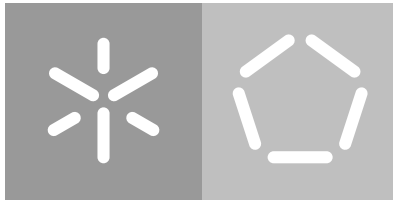


Universidade do Minho
Escola de Engenharia
Departamento de Informática

Inês Lucas Amorim Alves

**Intelligent Data Analysis from the Financial Execution
of Research Projects at University of Minho**

July 2021



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Inês Lucas Amorim Alves

**Intelligent Data Analysis from the Financial Execution
of Research Projects at University of Minho**

Master Dissertation
Master Degree in Informatics Engineering

Dissertation supervised by
Professor Cesar Analide
Professor Filipe Vaz

July 2021

AUTHOR COPYRIGHTS AND TERMS OF USAGE BY THIRD PARTIES

This is an academic work which can be utilized by third parties given that the rules and good practices internationally accepted, regarding author copyrights and related copyrights.

Therefore, the present work can be utilized according to the terms provided in the license bellow.

If the user needs permission to use the work in conditions not foreseen by the licensing indicated, the user should contact the author, through the RepositóriUM of University of Minho.

License provided to the users of this work



Attribution-NonCommercial

CC BY-NC

<https://creativecommons.org/licenses/by-nc/4.0/>

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 Code of Ethical Conduct of the University of Minho.

Insert name

Inês Alves

DEDICATORY

*To my grandmother who didn't see me become an engineer.
To my parents who gave me the chance to become one.*

*À minha avó que não me viu tornar-me uma engenheira.
Aos meus pais que me deram a oportunidade de o fazer.*

ACKNOWLEDGEMENTS

Joël Dicker wrote in his most recent book that "(...) *it is difficult to pay tribute to extraordinary people. Because we don't even know where to start*" and this is the moment when, in fact, I understand and feel these words. The words written here are not enough to thank you for all the love you gave me. **From the bottom of my heart, thank you all so much.**

Firstly, I would like to thank my supervisor, Professor Cesar Analide, for all the support he gave me, which started even before the beginning of this dissertation. Thank you for all the hours dedicated to helping me on the most diverse topics, for having believed in me, and for being a wonderful supervisor.

I would also like to thank my co-supervisor, Professor Filipe Vaz, for the opportunity to work with him and for always being available to help me. And, of course, thanks to Dr. Fátima Costa for having received and treated me so well since day one. It was a huge pleasure to work around such nice people.

To my parents, Anselmo and Delfina, and my sister Ana, for their unconditional support, patience, for making this dream come true, and for never letting me give up. If this cycle comes to an end today, it is thanks to you. A thank you full of love.

To my boyfriend, José Pedro, for all the hours spent looking for answers on google with me, for all his love and understanding, for his friendship and patience in the most difficult times. I'm so grateful to finish this step with you by my side. Thank you so much.

To the University of Minho for the research grant acquired that led to this research.

To Professor Pedro Rangel Henriques for his continuous concern and for being an amazing teacher. Some people specially mark us and Professor Pedro is one of them. Thanks for everything.

And last but not least, thank you to all my friends who accompanied me on this journey.

ABSTRACT

The number of research and development (R&D) projects underway has increased substantially in recent years, which derives from the recognition of the importance of these projects for the future success of the University of Minho and its scientific partners, not only from a financial perspective but also innovation and search for knowledge.

Any Higher Education Institution (HEI) needs a solid management base for many areas that are part of and complete its global organization, such as the area related to R&D projects. A large part of the financial management carried out by the University of Minho is intrinsically linked to project management, whose budgets are often in the thousands of euros.

The data used by the most diverse entities and support centers at the University of Minho are available to those responsible for them in an unintuitive and dispersed way. This dispersion, besides making access to information very difficult, does not sympathize with the organization that a higher education unit needs.

Therefore, getting detailed and reliable information is the key to success, both for researchers, who are directly responsible, and for the regulatory bodies that are implanted in the university. Thus, it was proposed to create a Data Visualization (DV) platform based on project execution data sources from the Financial and Patrimonial Services Unit (USFP) of the University of Minho to provide an organized and coherent data visualization platform, according to the needs of its stakeholders.

With the creation of this platform, through an Intelligent Data Analysis System, using a temporal and detailed observation of the data, it is possible to draw conclusions about the investments made in research projects that have occurred until now and to help in future investment decisions crucial to the healthy functioning of the educational institution. Thus, this analysis seeks not only to improve the financial management of the area in question but also to understand the extent to which the use of Machine Learning techniques can be useful in analyzing data related to the financial execution of R&D projects.

Furthermore, an area that is highly related to research projects, and cannot be ignored, is scientific production. The dissemination of scientific knowledge is an essential part of the research work carried out in any area, so this topic was also studied and introduced within the scope of this dissertation.

Keywords: Project Management, R&D Projects, Data Visualization, Business Intelligence, Intelligent Data Analysis, Machine Learning, Scientific Production

RESUMO

O número de projetos de investigação e desenvolvimento (I&D) em execução tem vindo a aumentar substancialmente nos últimos anos, o que deriva do reconhecimento da importância destes projetos para o sucesso futuro da Universidade do Minho e seus parceiros científicos, não só numa perspetiva financeira, mas também de inovação e procura pelo conhecimento.

Qualquer instituição de ensino superior necessita de uma base sólida de gestão para todos os tipos de áreas que fazem parte e completam a sua organização global, como é o caso da área relacionada com os projetos de I&D. Uma grande parte da gestão financeira realizada pela Universidade do Minho está intrinsecamente ligada à gestão de projetos, cujos orçamentos rondam, muitas vezes, os milhares de euros.

Os dados utilizados pelas mais diversas entidades e centros de apoio da Universidade do Minho encontram-se à disposição dos responsáveis das mesmas de uma forma pouco intuitiva e dispersa. Esta dispersão, para além de dificultar bastante o acesso à informação, não se compadece com a organização que uma unidade de ensino superior necessita.

Neste sentido, a obtenção de informação detalhada e fidedigna é a chave do sucesso, tanto para os investigadores, responsáveis diretos, como para as entidades reguladoras que se encontram implementadas na universidade. Assim, foi proposta a criação de uma plataforma de visualização de dados a partir de fontes de dados de execução de projetos provenientes da Unidade de Serviços Financeiro e Patrimonial (USFP) da Universidade do Minho com o intuito de fornecer uma plataforma de visualização de dados organizada e coerente, conforme as necessidades dos seus *stakeholders*.

Com a criação desta plataforma, através de um sistema de Análise Inteligente de Dados, isto é, fazendo uso de uma observação temporal e detalhada dos dados, é possível retirar conclusões sobre os investimentos feitos nos projetos de investigação ocorridos até à data e ajudar nas futuras decisões de investimento cruciais ao funcionamento saudável da instituição de ensino. Assim, com esta análise procura-se, não só melhorar a gestão financeira da área em questão, mas também perceber até que ponto a utilização de técnicas de *Machine Learning* pode ser útil na análise de dados relativos à execução financeira de projetos de I&D.

Para além disso, uma área que está altamente relacionada com os projetos de investigação, não podendo ficar alheia à mesma, é a produção científica. A disseminação de conhecimento científico é uma parte essencial do trabalho de investigação levado a cabo em qualquer área, pelo que é extremamente importante que também este tema seja estudado e introduzido no âmbito desta dissertação.

Palavras-Chave: Gestão de Projetos, Projetos de I&D, *Data Visualization*, *Business Intelligence*, Análise Inteligente de Dados, *Machine Learning*, Produção Científica

CONTENTS

1	INTRODUCTION	1
1.1	Motivation	2
1.2	Objectives	3
1.3	Research Hypothesis	4
1.4	Methodology Approach	5
1.5	Document Structure	5
2	RELATED WORK	7
2.1	Business Intelligence Systems	7
2.2	Business Intelligence in the Academic Environment	8
2.3	Background	9
3	STATE OF THE ART	13
3.1	Research and Development Projects	13
3.1.1	Factors of Success	14
3.1.2	Project Life Cycle	15
3.1.3	R&D at Portugal	16
3.2	Knowledge Discovery Methodologies	17
3.2.1	Knowledge Discovery in Databases	18
3.2.2	Cross Industry Standard Process for Data Mining	21
3.2.3	Sample, Explore, Modify, Model, Assess Process	23
3.3	Data Science	25
3.3.1	Statistics	27
3.3.2	Machine Learning	31
3.4	Scientific Production	40
3.4.1	Scientific Databases	41
3.4.2	Bibliometric Indicators	42
4	DEVELOPMENT	44
4.1	Data Visualization Platform	44
4.1.1	System Architecture	45
4.1.2	The Dataset	45

4.2	Covid-19 Impact on R&D Projects	47
4.3	The Relationship between Financial Execution in R&D and Scientific Production	50
4.3.1	Financial Execution	51
4.3.2	Scientific Production	53
4.3.3	Correlation between Financial Execution and Scientific Production	55
4.4	Scientific Production in Portuguese Public Universities	56
4.4.1	Analysis of the Scientific Production of Portuguese Public Universities	57
4.4.2	Comparison of Scientific Production Indicators	60
4.5	Making Predictions on R&D Projects	63
4.5.1	Expenses Forecast	63
4.5.2	Success Forecast	66
5	CONCLUSIONS AND FUTURE WORK	68
5.1	Conclusions	68
5.2	Scientific Publications	70
5.3	Future Work	71
A	APPENDIX	81
a.1	Level 1: Researcher	82
a.2	Level 2: Research Center	86
a.3	Level 3: Organic Unit	86
B	DETAILS OF RESULTS	87

LIST OF FIGURES

Figure 1	Volume, impact and quality of the UMinho's research output	10
Figure 2	Factors that Influence R&D Project Outcome	14
Figure 3	Project Life Cycle	15
Figure 4	Evolution of Gross domestic spending on R&D (1981 - 2019)	16
Figure 5	KDD process model according to Fayyad et al.	19
Figure 6	CRISP-DM life cycle	23
Figure 7	SEMMA Steps	24
Figure 8	Data Science Areas	26
Figure 9	Statistics Applications Worldwide	29
Figure 10	The main ingredients of ML	31
Figure 11	Example of voting technique	35
Figure 12	General Decision Tree	36
Figure 13	Evolution of scientific production indexed in the Scopus database	40
Figure 14	Platform System Architecture	45
Figure 15	Total percentage of budget spent on missions	48
Figure 16	Percentage of budget spent on missions per year	48
Figure 17	Merge illustration from UMinho papers with bibliometric indicators from Scopus	51
Figure 18	Sum of UMinho's expenditures per year	51
Figure 19	Expenses by rubric	52
Figure 20	Expenses by Research Center	52
Figure 21	Expenses by Organic Unit	52
Figure 22	Amount of scientific Production by Research Centre	54
Figure 23	Amount of scientific Production by Organic Unit	54
Figure 24	Amount of scientific production for the years 2017-2020	57
Figure 25	Correlation between the amount of papers and the indicators	58
Figure 26	Bibliometric Indicators behavior	59
Figure 27	Graphic representation of the behavior of the two algorithms	65
Figure 28	Platform Login Screen	81
Figure 29	Researcher Main Screen	82
Figure 30	Researcher Main Screen (continued)	82
Figure 31	Researcher Main Screen (continued)	83
Figure 32	Main screen of a specific project	83
Figure 33	Screen with expenses by rubric of a specific project	84

Figure 34	Screen with expenses by rubric of a specific project (continued)	84
Figure 35	Screen with the expenses of a specific rubric	85
Figure 36	Screen with the expenses of a specific rubric (continued)	85
Figure 37	Screen with payment requests and payment summary	85
Figure 38	Research Center Main Screen	86
Figure 39	Screen with researchers' names and project titles with very low execution rate	86
Figure 40	Organic Unit Main Screen	86
Figure 41	Expenses by Research Center: Complete Information	87
Figure 42	Amount of Scientific Production by Research Center: Complete Information	87

LIST OF TABLES

Table 1	Organic units and research centers of University of Minho	12
Table 2	CRISP-DM vs. KDD vs. SEMMA	25
Table 3	Descriptive vs Inferential Statistics Elements	30
Table 4	27 Scientific Areas described in Scopus	42
Table 5	Study Projects	49
Table 6	Project 19 Expenses	50
Table 7	Quantity and quality of Scientific Production at UMinho using Scopus indicators	53
Table 8	Bibliometric Indicators Top 5 Areas: UMinho	55
Table 9	Correlation between financial execution and scientific production	56
Table 10	Quantity and quality of Scientific Production from Portuguese Public Universities	57
Table 11	Correlation between quantity and quality of Scientific Production	58
Table 12	Correlation matrix of bibliometric indicators	59
Table 13	Bibliometric Indicators Top 5 Areas - Portuguese Public Universities	60
Table 14	Top 5 Universities per Bibliometric Indicator	61
Table 15	Bibliometric Indicators Weights	62
Table 16	RMSE and R^2 results	64
Table 17	Confusion Matrix (Act/Pred)	67

ACRONYMS

A

AI Artificial Intelligence

B

BI Business Intelligence

C

CRISP-DM Cross Industry Standard Process for Data Mining

D

DM Data Mining

DS Data Science

DV Data Visualization

G

GDP Gross Domestic Product

H

HEI higher education institution

K

KD Knowledge Discovery

KDD Knowledge Discovery in Databases

M

ML Machine Learning

O

OU Organic Unit

P

PMBOK Project Management Body of Knowledge

R

RC Research Center

RF Random Forest

RL Reinforcement Learning

R&D Research and Development

S

SEMMA Sample, Explore, Modify, Assess

U

UMINHO University of Minho

INTRODUCTION

"Every new beginning comes from some other beginning's end."

