), 213–222. <https://doi.org/10.1007/s41060-017-0093-7>
- Müller, O., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. *European Journal of Operational Research*, 263(2), 611–624. <https://doi.org/10.1016/j.ejor.2017.05.005>
- Nassis, G. P., Verhagen, E., Brito, J., Figueiredo, P., & Krstrup, P. (2023). A review of machine learning applications in soccer with an emphasis on injury risk. *Biology of Sport*, 40(1), 233–239. <https://doi.org/10.5114/biol sport.2023.114283>
- Peffer, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science research Methodology for Information Systems research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/mis0742-1222240302>
- Poli R., Ravenel L. and Besson R. (2015). How to build up a team for long term success? The CIES Football Observatory approach, CIES Football Observatory Monthly Report Issue no. 7 - September 2015
- Richau, L., Follert, F., Frenger, M., & Emrich, E. (2021). The impact of investors on transfer fees in the English Premier League: A study of the ownership structures [Special issue]. *Corporate Ownership & Control*, 18(3), 241–256. <http://doi.org/10.22495/cocv18i3siart2>
- Rouhani, S., Asgari, S., & Mirhosseini, S. V. (2012). Review study: Business intelligence concepts and approaches. *American Journal of Scientific Research*, 50(1), 62-75
- Rossi, A., Pappalardo, L., & Cintia, P. (2021). A Narrative review for a machine learning application in Sports: An example based on injury forecasting in Soccer. *Sports*, 10(1), 5. <https://doi.org/10.3390/sports10010005>
- Sharma, L., & Gera, A. (2013). A survey of recommendation system: Research challenges. *International Journal of Engineering Trends and Technology (IJETT)*, 4(5)
- Shin, J., & Gasparyan, R. (2014). A novel way to soccer match prediction. Stanford University: Department of Computer Science.

- Shobirin, K. A., Iskandar, A. P. S., & Swamardika, I. B. A. (2017). Data Warehouse Schemas using Multidimensional Data Model for Retail. *International Journal of Engineering and Emerging Technology*, 2(1), 84-86
- Stanojević, R., & Gyarmati, L. (2016). Towards Data-Driven Football Player Assessment. 2016 IEEE 16th International Conference on Data Mining Workshops. <https://doi.org/10.1109/icdmw.2016.0031>
- Thorat, P. B., Goudar, R. M., & Barve, S. (2015). Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *International Journal of Computer Applications*, 110(4), 31-36
- Tuyls, K., Omidshafiei, S., Müller, P., Wang, Z., Connor, J. T., Hennes, D., Graham, I., Spearman, W., Waskett, T., Steele, D. (2021). Game Plan: What AI can do for Football, and What Football can do for AI. *Journal of Artificial Intelligence Research*, 71, 41–88. <https://doi.org/10.1613/jair.1.12505>
- Vom Brocke, J., Hevner, A. R., & Maedche, A. (2020). Introduction to Design Science Research. In *Progress in IS* (pp. 1–13). https://doi.org/10.1007/978-3-030-46781-4_1
- Watson, H. J., & Wixom, B. H. (2007). The current state of business intelligence. *Computer*, 40(9), 96-99
- Yessad, L., & Labiod, A. (2016). Comparative study of data warehouses modeling approaches: Inmon, Kimball and Data Vault. In *2016 International Conference on System Reliability and Science (ICSRS)* (pp. 95-99). IEEE