— Seneca

The evolution of information technology and the continuous need to adapt to the environment in which it operates are determining factors in the way which organizations progress and their businesses develop. The information system is an intrinsic element to any organization, made up of professionals, procedures, computer applications, data, equipment, among other fundamental constituents that guarantee its correct use and innovation. Nowadays, all information systems have migrated to a technological aspect, leaving aside the most rudimentary methods, such as the filing rooms where thousands of paper files were stored. Thus, an institution's information system, such as a university, needs to ensure continuous support and a well-structured organization.

In Higher Education Institutions, a significant part of the information system is linked to the management of research and development (R&D) projects, where the data of hundreds of executed and in execution projects are stored, representing a large amount of national and host institution investment.

The financial management of R&D projects cannot be neglected, so the correct storage of data in information systems is not enough to obtain good levels of financial execution, since this management is carried out by personnel outside the Financial and Patrimonial Services Unit (USFP) of the University of Minho. This unit deals with the financial management the university, which includes the financial management of R&D projects. Project managers, researchers, and other regulatory bodies must have quick and easy access to the data of the projects they intend to manage, as well as a simple mechanism to understand them.

The data used by the various regulatory entities and management centers at the UMinho are available to those responsible for them in an unintuitive, dispersed and, consequently, poorly organized manner. This dispersion, besides making access to information very difficult, is not compatible with the data organization that a higher education unit needs to have.

Given this need, it was proposed to develop a Data Visualization (DV) platform in order to provide an organized and coherent data visualization, according to the needs of its stakeholders, in other words, interested parties in the scope of project management and its monitoring.

This platform works with 4 different levels of responsibility. Each level has access to a certain kind of information, according to the position's in the hierarchy. Thus, it is important to understand what information we must show at each level, as well as what knowledge is, in fact, useful knowledge to support the investment decision.

As already mentioned, nowadays, we live in a highly changed community and very dependent on technologies, in particular, on Artificial Intelligence (AI). This recent innovation brings us more and more chances to improve and facilitate the work done in our daily lives, being already widely used in areas such as medicine. For this reason, it is also very promising and interesting to study its influence in this field. So, this dissertation also focuses on the possibility of performing, correctly and coherently, better financial management of the R&D projects, using AI techniques.

Finally, the evolution and analysis of the scientific production produced to support research projects have also become very relevant, since these are two closely related areas. Thus, we intend, above all, to analyze the relationship between the financial execution of the University of Minho's projects and the scientific production produced by its researchers.

1.1 MOTIVATION

Nowadays, more important than having information is knowing how to analyze it and, a posteriori, understand it. For this reason, the available data must be aggregated and made available in the most direct way possible, for example, with graphs illustrating the aggregation of information, making its visualization easier and aesthetically appealing. This aggregation represents a powerful advantage in the statistical understanding of the data since without it a perspective of the surrounding environment is lost. With an adequate presentation and logistical aggregation, we achieve a good analysis and understanding of the data, essential to the success of the organization/institution. Institutions capable of providing their managers and regulators with data visualization platforms, such as the one suggested in this dissertation, tend to substantially increase their financial performance since the organization of information also becomes more careful and accurate.

Besides this development focusing on the area of project management and web development to better reproduce its monetary life cycle and thus provide an opportunity to improve the organization and understanding of these data, the artificial intelligence area also has an important role in monitoring these investments. In order to maximize the correct investment in the various projects in execution, it becomes very relevant to study their evolution over time, as well as what factors can be decisive for their success. Thus, we can apply Machine Learning techniques that allow us to forecast project expenditures and even, from there, study the probability of the project to have or not a positive financial execution rate.

Finally, the analysis of the scientific production is also a major focus of this dissertation, since the opportunity arises to relate the financial execution of research and development projects with the dissemination and production of scientific knowledge.

Reduce the time used on manual data aggregation by people responsible for project management, with many different and tools; minimize the time spent on its indirect analysis; increase the productivity of the management team; using Machine Learning techniques as a way of forecasting and supporting investment decisions; and the

analysis of the relationship between the project's financial execution and their scientific production, are the main reasons for the development of this dissertation.

1.2 OBJECTIVES

This dissertation has as main objective the development of a DV platform with project execution data sources from the Financial and Patrimonial Services Unit of the University of Minho in order to build a system for visualization, maintenance, and analysis of data related to the financial execution of R&D projects at the University of Minho. We can summarize the principal objectives of this dissertation in the following guidelines:

- Research on Project Management and its financial execution;
- Research on success factors in the execution of R&D projects;
- Survey of requirements for the development of a data visualization platform;
- Development of an organized data visualization platform that allows an intuitive data understanding;
- Assist project management in investment decisions:
 - Study the financial evolution of projects;
 - Make expenses and costs forecasts to better manage financial investment overtime.
- Scientific production analysis:
 - Research on the relationship between the financial execution of R&D projects and the quantity and quality of the scientific production;
 - Study the scientific production made by UMinho;
 - Study the scientific production in Portugal.
- Answer the questions:
 - Can we predict future financial difficulties in R&D projects using ML techniques?
 - What are the advantages of using data visualization platforms?
 - Is there any relationship between the quantity and quality of scientific production and the financial execution of R&D?

1.3 RESEARCH HYPOTHESIS

The introduction of a new tool in a person's daily life can be quite time-consuming. There are countless reasons for this aversion, such as the need to adapt to it, the different design, among others. To combat the possible reasons for disagreement in the use of new platforms it is important to know, first of all, the needs of its users and reconcile this with a visually appealing and easy to use graphic interface.

Besides, it is equally important that the platform is an asset, that is, that it brings something to users that it would be much more costly to do otherwise or that it could not do at all. In the case of this dissertation, this added value is found in three distinct areas:

- **Data aggregation:** Users currently have a large amount of data dispersed. Several users have developed some data aggregation techniques in order to facilitate their own work, however, they suffer from some limitations:
 1. the time spent in its aggregation is abysmal, leading users to invest part of their time outside of work in this solution, both in its development and in the resolution of any errors;
 2. it is done using standard tools and they are not prepared to present the results of this aggregation cohesively, contrary to what is expected with a web application.

With the development of a DV platform, this concern disappears. The data will be aggregated automatically, coming directly from the source, obtaining a much lower probability of error.

- **Understanding the data:** Although those who deal with the data have great knowledge about the subject, understanding the data is often not so trivial. This factor is aggravated by the dispersion of the data. Thus, with the aforementioned aggregation of data and subsequent visualization of these data in a more prone environment, with greater visual and statistical organization, the understanding of results by users will be greatly facilitated.
- **Forecasting financial characteristics:** While data aggregation and visualization can be considered a simpler activity, the attempt to forecast the future is not so seen. In this way, this dissertation also intends to provide results about possible financial features of R&D projects, thus becoming an enormous asset in the financial management of projects.

Our research hypothesis are: (1) if we use Machine Learning techniques, we can predict future financial characteristics in R&D projects; and (2) the use of data visualization platforms is a huge advantage for its users.

1.4 METHODOLOGY APPROACH

Once this preamble has been made, the case study of the University of Minho is used as a research strategy, according to a partnership with the Rectory of the University.¹

The need to develop a system capable of responding to the difficulties of all those involved in the financial management of R&D projects led to the creation of this partnership through a research grant with an open call made available to carry out work in the Pro-Rectory for Research and Projects.

Today, the University of Minho (UMinho) is undoubtedly a university of research. That can be seen in the position the university occupies in the most important international rankings, such as the Leiden Ranking, which is the main bibliometric ranking associated with scientific research. Other important rankings in which the University of Minho is well-positioned are the Times Higher Education (THE) ranking, both in their global ranking and in the THE ranking of universities under-50 years old.

There has also been a clear increase in scientific publications and other diverse scientific outputs coming from the University of Minho throughout the years. UMinho's researchers are considered to be highly cited, with a big number of citations and h-indexes in their scientific areas.

UMinho is one of the Portuguese HEIs with the highest success rate in winning large European projects and it is the only university in the country coordinating projects in all the fields of the European Commission's (EC) Widening programme.

In 2018, UMinho had 25 258 611€ of European funding, 5 808 480€ of Internacional funding, 122 713 451€ of National funding from PT2020 and 7 404 111€ of other National funding, making a total funding of 161 184 653€ from the different Organic Units of UMinho, as [University of Minho \(2018\)](#) reported.

1.5 DOCUMENT STRUCTURE

This document was structured into six chapters, organized as follow:

1. Introduction

The first chapter begins with a brief description of the problem and an introduction to the main concepts of the thesis theme. An explanation about the motivation behind this work is also given, as well as the main objectives, the methodology approach and the document structure.

2. Related Work

This chapter presents a short disclosure about some related work that was already done, in order to serve as a basis for the work developed. Here are described some solutions similar to the proposal for this dissertation. It also describes the background of this dissertation, that is, the explanation of how the theme of this thesis arose and the higher education entity that hosted this investigation.

3. State of the Art

The state of the Art chapter is mainly related to project management, R&D projects, machine learning

¹ Supported by University of Minho through a research grant - PRT-FV-001/2020.

techniques and scientific production analysis worldwide. It's explained how some techniques are used nowadays and what can we do with them to improve the financial execution of projects, supporting the theme of this dissertation.

4. **Development**

After all the necessary literature review, we proceed to the presentation of all the work developed. This chapter detailedly explains how the work was developed and what themes were explored to get satisfactory conclusions.

5. **Conclusions and Future Work**

The last chapter of this dissertation summarizes the work that has been done and the main conclusions to be drawn from it, as well as what can be improved. The limitations found throughout the research, the future work that can be done and the scientific publications carried out in support of this dissertation are also mentioned.

RELATED WORK

“How can we prepare for the future if we won’t acknowledge the past?”

— N.K. Jemisin

This chapter aims to demonstrate work related to the topic in question. Despite this being a very broad topic and with numerous applications, this chapter focuses on studying Business Intelligence systems already in place and describing their comprehensive use.

2.1 BUSINESS INTELLIGENCE SYSTEMS

According to [Zulkefli et al. \(2015\)](#), decision-making is a critical issue in every organization to achieve business performance indicator goals and improve business processes. Therefore, appropriate data plays a crucial role in effective decision-making.

An increase in the number of data in Higher Education Institutions (HEI) leads to an increase in the complexity of its management and handling. Each institution has several sectors that, in turn, have very important data for the performance of the university. Thus, a Business Intelligence (BI) tool can help in the organization and management of data.

Using BI tools has grown immensely since most organizations have numerous data and information that are random, dispersed, and unrelated. [Bentley \(2017\)](#) defined Business Intelligence as *a set of techniques and tools for the acquisition and transformation of raw data into meaningful information for business analysis purposes*. The main goal of BI is to allow easy interpretation of large volumes of data. Usually uses reports, analytics, business performance management, among others.

Thus, according to Primak’s work [Primak \(2008\)](#), BI tools bring some advantages to institutions, such as:

- software cost reduction;
- reduced administration and support costs;
- cost reduction in project evaluation;
- cost reduction with training for employees;

- greater control and less incorrect data;
- greater information security;
- ease of access control and definition of management levels;
- better user management;
- greater efficiency in obtaining information for decision-making;
- consistent information across multiple dispersed locations.

However, Primak (2008) also lists some difficulties when implementing BI systems:

- dispersed and often inconsistent operational data;
- deficiency in the operating systems used by organizations, which often do not store useful data for future decision-making;
- need for a good relationship between the business area and the technological team;
- obtaining information from several external sources is done in a way that may not be favorable, that is, it does not compensate in terms of cost-benefit;
- the treatment and storage of data is a process that must be well planned, since it is complex, so capable professionals are needed to guarantee the success of this stage;
- implementation of a BI system is not cheap.

2.2 BUSINESS INTELLIGENCE IN THE ACADEMIC ENVIRONMENT

There are numerous possible applications for Decision Support, Data Visualization and Business Intelligence systems, but these have been increasingly adopted in the academic/higher education environment. In higher education, BI is a very promising solution for adding much-needed efficiency at the operational level - Guster and Brown (2012). In fact, Dell'Aquila et al. (2008) claimed that the management of a University is as critical as the management of a big business company since the factors affecting a great University's management are the same as those involved in the business processes. Authors like Guster and Brown (2012) even defend that the problem is often even more difficult when it is applied to higher education, particularly if the institution is public. Most institutions have an extensive amount of data and, given the current economic environment, it is absolutely necessary to run the institution more efficiently.

Universities are organizations with an important social responsibility since it's in them that a large part of the knowledge that supports the economic development of any society is generated and transmitted - Dixson et al. (2015). They can use these systems for various purposes, even having several systems for various sectors, but, in general, the objective is to obtain data and turn it into useful information for decision-makers, such as Drake

and Walz (2018) affirmed. There are still some systems created by companies whose objective is exactly the same, however, each educational institution has its own internal organization, and it is often not possible to use pre-designed systems. Thus, there is a great need for each educational institution to have its own systems, in order to obtain the necessary flexibility and functionality.

Several countries and institutions have already implemented systems that aim to respond to their needs. For example, Mansmann and Scholl (2007) created a Decision Support System (DSS) for Managing Educational Capacity, i.e., for assessing educational capacity and planning its distribution and utilization; Feghali et al. (2011) developed a web-based Decision Support Tool for Academic Advising for the American University of Beirut; in Thailand, Kleesuwan et al. (2010) developed a BI tool to manage Thailand's Higher Educational Resources; and even a DSS for forecasting student's grades making use of machine learning techniques were implemented by Kotsiantis (2012).

Although there are some implementations of these systems in the scope of Research and Development (R&D) projects such as Khatibi et al. (2017), there is a lack of information about systems implemented in higher education institutions in this area. The same happens concerning the financial management of these projects in the university context, that is, although there is research in financial management using DSS like EDDIE-Automation, explained in Tsang et al. (2004), research with a focus on financial data from R&D projects in universities is scarce. This lack can be easily explained by the sensitivity and privacy of the processed data, which does not mean that other universities (mainly Portuguese) have not already developed identical platforms. However, since there is a failure to obtain more examples, we consider the creation of a DV platform for the financial management of research projects quite innovative and fundamental to the success of a higher education institution.

2.3 BACKGROUND

Founded in 1973, the University of Minho (UMinho) is renowned for its competence and quality in several fields.

UMinho has 32 Research Units and 75% of those were recognized with the three highest ratings in the International Research Unit Evaluation Programme, conducted by the National Science Foundation (FCT). In fact, as we can see in the image below, UMinho has been improving itself in the field of research, being nowadays a research university, committed to the value chain of knowledge, namely: Research, Development, and Innovation, as stated in University of Minho (2018). In this chart, the scientific performance evolution of the University of Minho (2009 - 2021) in the Research Rank made by SCImago Institutions Rankings is represented.

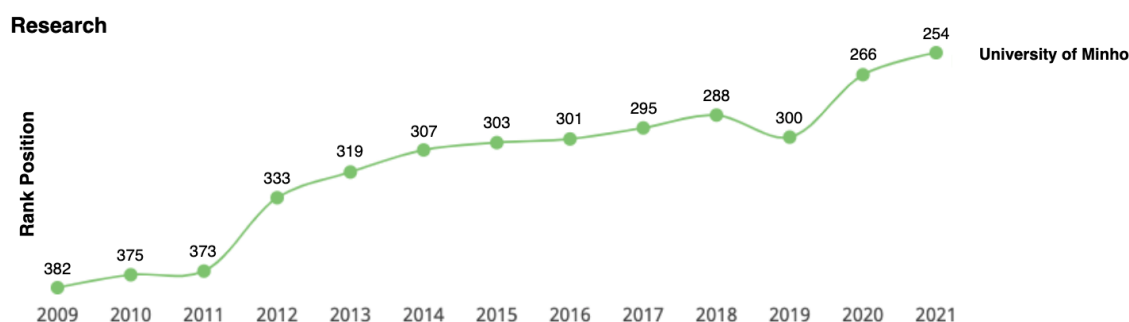






Figure 1: Volume, impact and quality of the UMinho's research output

To achieve these levels of success, an exemplary organization is essential. Many factors influence the success of projects, but one of the most frequently mentioned by authors like [Cooke-Davies \(2002\)](#); [Goldstein and Kearney \(2020\)](#); [Mahmood et al. \(2014\)](#); [Nagesh and Thomas \(2015\)](#) is research funding.


A project may contribute substantially to the economic welfare of a country, but if the implementing institution lacks the funds to finance it, project implementation will suffer – Belli (1999).

That said, we can say that an educational institution must have a great deal of control over its finances so the financial execution of the projects to be carried out successfully.

Organic Unit	Research Center
 Institute of Arts and Humanities	Centre for Humanistic Studies (CEHUM)
	Centre for Ethics, Politics and Society (CEPS)
 Institute of Education	Research Centre on Education (CIEd)
	Research Centre on Child Studies (CIEC)
 Institute of Social Sciences	Communication and Society Research Centre (CECS)
	Centre of Studies on Geography and Spatial Planning (CEGOT)
	Interdisciplinary Centre of Social Sciences (CICS)
	Centre for Research in Anthropology (CRIA)
	Landscape, Heritage and Territory Laboratory (Lab2PT)
 Research Institute I3B's	3B's Research Group (3B's)

Continued on next page

Table 1 – Continued from previous page

Organic Unit	Research Center
 School of Architecture	Landscape, Heritage and Territory Laboratory (Lab2PT)
 School of Economics and Management	Research Centre in Political Science (CICP) Centre for Research on Economics and Management (NIPE)
 School of Engineering	Mechanical Engineering and Resource Sustainability Centre (MetRICs) Algoritmi Centre (CAIlg) Centre of Textile and Science Technology (2C2T) Centre of Biological Engineering (CEB) Centre for Territory, Environment and Construction (CTAC) Institute for Polymers and Composites (IPC) Institute for Sustainability and Innovation in Structural Engineering (ISISE) High-Assurance Software Laboratory (HASLab) Centre for Microelectromechanical Systems (CMEMS)
 School of Law	Research Centre for Justice and Governance (JusGov)
 School of Medicine	Life and Health Sciences Research Institute (ICVS)
 School of Nursing	Health Sciences Research Unit: Nursing (ESE/UICISA)
 School of Psychology	Psychology Research Centre (CIPsi)
 School of Sciences	Centre of Plant Functional Biology (CBFP) Centre of Molecular and Environmental Biology (CBMA)

Continued on next page

Table 1 – *Continued from previous page*

Organic Unit	Research Center
	Centre of Earth Sciences (CCT)
	Centre of Physics (CFUM)
	Centre of Mathematics (CMAT)
	Centre of Chemistry (CQUM)
	Laboratory for Instrumentation and Experimental Particle Physics (LIP)

Table 1: Organic units and research centers of University of Minho

STATE OF THE ART

"Any knowledge that doesn't lead to new questions quickly dies out."

— Wislawa Szymborska

In order to explore in more detail other areas that are related, in some way, to the financial execution of R&D projects and what benefits they can bring, some literature has been studied. In other words, this research aims to better understand the projects and their management, as well as their relationship with other areas. Thus, this chapter focuses largely on the technological enrichment derived from the development of the platform. The use of ML techniques to make predictions, the process of knowledge discovery, and the relationship between research projects and scientific production were the most discussed topics, as they are crucial for the correct progress in the development and exploration of the case of study.

3.1 RESEARCH AND DEVELOPMENT PROJECTS

An R&D Project integrates several activities that are related to innovation and are directed towards the development of new products or applied research, which may eventually enable the future development of a product.

According to the PMBOK Guide [PMI \(2008\)](#) a project can be defined as a *temporary endeavor undertaken to create a unique project service or result*. However, it is also important to realize that Project Management is a different thing, since it is defined by the Association for Project Management [[Murray-Webster and Dalcher \(2019\)](#)] as *the application of processes, methods, skills, knowledge and experience to achieve specific project objectives according to the project acceptance criteria within agreed parameters*.

According to [Sugandhavanija et al. \(2011\)](#), universities play a very important role in the development of technology and the knowledge base that underpins the process of economic growth, both in developed and developing countries. Furthermore, as [Closs et al. \(2012\)](#) stated, a great effort is needed between universities, the government and companies to provide an adequate scientific environment and apt for the development and innovation of knowledge dissemination.

The area of research, of its projects and management is a very vast area and, consequently, complex when it comes to measuring its performance. This performance is difficult to measure because it is associated with

great uncertainty and risk and, for this reason, the research related to the success factors of project management has been growing considerably, as mentioned by Barnes et al. (2006). Therefore, it is essential that the entities where the research projects take place know their success factors, so they implement projects capable of being successful.

3.1.1 Factors of Success

Firtsly, project success is not the same as project management success. The latter is commonly measured by the management of the following constraints: budget, time and target/objective; project success depends on its outcome, impact and the satisfaction of its stakeholders. Also, studies shows that the project success or failure depends on the project management process.

There are many studies such as Mahmood et al. (2014); Cooke-Davies (2002); Saito and Lezana (2015); Nagesh and Thomas (2015) that aim to determine the success factors of the projects and their management, since this is a topic of interest since the early 1960s.

Despite the need to continue studying this topic and the existence of numerous success factors that differ from project to project, there are some factors that we can consider "essential", since they appear many times in the studies carried out. The management and monitoring of projects, the availability of resources and their adequacy, the management of human resources and the budgeting/financial support are some of the most frequently listed. All these factors are closely linked to financial management, which is why we reiterate the importance of developing a platform capable of managing research projects.

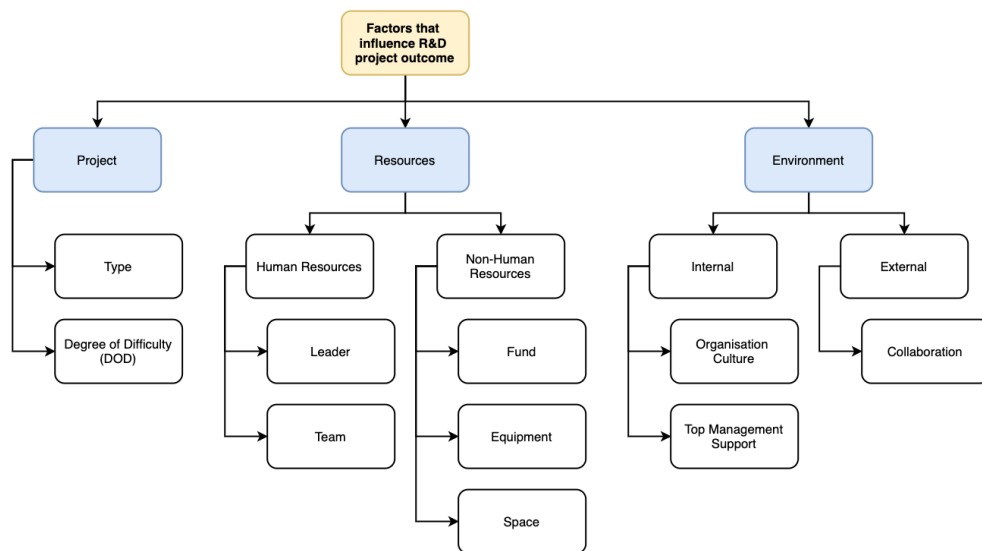


Figure 2: Factors that Influence R&D Project Outcome

3.1.2 Project Life Cycle

The Project Life Cycle is characterized by [Ward and Chapman \(1995\)](#) as a convenient way of conceiving the generic structure of projects over time. Since these are associated with uncertainty, organizations that perform projects tend to divide each project into several phases in order to improve its management and monitoring. Regardless of their genesis, projects go through a very similar life cycle from the beginning to the end, characterized by identical work phases. The set of these phases is known as the life cycle of a project, as claimed in [PMI \(2008\)](#).

For this work, we admit five phases for the project life cycle: Initiation, Planning, Execution, Monitoring and Controlling and Closure. Each of the phases is characterized by the completion and approval of one or more deliveries, related to final or intermediate outputs.

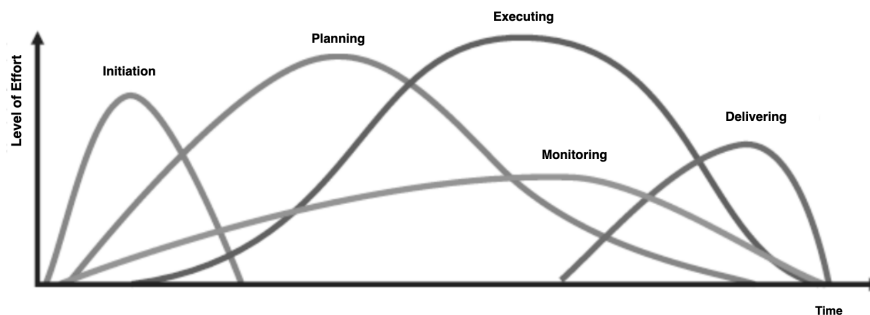


Figure 3: Project Life Cycle

1. **Initiation Phase:** characterized by the beginning of the project, definition of its specifications, establishment of objectives, formation of project teams and the assignment of the main responsibilities;
2. **Planning Phase:** characterized by an increase in the level of activity/intensity and the development of plans to determine the implications of the project, its scheduling, the key tasks, its costumers, level of quality and the budget;
3. **Execution Phase:** characterized by a high level of activity due to the production of result;
4. **Monitoring & Controlling Phase:** monitoring and control of previously defined parameters, such as time, cost and specification measures defined by stakeholders;
5. **Closure Phase:** delivery of the project's product to the customer, redistribution of project resources and the revision after its completion.

3.1.3 R&D at Portugal

Álvarez Bornstein and Bordons (2021) stated that research is an important driver of social and economic development of countries, therefore every year, a lot of money is spent on research and development through various funding agencies. This investment is made to provide the best possible working conditions (such as technical and human resources) to achieve positive and promising results. In 2019, Portugal invested 2.991.864,1 thousand euros in R&D, the highest value ever, such as described in PORDATA (2021); OECD (2021). This value corresponds to 1.40% of GDP, but the highest value of GDP was 1.58% reached in 2009 as shown in figure 4. Besides, the amount corresponding to expenses in higher education is 1.210.653,0 thousand euros, which is also the highest amount of expenditure on R&D for this sector. This leads us to believe that research has come to be valued and occupies, more and more, an important place in the researchers' lives and scientific advancement of the world.

R&D investments are not made on the expectation of immediate payoffs but rather on the expectations of creating future investment opportunities that will be profitable – Herath and Park (1999).

In Portugal, the Higher Education sector includes all universities, higher institutes, polytechnic institutes, and other post-secondary education institutions, regardless of the source of their financial resources and legal status. It also includes all institutions that operate under the direct control of or are managed by higher education establishments. Still according to PORDATA (2021), the sector also includes private non-profit institutions controlled and largely funded by Higher Education.

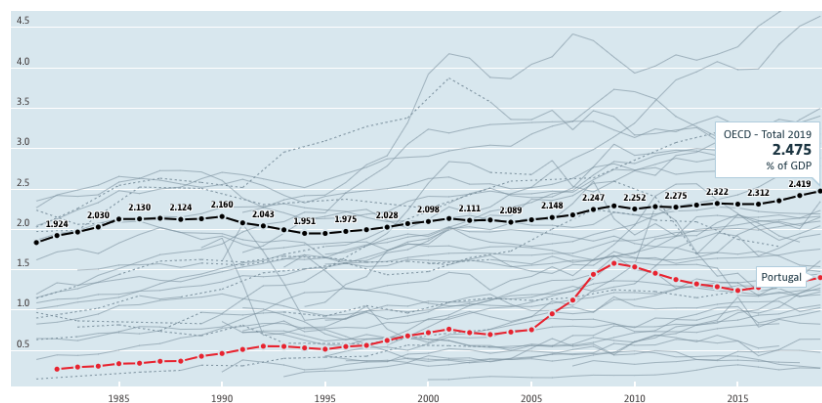


Figure 4: Evolution of Gross domestic spending on R&D: comparison between the OECD total and Portugal for the years 1981 - 2019. Image taken from OECD (2021).

The Organisation for Economic Co-operation and Development OECD (2021) defined Gross Domestic Spending on R&D as *the total expenditure (current and capital) on R&D carried out by all resident companies, research institutes, university and government laboratories, etc., in a country. It includes R&D funded from abroad, but excludes domestic funds for R&D performed outside the domestic economy.*

3.2 KNOWLEDGE DISCOVERY METHODOLOGIES

As already mentioned, this dissertation is also related to the acquisition of knowledge in order to extract useful information for users. This extraction is now more than ever essential to the development of the industry, which is increasingly dependent on information technology and, consequently, information computerization. Nowadays any Small and Medium-sized Enterprise (SME) or even any worker needs to organize their data easily and intuitively, often going through the storage of data in large databases. However, the evolution of this sector leads us to the need to draw conclusions from these data: knowledge extraction.

Discovery (or extraction) of knowledge in large datasets is made possible thanks to the so-called Knowledge Discovery (KD) methods (also known as Data Mining methods). There are several methods of achieving this goal, all of which are based on Artificial Intelligence. As such, this dissertation presents the three best known methods for discovery knowledge from data in the following subchapters.

Still, before explaining each of these methods, it is important to realize the growing use that these methods have been achieving. As such, based on the research work of Rogalewicz and Sika (2016) and Chawla (2005), we can affirm that at the beginning of the 21st century, the KD Methods are increasingly used because of the following reasons:

- The number of datasets and their size have been increasing a lot, so the computational resources need to accompany this growth. The only problem lies in discovering useful knowledge in these sets.
- Human beings are not able to process large amounts of data like this in a manual and fast way. Thus, automatic ways of extracting knowledge (such as KD methods) are essential.
- KD methods help in decision making (e.g.: preparing prognoses and detection of frauds)
- KD methodologies don't require performing very expensive experiments since they are based on already gathered data.
- KD methodologies allow obtaining knowledge from datasets that are noisy, contain missing values or correlated variables. These sets aren't dealt well by the traditional data analysis methodologies.
- These methods can be applied to a wide spectrum of problems.
- The increase in literature about the theme is notable and presents successful implementation of KD methods. Also, dedicated methods were created to realize the knowledge extraction process, presenting step by step ways of conduct.

However, as with everything, there are also disadvantages/limitations in these methods:

- Ensuring the security of data stored in databases can be a problem, as can extracting knowledge from them.

- Knowledge extraction can be used for the wrong reasons (such as against safety of a company, a country or its citizens).
- The incorrect use of KD methods can lead to wrong results, which means inadequate conclusions and decisions made based on them.
- Implementation of a well-functioning system, which systematically utilizes the Knowledge Discovery methods, requires the application of large resources (maintenance and updating of databases, hiring specialists for knowledge extraction, ...), and not every company is capable of doing it.
- Users may have wrong expectations about using these methods. For example, they may expect that these methods will replace their conclusions drawing and decision making.

3.2.1 Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD), coined in 1989, is a broad area that integrates concepts and methods from several disciplines, trying to make sense of data and extract useful knowledge from them. It is the process of extracting hidden knowledge from databases. Research work in this field has been done by some researchers like Ioannis Kavakiotis and Ruhul Sarker in [Kavakiotis et al. \(2017\)](#); [Sarker et al. \(2000\)](#). In consonance with their work, it is considered to be a multistep iterative process composed of the following five steps: (1) data selection, (2) preprocessing, (3) transformation, (4) data mining and (5) interpretation, as we can observe in figure 5.

A generally accepted definition is given by Usama Fayyad in [Fayyad et al. \(1996a\)](#) in which KDD is defined as the *nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*.

According to this definition and David Taniar and Usama Fayyad's work [Taniar \(2008\)](#); [Fayyad et al. \(1996b\)](#), we can withdraw some assumptions:

1. Data is a set of facts that is somehow accessible in electronic form;
2. The term pattern indicate an expression in some language describing a subset of the data or a model applicable to that subset;
 - ⇒ Patterns have to be valid (i.e., they should be true on new data with some degree of certainty);
 - ⇒ We want patterns to be novel (at least to the system, and preferably to the user);
 - ⇒ A novel pattern is not previously known or trivially true ;
 - ⇒ The potential usefulness of patterns refers to the possibility that they lead to an action providing a benefit;
 - ⇒ Extracting a pattern also designates fitting a model to data, finding structure from data, or in general any high-level description of a set of data;
 - ⇒ Patterns should be understandable, if not immediately then after some post-processing;

⇒ A pattern is understandable if it is interpretable by a human user;

3. The term process implies that KDD is comprised of many steps, which involve data preparation, search for patterns, knowledge evaluation, and refinement, all repeated in multiple iterations;
4. Nontrivial means that some search or inference is involved, i.e., it is not a straightforward computation of predefined quantities like computing the average value of a set of numbers.

However, it is still necessary to understand the various steps that make up the KDD Process, especially because there are different views of what should be the steps for KDD. While [Fayyad et al. \(1996a\)](#) identify nine steps, [Sarker et al. \(2000\)](#) identify thirteen and many other researchers have identified others. Although, the universally accepted steps are five and meet Fayyad's definition. They are just less detailed, encompassing more phases in fewer steps.

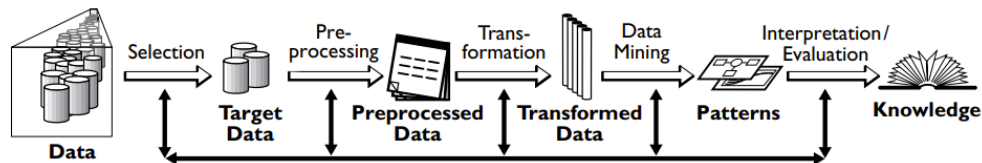


Figure 5: KDD process model according to Fayyad et al.

- **Data Selection**

Consists on creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.

Much of the available data nowadays may be redundant and therefore, it hinders the knowledge extraction as well as making it more time and resource consuming. After an initial study, suitable and unsuitable data should be identified. Therefore, in order to ease the upcoming KDD phases, it is very important to select the potentially useful data.

- **Preprocessing**

This stage consists on the target data cleaning and preprocessing in order to obtain consistent data.

Nowadays, datasets are prone to suffer from noise, outliers, missing values, and inconsistencies due to their big size and their probable origin from multiple and heterogeneous sources. Not only do these data quality issues compromise knowledge extraction algorithms' performance, but they also may have a negative impact on decision-making processes.

Different methods are applied to ensure quality of the data and prepare the data for a subsequent analysis such as the removal of noise if appropriate, deciding on strategies for handling missing data fields, and so on.

- **Transformation**

Transformation of the data into a form which data mining algorithms can work with and improve their performance.

Data transformation means reflecting the logical relations between the tables into a single table that contains all the information needed for the mining process. Many of the mining algorithms do not work on multiple-tables and therefore we need somehow to combine the tables into one.

- **Data Mining**

Searching for patterns of interest in a particular representational form or a set of such representations, depending on the data mining objective.

The user's role in this phase consists of selecting the suitable algorithm and fine-tuning it with the appropriate parameters. Furthermore, as each algorithm's performance may vary depending on the input data, the user expertise and even intuition at times also play a role in this phase. This phase is explained in [3.2.1](#).

- **Interpretation/Evaluation**

It's the final phase where the results, patterns and models derived are used to support decision-making processes.

This step can also involve visualization of the extracted patterns/models or visualization of the data given the extracted models.

The KDD process can involve significant iteration and may contain loops between any two of the mentioned steps, as demonstrated in [5](#). The necessity of having such a flexible process arises from the wide range of methods and parameter selections that can be applied in each step.

Additionally, the KDD process must be preceded by the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user. It also must be continued by the knowledge consolidation by incorporating this knowledge into the system, conforming to [Fayyad et al. \(1996b\)](#).

We can distinguish two types of knowledge discovery goals: (1) Verification, where the system is limited to verifying the user's hypothesis; and (2) Discovery, where the system autonomously finds new patterns.

Data Mining

The most important step within the process of KDD is Data Mining (DM). Data Mining consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data.

The terms Data Mining and Knowledge Discovery in Databases have been used interchangeably in practice, and the literature therefore creates some confusion, making it difficult to pinpoint each of the concepts. However, strictly speaking, and in agreement with [Sarker et al. \(2000\)](#), KDD refers to the overall process of deriving useful knowledge from data, while DM refers to a specific step in that process.

Most DM algorithms can be viewed as compositions of a few basic techniques and principles. In particular, as Sarker and Fayyad in [Sarker et al. \(2000\)](#); [Fayyad et al. \(1996a\)](#) reported, there are three primary components:

- **Model Representation:** Model representation is the language used to describe discoverable patterns. It determines both the flexibility of the model in representing the data and the interpretability of the model in human terms;
- **Model Evaluation Criteria:** Model evaluation criteria determine how well a particular model and its parameters meet the goals of the KDD process;
- **Search Method:** Consists of two components: parameter search and model search.

In parameter search, the algorithm searches for the parameter set for a fixed model representation that optimizes the model evaluation criteria based on the observed data. The model search takes place as a loop using the parameter search method: the model representation is changed so that a family of models is considered. For each specific model representation, the parameter search method is instantiated in order to evaluate the quality of that particular model.

3.2.2 Cross Industry Standard Process for Data Mining

Cross Industry Standard Process for Data Mining (CRISP-DM) is a comprehensive knowledge discovery methodology and process model that provides anyone from novices to data mining experts with a complete blueprint for conducting a data mining project. It breaks down the process of KD into six different phases shown in figure 6. Pete Chapman in [Chapman et al. \(2000\)](#) explains all these phases:

1. **Business understanding:** What does the business need?
This phase focuses on understanding the goal and the requirements of the project from a business perspective. In other words, uncovers important factors including success criteria, business and data mining objectives and requirements as well as business terminologies and technical terms. This understanding will then be transformed into a definition of data mining problems to create a project plan and achieving the goals.
2. **Data understanding:** What data do we have/need? Is it clean?
This phase starts with an initial data collection and then proceed with activities that enable you to become familiar with the data, identify data quality problems, discover first insights into the data and detect interesting subsets.
3. **Data preparation:** How do we organize the data for modeling?
Once the data collected, it is necessary to prepare them to construct the final dataset. Data preparation tasks are likely to be performed multiple times and not in any specific order. These may include many

tasks records, table and attributes selection as well as cleaning and transformation of data (e.g.: cleaning data from noise).

4. **Modeling:** What modeling techniques should we apply?

In this phase, various modeling techniques are selected and applied to the project and parameters get calibrated for the models to get the optimal value. Typically, there are several techniques for the same DM problem type and some of these techniques have specific requirements on the form of data. Therefore, going back to the data preparation phase is often necessary.

5. **Evaluation:** Which model best meets the business objectives?

At this stage in the project, you have created a model (or models) that appears to have high quality from a data analysis perspective. Before the model can be deployed and implemented, the work must be evaluated to ensure that the results meet the business requirements. The steps taken to produce the result are thoroughly reviewed and evaluated. The main objective is to determine if there is some important business issue that has not been sufficiently considered.

At the end of this phase, a decision on the data mining results should be reached.

6. **Deployment:** How do stakeholders access the results?

In this phase, the final model is deployed. This focuses on organizing, reporting and presenting the knowledge gained when needed. Depending on the requirements, the deployment phase can be as simple as creating a report or as complex as implementing a repeatable data mining process across the enterprise. In many cases, it is the customer, not the data analyst, who carries out the deployment steps. But even if the analyst takes over the deployment, it is important for the customer to understand in advance which actions have to be carried out in order to actually be able to use the created models. In other words, it is important to create a deployment plan so that it is clear what actions are required to carry out the deployment.

As mentioned in [Chapman et al. \(2000\)](#), there are no strict ways of moving between different phases of the processes, in fact, moving back and forth between them are required. It is the result of each phase that determines whether you should move on to the next step or repeat with the one above. The outer circle in figure 6 symbolizes the cyclical nature of data mining. Data mining doesn't end with deploying a solution. Even after a solution has been deployed, the process continues to create a better version, so the insights gained during the process and from the deployed solution can trigger new, often more focused, business questions. Subsequent data mining processes will benefit from the experiences of previous ones.

According to [Mariscal et al. \(2010\)](#), CRISP-DM is considered the standard for developing Data Mining and Knowledge Discovery projects since 1999.



Figure 6: CRISP-DM life cycle

3.2.3 Sample, Explore, Modify, Model, Assess Process

SEMMA is an acronym for Sample, Explore, Modify, Assess and it was developed by the SAS Institute. SAS Institute defines SEMMA as a logical organization of the functional tool set of SAS Enterprise Miner for carrying out the core tasks of DM.

As Umair Shafique argued in his research [Shafique and Qaiser \(2014\)](#), SEMMA offers and allows understanding, organization, development and maintenance of data mining projects, as well as providing the solutions for business problems and goals.

The main difference between the original KDD process and SEMMA is that SEMMA is integrated into SAS tools and it's unlikely to use SEMMA methodology out of them, while KDD is an open process and can be applied in very different environments. But, according to [Mariscal et al. \(2010\)](#), these are not the only differences. First, SEMMA skips the first step of the KDD process (learning the application domain) and proceeds directly to the sample step. Furthermore, SEMMA doesn't include an explicit step to use the discovered knowledge, while KDD includes this after the Interpretation/Evaluation step. These two steps are considered essential to successfully running a data mining project.

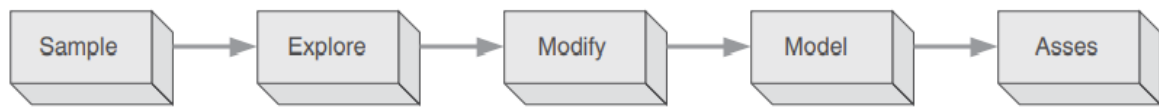


Figure 7: SEMMA Steps

As its name implies, SEMMA consists of five different steps, explained by Ana Azevedo and Manuel Santos in their research [Azevedo and Santos \(2008\)](#):

1. **Sample**

This phase focuses on sampling the data that will be used for modeling by extracting a portion of a large dataset. This portion must be large enough to contain important information, but small enough to be edited quickly and easily. This step also includes partitioning of the data to create the training, validation, and test samples.

This phase is pointed out as being optional.

2. **Explore**

In this step, the data is explored by searching for unexpected trends and anomalies to gain an understanding of the data and draw conclusions and ideas. As stated by Colleen McCue in [McCue \(2015\)](#), this can be done with the help of data visualization, but if the visualizations do not show clear trends, statistical analysis can be used instead.

3. **Modify**

This step builds on the data exploration (the previous step). It consists in modifying the data by creating, selecting and transforming the variables to focus the model selection process. This stage may also look for outliers and the number of variables to be reduced.

4. **Model**

In this step, the model is starting to be created. According to [McCue \(2015\)](#), it includes the use of machine learning algorithms to create models of the sample data that can be used to reliably classify unknowns and/or predict outcomes.

5. **Assess**

The final step focuses on the evaluation of the reliability and usefulness of findings from the data mining process and estimates its performance. Therefore, with this evaluation, a decision is made if the model is useful and reliable.

The movement between the different steps is not strict, so you can go both back and forth throughout the project and repeat steps several times before you are satisfied with the result. Also, as explained in the developer documentation - [SAS Institute \(2006\)](#) -, it isn't mandatory to include all the steps, i.e., you may or may not

include all of the SEMMA steps in your analysis.

Table 2: CRISP-DM vs. KDD vs. SEMMA: Knowledge Discovery Methodologies Comparison

KD Methodology	Pre-Processing	Main-Processing	Post-Processing
CRISP-DM <i>Cross-Industry Standard Process for Data Mining</i>	Business Understanding Data Understanding Data Preparation	Model	Evaluation Deployment
KDD <i>Knowledge Discovery in Database</i>	Selection Pre-Processing Transformation	Data Mining	Interpretation/ Evaluation
SEMMA <i>Sample, Explore, Modify, Model, Assess</i>	Sample Explore Modify	Model	Assess

3.3 DATA SCIENCE

Data Science (DS) is a subject that has aroused growing interest all over the world. Although there is no well-defined definition, it is the current term for the field of study that combines domain expertise, computer science and knowledge of mathematics and statistics to extract meaningful insights from data. As Stanton indicated in his well-known work [Stanton \(2013\)](#), this is an area that is associated to collection, preparation, analysis, visualization, management and preservation of large amount of information.

According to Longbing Cao in [Cao \(2017\)](#), Data Science can be defined by the following formula:

$$\text{Data Science} = (\text{statistics} + \text{informatics} + \text{computing} + \text{communication} + \text{sociology} + \text{management}) \mid (\text{data} + \text{environment} + \text{thinking})$$

where "|" means "conditional on", i.e., that all the mentioned sciences act on the basis of data, the environment and the so-called data-to-knowledge-to-wisdom thinking, as explained in [Weihs and Ickstadt \(2018\)](#).

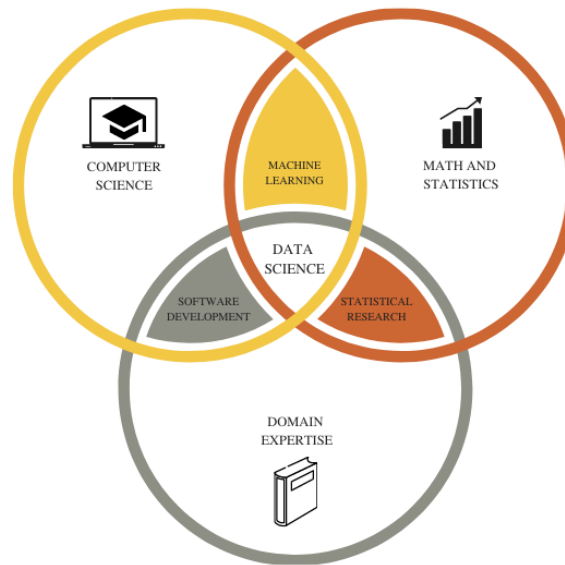


Figure 8: Data Science Areas

Data Science is a complex, multifaceted field that can be approached from several points of view such as ethics, business models, how to deal with big data, data governance, etc. Each point of view becomes a long and interesting discussion, but the approach adopted in this dissertation focuses on analytical techniques, because such techniques constitutes the core toolbox of every data scientist. Also, they are the key ingredients in predicting future events, discovering useful patterns, and probing the world, and, as we will see in this chapter, these are very important points to explore the world using data.

Data scientists apply machine learning algorithms to numbers, text, images and others to produce AI systems to perform tasks that ordinarily require human intelligence. In turn, these systems generate insights that analysts and business users can translate into tangible business value.

The novelty of DS is not rooted in the latest scientific knowledge, but in a disruptive change in our society that has been caused by the evolution of technology: datification.

M. Lynne Markus, in the research [Markus \(2017\)](#) made with Bentley University, defended that datification, also called the heart of DS, is basically the process of rendering into data aspects of the world that have never been quantified before, i.e., the technologies and work practices by which individuals and organizations are sorted and classified, scored and ranked on various dimensions, prescribed or predicted and manipulated. It is about taking a process or activity that was previously invisible and turning it into data. That data can then be tracked, monitored, and optimized, leading to new opportunities — and new challenges.

However, datification is not the only ingredient of the data science revolution. The other is the democratization of data analysis. Some companies used it to take advantage of datification by using analytical techniques to develop innovative products and even to take decisions about their business.

Laura Igual, in [Iguar et al. \(2017\)](#), argues that, in general, DS allows us to adopt four different strategies to explore the world using data:

1. **Probing reality:** There are some cases where the data represents the response of the world to our actions. Analyzing these responses can be extremely valuable in making decisions about how we should proceed. A good example is the use of A/B testing for web development: what is the best button size and color? The best answer can only be found by probing the world;
2. **Pattern discovery:** Dated problems can be analyzed automatically to discover useful patterns and natural clusters that can greatly simplify their solutions. The use of this technique to profile user is a critical ingredient today in such important fields as programmatic advertising or digital marketing;
3. **Predicting future events:** Since the dawn of statistics, one of the most important scientific questions has been how to build robust data models that are capable of predicting future data samples. Predictive analytics enables decisions to be made in response to future events, not just reactively. Obviously, there is no way of predicting the future in any environment and there will always be unpredictable events, but the identification of predictable events represents valuable knowledge. For example, predictive analytics can be used to optimize the tasks planned for retail store staff during the following week, by analyzing data such as weather, historic sales, traffic conditions, and so on;
4. **Understanding people and the world:** This is an objective that at the moment is beyond the scope of most companies and people, but large corporations and governments are investing considerable sums in research areas such as understanding natural language, computer vision, psychology and neuroscience. The scientific understanding of these areas is important for data science, because in order to make optimal decisions, it is ultimately necessary to know the real processes that determine decisions and behavior of people.

Traditionally, the data that we had was mostly structured and small in size, which could be analyzed by using simple BI tools. Unlike data in the traditional systems, today most of the data is unstructured or semi-structured. This data is generated from different sources. Simple BI tools are not able to process this huge volume and variety of data. This is why we need more complex and advanced analysis tools and algorithms for processing, analyzing and extracting acknowledgments.

According to [Baier et al. \(2005\)](#) the greatest challenge for Data Science is finding out what available evidence are useful for the task.

3.3.1 *Statistics*

Statistical analysis is the process of generating statistics from stored data and analyzing the results to deduce or infer meaning about the underlying dataset or the reality it is trying to describe.

Statistics can be defined as the science of data. It includes collecting, classifying, summarizing, organizing, analyzing, and interpreting numerical and categorical information, summed up Darius Singpurwalla in [Singpurwalla \(2013\)](#).

Statistics and machine learning play a central role in data science, as explained in [van Dyk et al. \(2015\)](#), and is defined as the study of the collection, analysis, interpretation, presentation, and organization of data. If we strictly observe this definition, we realize that it is not far from the definition of Data Science.

Bruce Ratner, in [Ratner \(2017\)](#), developed a study to answer the question: "Are Statistics and Data Science Different?". Ratner proposed two hypotheses based on 19 perspectives from several important studies in the area:

H0: Data Science and Statistics are identical ($p = p_0$)

H1: Data Science and Statistics are not identical ($p \neq p_0$)

According to his research, 8 scientists consider that H0 is the most correct hypothesis, while 11 consider that H1 is the correct hypothesis.

When observing these results, it is not possible to draw a solid conclusion, since these results are due, in large part, to a gap in the coherence of the definition of the various areas. What we can tell is that the areas intersect a lot with each other and, therefore, they are fundamental in the existence of each other, promoting a symbiosis relationship.

According to [Weihs and Ickstadt \(2018\)](#) and other researchers, statistical methods are essential to finding structure in data and making predictions since they are able to handle many different analytical tasks.

3.3.1.1 *Statistics Applications*

Statistics is used to solve complex problems so that they can be analyzed, looking for trends and significant changes in the data. In simple words, statistics can be used to derive meaningful insights from data by performing mathematical computations on it, as said in [Lateef \(2020\)](#).

Statistical analysis in Data Science may be used to a lot of things, mainly:

- Present key findings revealed by a dataset;
- Summarize information;
- Calculate measures of cohesiveness, relevance or diversity in data;
- Make future predictions based on previously recorded data;
- Test experimental predictions.

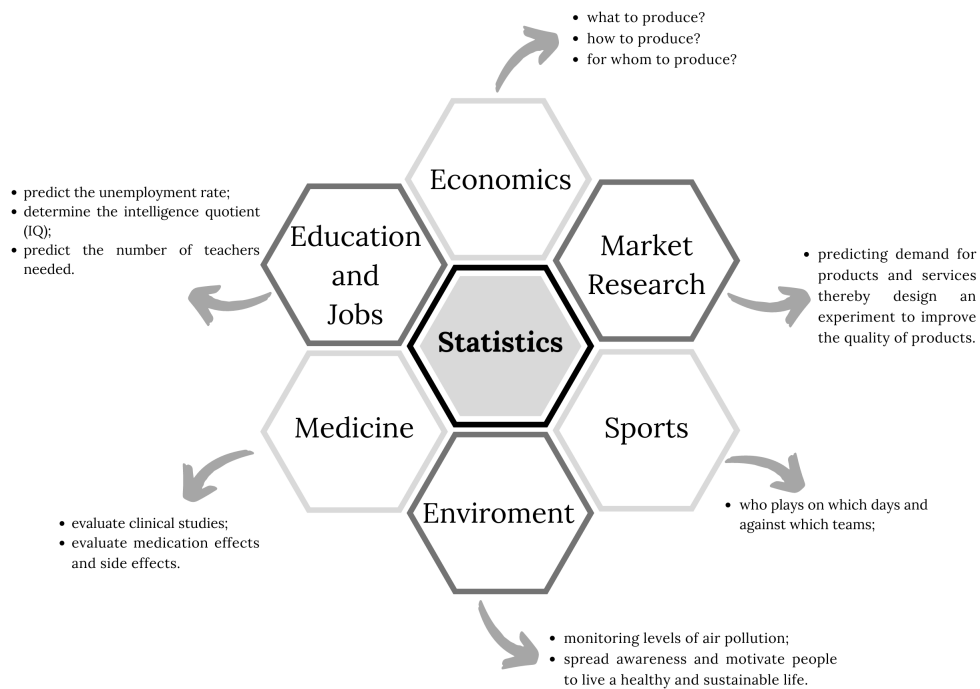


Figure 9: Statistics Applications Worldwide

3.3.1.2 Types of Statistics

According to Darius Singpurwalla in [Singpurwalla \(2013\)](#), there are two types of statistics that are often referred to when making a statistical decision or working on a statistical problem:

- **Descriptive Statistics**

Utilizes numerical (e.g.: mean, median, ...) and graphical (e.g.: pie charts, bar charts, ...) methods to explore data, i.e., to look for patterns in a data set, to summarize the information revealed in a data set, and to present the information in a convenient form, making possible to make decisions. [Iguar et al. \(2017\)](#) also defends that it helps to simplify large amount of data in a sensible way. In this type of statistics we do not draw conclusions beyond the data we are analyzing; neither do we reach any conclusions regarding hypothesis we may make. We do not try to infer characteristics of the population of the data, but claim to present quantitative descriptions of it in a manageable form. The main goal is to describe a data set.

It is based on two main concepts:

1. a **population** is a collection of objects, items or units about which information is sought;
2. a **sample** is a part of the population that is observed.

- **Inferential Statistics**

Utilizes sample data to make estimates decisions, predictions, or other generalizations about a larger set

of data (e.g.: z-statistics, t-statistics, ...). Inferential statistics generalizes a large data set and applies probability to arrive at a conclusion. It allows you to infer parameters of the population based on sample stats and build models on it. The main goal is to make a conclusion about a population based off of a sample of data from that population.

However, there is not only one way to address the problem of statistical inference. In fact, there are two main approaches to statistical inference:

- **Frequentist approach:** the main assumption is that there is a population, which can be represented by several parameters, from which can obtain numerous random samples. Population parameters are fixed but they are not accessible to the observer. Thus, frequentist methods regard the population value as a fixed, unvarying (but unknown) quantity, without a probability distribution. The only way to derive information about these parameters is to take a sample of the population, to compute the parameters of the sample, and to use statistical inference techniques to make probable propositions regarding population parameters.
- **Bayesian approach:** bayesian methods are based on the idea that unknown quantities, such as population means and proportions, have probability distributions. The probability distribution for a population proportion expresses our prior knowledge or belief about it, before we add the knowledge which comes from our data. That said, it is based on a consideration that data are fixed, not the result of a repeatable sampling process, but the parameters describing data can be described probabilistically.

J. Martin Bland defended in [Bland and Altman \(1998\)](#) that the major difficulty is deciding on the prior distribution. This is going to influence the conclusions of the study, yet it may be a subjective synthesis of the available information, so the same data analysed by different investigators could lead to different conclusions.

Table 3: Descriptive vs Inferential Statistics Elements

Types of Statistics		
	Descriptive	Inferential
1)	define the population of interest	define the population of interest
2)	select the variables that are going to be investigated	select the variables that are going to be investigated
3)	select the tables, graphs, or numerical summary tools	select a sample of the population units
4)	identify patterns in the data	run the statistical test on the sample
5)	Generalize the result to your population and draw conclusions	

3.3.1.3 Types of Data

Conforming to Devin Pickell in [Pickell \(2019\)](#), there are two types of data in statistical analysis:

- **Qualitative**

Qualitative data is descriptive and conceptual. This data cannot be measured on numerical scale. Instead, it is categorized based on properties, attributes, labels, and other identifiers;

- **Quantitative**

Quantitative data can be counted, measured, and expressed using numbers.

3.3.2 Machine Learning

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that focuses on programming computers to learn from data and seeks to answer the question: "How do you create a computer system that automatically improves through experience?"

ML is all about using the right features to build the right models that achieve the right tasks.

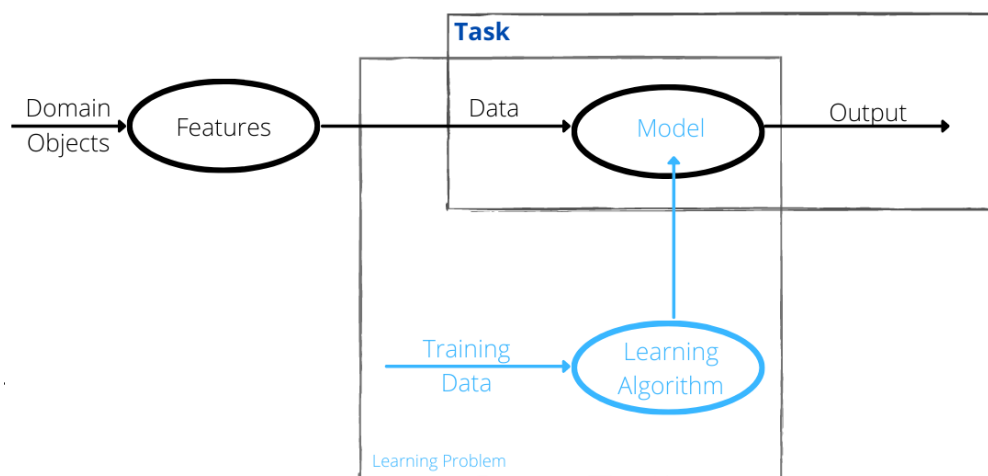


Figure 10: The main ingredients of ML

As Peter Flach in his work [Flach \(2012\)](#) argued, in resume, features define a 'language' in which we describe the relevant objects in our domain; a task is an abstract representation of a problem we want to solve regarding those domain objects; the model is produced as the output of a machine learning algorithm applied to training data.

Each machine learning problem can be precisely defined, according to [Mitchell \(2017\)](#), as the problem of improving some measure of performance P when executing some task T , through some type of training experience E . Once the three components $\langle T, P, E \rangle$ have been specified fully, the learning problem is well defined.

3.3.2.1 *Machine Learning Lifecycle*

As Ritvik Voleti in Voleti (2020) said, *Data is equal to Knowledge*. Data is the vital "organ" of any ML application, so that's where the entire ML life cycle starts. There are many researchers in this field, such as Sean Kandel - Kandel et al. (2011) -, and Rob Ashmore - Ashmore et al. (2019). According to their research, data management is represented here as a combination of several data-related steps. This phase is responsible for collecting data from the various sources, as well as integrating them. This also encompasses data analysis and pre-processing and the Data Wrangling process, which aims to make the data usable. It is necessary to resort to cleaning and transform the data in a suitable format to make them more suitable for analysis in the next step. It is one of the most important steps in the whole process, since without it all the next work may be in vain and the algorithm will not arrive at a reliable result. This phase produces the training dataset and testing dataset, i.e., the datasets that are used to train and verify the model, respectively.

The next two phases are related to the model to be used. It starts with choosing the model based on the nature of the problem and the volume and structure of the training data. A loss function is constructed as a measure of training error. The purpose of the training being performed is to create a model that minimizes the error. When the resulting ML model obtains satisfactory results for the context in question, a phase of verification of the model comes. The key challenge in this phase is to ensure that the trained model works well with new, previously unseen inputs.

After these phases, we receive the result of the review. If the result is good and contains all the necessary requirements, the Model Deployment phase begins.

The Model Deployment phase includes activities related to integrating the successfully verified model with other developed and verified components. The outcome of the Model Deployment stage is a fully-fledged deployed and operating system, as described in Ashmore et al. (2019).

3.3.2.2 *Supervised and Unsupervised Learning*

Mariette Awad argued in her work Awad and Khanna (2015) that we can classify machine learning systems according to the type and amount of human supervision during the training.

1. **Supervised Learning**

As stated by some authors like Gron (2017); James et al. (2013); Sathya and Abraham (2013), supervised learning is a learning mechanism that infers the underlying relationship between the input data and a target variable (a dependent variable or label) that is subject to prediction. It is based on training a data sample from data source with correct classification already assigned. For each observation of the predictor measurement(s) $x_i, i = 1, \dots, n$ there is an associated response measurement y_i . The goal is to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference). So, that said, supervised algorithms use datasets that contain a label y for each

example x and these algorithms try to correlate the features with a given label by estimating $p(y|x)$, i.e., the conditional probability of y given x .

Also, Glenn J. Myatt and Wayne P. Johnson, in their research work [Myatt and Johnson \(2009\)](#), pointed that the two typical tasks of supervised learning are classification and regression. While in a classification problem a model is built to predict a categorical variable (e.g.: filtering spam: classifying emails as spam or not, leading the model to learn how to classify new emails); a regression problem (also called estimation, forecasting or prediction) refers to building models that generate an estimation/prediction for a continuous variable given a set of features (predictors) (e.g.: predicting the sales for a given quarter).

The **Random Forest Algorithm** and the **Linear Regression Algorithm** are explained in [3.3.2.3](#) and [3.3.2.4](#) as examples of supervised learning.

2. Unsupervised Learning

Based on Aurlien Gron and Gareth James works [Gron \(2017\)](#); [James et al. \(2013\)](#), Unsupervised Learning means there's no supervision or guidance and unsupervised algorithms don't have access to a label. It's often thought of as having no correct answers, just acceptable ones. For every observation $i = 1, \dots, n$, we observe a vector of measurements x_i but no associated response y_i . It is not possible to fit a linear regression model, since there is no response variable to predict. Thus, these algorithms attempt to estimate $p(x)$.

Visualization algorithms are a good examples of unsupervised learning algorithms: you feed them a lot of complex and unlabeled data, and they output a 2D or 3D representation of your data that can easily be plotted. Another important unsupervised task is anomaly detection. The system is trained with normal instances, and when it sees a new instance it can tell whether it looks like a normal one or whether it is likely an anomaly.

[Iguar et al. \(2017\)](#) also defends that the problem of unsupervised learning is trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate the goodness of a potential solution. This distinguishes unsupervised from supervised learning.

- **Clustering**

Clustering is a process of grouping similar objects together, i.e., to partition unlabeled examples into disjoint subsets of clusters, such that:

- Examples within a cluster are similar;
- Examples in different clusters are different.

3. Reinforcement Learning

Reinforcement learning (RL) methodology involves exploration of an adaptive sequence of actions or behaviors by an intelligent agent (RL-agent) in a given environment with a motivation to maximize the cumulative reward. In general, this methodology can be viewed as a control-theoretic trial-and-error learning

paradigm with rewards and punishments associated with a sequence of actions. The RL-agent changes its policy based on the collective experience and consequent rewards. RL seeks past actions it explored that resulted in rewards.

RL algorithms are those that learn via reinforcement from criticism that provides information on the quality of a solution, but not on how to improve it.

James et al. (2013) also defended that variables can be characterized as either quantitative or qualitative (also known as categorical). We tend to refer to problems with a quantitative response as regression problems, while those involving a qualitative response are often referred to as classification problems. However, some regression algorithms can be used for classification as well, and vice versa.

3.3.2.3 *Random Forest*

There are many learning models in the ML literature. In this dissertation we are going to focus on Random Forest, based on Iguar et al. (2017).

As the name suggests, Random Forest (RF) consists of a large number of individual decision trees that operate as an ensemble, so RF is an ensemble technique. Ensemble techniques rely on combining several base models in order to produce one optimal predictive model using some aggregation technique, such as majority voting. These techniques usually have good properties for combating overfitting. The aggregation of models using a voting technique reduces the variance of the final model. This increases the robustness of the model and usually achieves a very good performance. The forest mechanism is versatile enough to deal with both supervised classification and regression tasks. A critical issue is that the combination to be successful, the errors made by the members of the ensemble should be as uncorrelated as possible.

Random Forest Steps

First of all, it is clear that, to make use of the RF algorithm, we want to categorize a sample. For example, whether a patient has heart disease or not.

1. **Create a bootstrapped dataset.** To create a bootstrapped dataset that is the same size as the original, we just randomly select samples from the original dataset.

The important detail is that we are allowed to pick the same sample more than once.

2. **Create a Decision Tree using the bootstrapped dataset,** but only use a random subset of variables (or columns) at each step. Thus, instead of considering all variables to figure out how to split the root node, we randomly select a subset.

Basically, each sample of the bootstrapped dataset is going to create various decision trees, depending on the number of variables we chose.

Using a bootstrapped sample and considering only a subset of the variables at each step results in a wide variety of trees. The variety is what makes Random Forests more effective than individual decision trees.

3. Now, our patient data is going to be used. After running the data down all of the trees in the RF, we **see which option received more votes**, and draw a conclusion.

Heart Disease	
Yes	No
5	1

Figure 11: As an example, suppose we want to know whether a patient has heart problems or not. From the mentioned algorithm, the data are run by all the trees that were previously created and, each one of them gives a result. In the end, the results are aggregated, as shown in the figure, and the result with the most votes is the "winner". In this case, "Yes" received the most votes, so we will conclude that this patient has heart disease.

Bootstrapping the data plus using the **aggregate** to make a decision is called **Bagging**.

4. **Estimate the accuracy of a RF.** Typically, about $\frac{1}{3}$ of the original data does not end up in the bootstrapped dataset. This is called the Out-of-Bag dataset.

Since the Out-of-Bag dataset was not used to create our trees in the RF, we can now run it through and see if it correctly classifies the samples. So, in summary, this dataset is also run by the decision trees of our RF algorithm and, in the end, the vote takes place to reach a conclusion.

Ultimately, we can measure how accurate our RF is by the proportion of Out-of-Bag samples that were correctly classified by the RF. The proportion of Out-of-Bag samples that were incorrectly classified is "Out-of-Bag error".

5. **Achieving better results.** Now, we can compare the Out-of-Bag error for an RF built using different variable numbers per step, and we test a bunch of different settings and choose the most accurate RF. Typically, we start by using the square of the numbers of variables and then try a few settings above and below that value.

Decision Trees

We can not talk about RF without talk about Decision Trees. A decision tree is one of the most simple and intuitive techniques in ML, based on the "divide and conquer" paradigm. A decision tree determines the predictive value based on series of questions and conditions. The basic idea behind them in to partition the space into patches and to fit a model to a patch. There are two questions in order to implement this solution:

1. **How do we partition the space?**

There are different strategies for creating decision trees. Most techniques share the axis-orthogonal hyperplane partition policy, i.e., a threshold in a single feature.

2. What model shall we use for each patch?

Each patch is given the value of a label (e.g.: the majority label), and all data falling in that part of the space will be predicted as such.

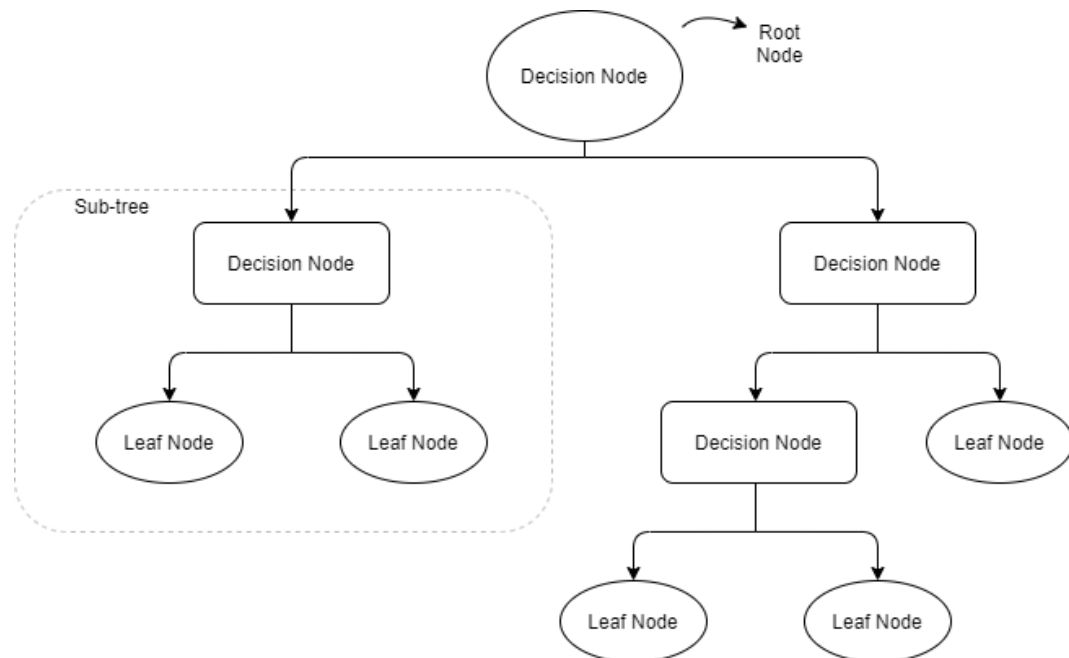


Figure 12: General Decision Tree

So, decision trees are easy to build, easy to use and easy to interpret, but, quoting [Hastie et al. \(2001\)](#), *trees have one aspect that prevents them from being the ideal tool for predictive learning, namely inaccuracy. They seldom provide predictive accuracy comparable to the best that can be achieved with the data at hand.* In other words, they work great with the data used to create them, but they are not flexible when it comes to classifying new samples. RF combine the simplicity of decision trees with flexibility resulting in a vast improvement in accuracy.

RF technique creates different trees over the same training dataset. The word "random" in Random Forest refers to the fact that only a subset of features is available to each of the trees in its build process. The most important parameters in this technique are the number of trees in the ensemble and the number of features each tree is allowed to check.

3.3.2.4 Linear Regression

Linear Regression (LR) is used when we want to predict the value of a variable (independent variable) based on the value of another variable (dependent variable). As stated by [Su et al. \(2012\)](#), it provides the simplest model form to model the regression function as a linear combination of predictors and the models parameters are easily interpretable.

LR helps to determine if an independent variable does a good job in predicting the dependent variable or which independent variable plays a significant role in predicting the dependent variable.

According to [Wrong \(2020\)](#), when performing a simple linear regression there are four main components:

- **Dependent Variable:** target variable that will be estimated and predicted;
- **Independent Variable:** predictor variable that is used to estimate and predict;
- **Slope:** angle of the line, denoted as m or β_1 ;
- **Intercept:** where the function crosses the y-axis, denoted as c or β_0 .

And the formula can be described as:

$$y = \beta_0 + \beta_1 X + \epsilon$$

Linear regression finds the line of best fit line through your data by searching for the regression coefficient (β_1) that minimizes the total error (ϵ) of the model.

3.3.2.5 Machine Learning in Medicine

Machine learning is ubiquitous and essential for solving complex problems in most sciences. Essentially, from an initial training dataset of features and outcomes, an algorithm learns how the features relate to and are predictive of the outcomes.

As in other areas, ML techniques have already partially revolutionized medicine and are showing no signs of slowing down. The advances that have already been made would have been inconceivable without machine learning to process high-resolution physiological data in real time.

ML excels particularly in identifying patterns in large and noisy data sets, which makes it useful for analyzing complex biological data. However, machine learning algorithms are very "data hungry" and often require millions of observations to achieve an acceptable level of performance.

According to [Obermeyer and Emanuel \(2016\)](#), the ability to transform data into knowledge will disrupt at least three areas of medicine.

First, machine learning will dramatically improve prognosis, in consonance with [Rahul C. Deo and Chloe Gui's](#) analysis in [Deo \(2015\)](#) and [Gui and Chan \(2017\)](#), respectively.

Prognosis prediction includes approximate outcomes such as a patient's susceptibility to disease, the likelihood of disease recurrence, life expectancy, and response to treatment. The factors involved are complex

and a definitive prognosis for many diseases is difficult. Accurate prognosis forecasting is valuable as it helps healthcare providers make informed decisions about resource allocation and best treatment practices.

Oncology in particular deals intensively with the use of ML methods to predict the prognosis. One review reported that ML improves cancer prognostic predictions by 15-25%, according to [Kourou et al. \(2014\)](#).

Second, pursuant to [Wang and Summers \(2012\)](#); [Gui and Chan \(2017\)](#), machine learning will replace much of the work of radiologists and anatomical pathologists.

ML already forms the basis of radiology tools such as image segmentation to isolate areas of interest. However, increased ML skills can oust physicians from some of their roles. These physicians mainly focus on interpreting digitized images, which can instead be fed directly into algorithms without any problems. Massive image data sets combined with recent advances in computer vision will cause rapid increases in performance, and machine accuracy will soon surpass that of humans.

ML would ease the burden and provide consistent 24-hour service while human radiologists can make mistakes in circumstances like night shifts. Besides, physicians of other image-based specialties, such as pathology, may also perform fewer image analysis tasks as ML algorithms improve.

And **third**, machine learning will improve diagnostic accuracy. A recent Institute of Medicine (IOM) report - [Monegain \(2015\)](#) - drew attention to the alarming frequency of diagnostic errors and the lack of interventions to reduce them.

ML has been demonstrated to be capable of screening patients, stratifying patients by risk, and assisting physicians in decision making. Screening models have been built to detect diseases such as congenital cataracts, skin cancer, heart disease, hepatitis disease, and autism. Given the high stakes of medical decisions, a model built with particularly high sensitivity could be an inexpensive tool to rule out diagnoses, leaving potentially positive cases for physicians to investigate.

Also, screening could be particularly advantageous for rare diseases. Rare diseases are typically difficult to identify, and this may lead to delayed or incorrect treatment, possibly with harmful consequences to the patient. Thus, [Long et al. \(2017\)](#) examined whether AI could provide a unique management system for rare diseases. They built ML models to screen patients for congenital cataracts, perform risk stratification, and suggest treatment.

In resume, ML is a versatile and powerful tool that potentiates personalized medicine, providing a more precise understanding of individual patients and their needs.

3.3.2.6 *Machine Learning in Finance*

Finance, technology, and data analytics have long existed in symbiosis. Artificial Intelligence and more specifically machine learning are just some of the most recent innovations in data analytics to be leveraged by the financial sector, as [Gensler and Bailey \(2020\)](#) claimed.

Recently, [Emerson et al. \(2019\)](#) mentioned that there has been a proliferation of ML techniques and growing interest in their applications in finance, where they have been applied to sentiment analysis of news, trend analysis, portfolio optimization, risk modelling among many use cases supporting investment management.

As [Dixon and Halperin \(2019\)](#) stated, finance has all the ingredients needed to make ML work. In general, the more data you enter, the more accurate the results will be. Coincidentally, huge data sets are widespread in the financial services industry. There are a lot of data on transactions, customers, invoices, money transfers, and others. That is a perfect fit for machine learning and that's why so many financial companies invest heavily in machine learning R&D.

We can also see that the future of financial services is hard to imagine without machine learning as technology evolves and the best algorithms are open source.

Konstantin Didur summarized in [Didur \(2018\)](#) that, thanks to the quantitative nature of finance and the large amount of historical data, ML is poised to improve many aspects of the financial ecosystem.

Principles for Effective Machine Learning in Finance

Although it is important to note that there is no possible substitution for human judgment and experience in the case, in line with [Dixon and Halperin \(2019\)](#), we can follow some guidelines for improving the success of machine learning in investment management:

1. **Problem definition:** Define the problem that is being solved;
2. **Modeling assumptions:** State the main assumptions of the modeling approach. Does the machine learning method assume that the observations are identical and independently distributed? If the data is a time series, does the model assume stationarity?;
3. **Develop intuition:** Establish a toy model, a subset of the full problem, in order to gain comfort with machine learning in a setting where less can go wrong;
4. **Defensible results:** Design the experiment so that the model output can be explained and is defensible by reduction ad absurdum²;
5. **Diagnostics:** Employ several diagnostics, both from machine learning and statistics, to characterize the data;
6. **Keep it Simple:** Minimize the use of elaborate algorithms and widgets, in favor of approaches that rely on as few parameters and are as transparent as possible;
7. **Choice of utility function:** Ahead of analyzing the performance of specific algorithms, review the appropriateness of your assumption about a loss function you use to train your machine learning algorithms;
8. **Solution constraints:** Analyze all possible prior views or constraints that the expected solution should satisfy and enforce these prior views on the solution;

² Reductio ad absurdum is a mode of argumentation that seeks to establish a contention by deriving an absurdity from its denial, thus arguing that a thesis must be accepted because its rejection would be untenable.

- 9. **Is your data biased?:** In many cases, datasets available for training machine learning algorithms are misbalanced. A misbalance of data may propagate into biases of your predictive model;
- 10. **Feature Engineering:** If possible, avoid the temptation to prematurely feature engineer, switch between various machine learning models or live simulation environment.

Like these, there are many other viable ML applications in the industry. Therefore, and following what was found in this section, we easily realize the importance of trying to make the most of ML in monitoring the financial execution of UMinho's R&D projects.

3.4 SCIENTIFIC PRODUCTION

As already mentioned, research is an area where a lot of money is invested. The knowledge dissemination and the publication of scientific results is an essential part of scientific research and, according to Ugolini et al. (2012), this production has increased significantly. Scientific production is carried out based on research, so it is highly dependent on the latter. In turn, as said in Gulbrandsen and Smeby (2005), research has very specific needs depending on its focus such as obtaining specific equipment and material or hiring staff, which is often quite expensive. Therefore, it is important to cross the scientific production and the financial execution of R&D projects since they are so connected with each other.

According to the Scopus database, scientific production in Portugal has been increasing over the years, which reveals, once again, the importance that this theme has been acquiring.

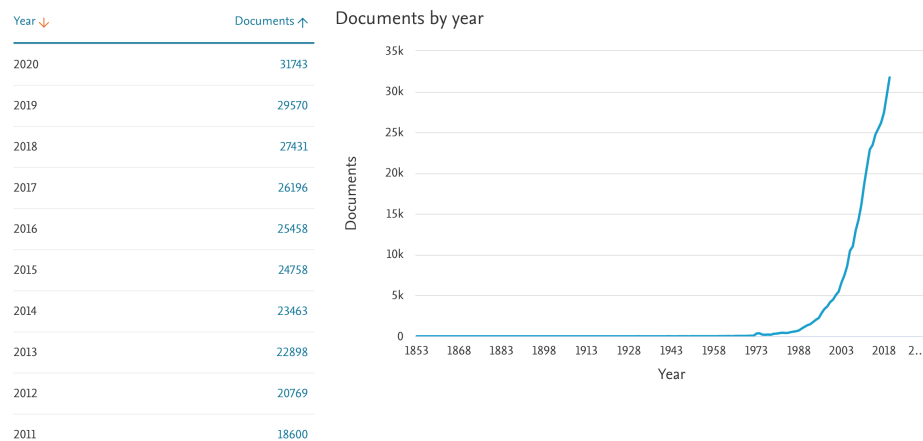


Figure 13: Evolution of scientific production indexed in the Scopus database

Richard Van Noorden, a well-known journalist specializing in science journalism, proposed the following question in Noorden (2014): "How fast is the mass of scientific production really increasing?" In the same paper, he stated that, according to some bibliometric analysts, *it is impossible to know for sure, but the real rate is closer to 8-9% per year*. This means that **global scientific production doubles every nine years**.

Despite this great development in the quantity of scientific production, evaluating its quality is a much more complex problem and, according to Seglen (1997), has no standard solution. Ideally, scientific work should be evaluated by real experts in the field according to universal rules. In practice, however, the assessment is carried out by the committees, which, despite having a solid knowledge of the field, are usually not experts. For this reason, it is not surprising that other methods of assessing the quality of scientific work are being developed.

Therefore, as Montoya et al. (2018) claimed, the evaluation of these works is often based on bibliometric indicators provided by certain information sources (such as Scopus, Web Of Science or Google Scholar), which not only contain several indicators for quantity and quality, but also facilitate the search and grouping of information (by author, journal, affiliation and others). According to Merediz-Solà and Bariviera (2019), these methods have become an emergent and buoyant discipline and research works like Larivière et al. (2006) show that they are increasingly used in research evaluation due to concerns about public spending in science.

Much of the scientific world has accepted the Impact Factor (IF) as the standard quality indicator for evaluating scientific work. However, in consonance with Ramin and Shirazi (2012); Seglen (1997), it has come under some criticism: among other things, poor quality assessment of the citations, influence of self-citation, bias of the English language. Furthermore, some researchers such as Salimi (2017) reported that not all metrics rate an area equally, that is, some metrics rate some areas more positively than others.

Actually, Jonas Lundberg in Lundberg (2007) affirmed that there are *differences in average citation rates for publications of different types, of different age, published in journals within different fields*. In other words, in agreement with Guerrero-Bote and Moya-Anegón (2012), not all papers or journals have the same quality values. And this type of metric is usually lower in the areas of Engineering, Social Sciences, and Humanities as reported by Guerrero-Bote et al. (2007); Lancho-Barrantes et al. (2010a). Some authors as Lancho-Barrantes et al. (2010b) have also expressed concern about the applicability of certain indicators to certain scientific areas. In addition, citation indexes like WoS or Scopus concentrate mainly on journals and less on books, proceedings or reports. So, in line with Larivière et al. (2006); Mongeon and Paul-Hus (2016), this is a problem because there are scientific areas like Social Sciences and Arts and Humanities where the publication of books is more common and has great importance for researchers' careers, even more than publishing papers.

3.4.1 Scientific Databases

In 2000, M. K. Michener and J. W. Brunt Michener and Brunt (2009), defined a scientific database as a *computerized collection of related data organized so as to be accessible for scientific inquiry and long-term stewardship*.

Nowadays there are several databases designed to store scientific work, that is, scientific databases. Some of them are open access, while others only allow access by organizations that make a subscription, as Montoya et al. (2018) claimed. There are also some databases that only focus on one scientific area, such as PubMed. Since there are many sources of information available, some studies such as Bar-Ilan (2008); Kulkarni et al. (2009); Levine-Clark and Gil (2008) have been done to compare them.

Until 2004, according to Bar-Ilan (2008), the Web of Science was the only comprehensive database that provided citation data, but that changed with the introduction of Scopus and Google Scholar.

These sources of information, while used to store and organize scientific studies, also began to recognize that there was an increasing need to classify these works qualitatively. Therefore, they developed their own performance metrics.

The correspondence base used for this work is Scopus since it is the database used by UMinho to manage the quantity and quality of its scientific production. Not only it is the scientific database that contains most of the journals in which UMinho articles are published, it is also the most reliable information, along with the Web of Science [Montoya et al. \(2018\)](#). Furthermore, the Scopus Database API enables us to search and collect data very easily, as it is accessible to the public and enables the collection of raw data according to specific criteria.

Thus, for a classification of the scientific production quality by area, the 27 scientific areas listed in Table 4 described in Scopus were considered.

Scientific Area	
1. Agricultural and Biological Sciences	2. Arts and Humanities
3. Biochemistry, Genetics and Molecular Biology	4. Business, Management and Accounting
5. Chemical Engineering	6. Chemistry
7. Computer Science	8. Decision Sciences
9. Dentistry	10. Earth and Planetary Sciences
11. Economics, Econometrics and Finance	12. Energy
13. Engineering	14. Environmental Science
15. Health Professions	16. Immunology and Microbiology
17. Materials Science	18. Mathematics
19. Medicine	20. Multidisciplinary
21. Neuroscience	22. Nursing
23. Pharmacology, Toxicology and Pharmaceutics	24. Physics and Astronomy
25. Psychology	26. Social Sciences
27. Veterinary	

Table 4: 27 Scientific Areas described in Scopus

3.4.2 Bibliometric Indicators

According to [Ugolini et al. \(2012\)](#), bibliometric studies are systematically carried out to assess the relative importance of scientific production in a given scientific area. The Impact Factor (IF) is one of the most commonly used measures of quality, but we have used this and others to get a more general view.

SJR: The SCImago Journal Rank (SJR) indicator expresses the average number of weighted citations received during the selected year from articles published in the selected journal in the last three years, i.e. weighted citations received in year X to documents published in the journal in years $X-1$, $X-2$ and $X-3$;

H-index: The H-index expresses the journal's number of articles (h) that have received at least h citations;

Cites/Document (2 years): Cites per Document is the average citations per document. It is calculated considering the number of citations a journal has received in the current year compared to the documents published in the two previous years, i.e., citations received in year X to documents published in years X-1 and X-2;

IPP: The Impact Per Publication (IPP) is calculated as the number of citations given in the present year to publications in the past three years divided by the total number of publications in the past three years;

SNIP: The Source Normalized Impact per Publication (SNIP), calculated as the number of citations given in the present year to publications in the past three years divided by the total number of publications in the past three years. The difference with IPP is that in the case of SNIP citations are normalized in order to correct for differences in citation practices between scientific fields;

JIF: The Journal Impact Factor (JIF) is calculated as the average of the sum of the citations received in a given year to a journal's previous two years of publications, divided by the sum of "citable" publications in the previous two years;

Eigenfactor: The Eigenfactor score intends to give a measure of how likely a journal is to be used, and is thought to reflect how frequently an average researcher would access content from that journal;

CiteScore: The CiteScore is a measure reflecting the yearly average number of citations to recent papers published in that journal;

There are several indicators from different databases that could be used for this study. However, many data sources are not public, do not have a large part of the papers or have a more complex mass extraction process, which complicates the evaluation task intended here.

DEVELOPMENT

“The best way to predict the future is to create it.”

— Alan Kay

This chapter describes all the development work carried out throughout the dissertation. As previously mentioned, the main focus of this thesis was the construction of a Data Visualization platform that would allow a correct and organized visualization of the financial data of the University of Minho R&D projects. However, as the theme of this master’s thesis is quite wide and with a wide variety of adjacent themes, there was the possibility to explore some of these themes, going beyond the development of the platform. Thus, work was also carried out in the scope of the exploration of scientific production produced with the support of UMinho, as well as its relationship with the financial execution of the university’s research projects. In addition, the prediction of certain variables also proved to be a challenge that was explored in this development.

4.1 DATA VISUALIZATION PLATFORM

According to [Albertin and Amaral \(2010\)](#), there is a high rate of failures and unsuccessful projects, which corroborates the need for organizations to update and improve their management techniques and platforms and highlights the need to create this DV platform.

The creation of this platform required a very detailed survey of requirements from UMinho rectory and interested parties in order to develop a platform capable of responding to the widest needs.

After some meetings with several researchers, directors of Research Centers (RC), presidents of Organic Units (OU), and other competent entities from UMinho, some key pieces were proposed for the correct development of the platform.

In the first phase, the stakeholders expressed a special interest in gathering information on a single page, containing a dashboard very rich in information. However, this interest quickly proved inconceivable and even detrimental to the removal of information from the page by users. The amount of information that the platform needed to host was immense, so the best chance was to spread the information across several pages.

Even so, the most relevant idea was to make a global summary on the home page, that is, the first page to be viewed should contain the user’s summary information. Thus, and according to the level of responsibility

of the user, when performing the login you would be redirected to the page with the most comprehensive and summarized set of information.

As an example, an investigator, after entering his credentials and entering the platform, is redirected to a page that has a summary of the researcher's ongoing projects, where he can see which projects are in charge, what the overall budget is of each one, as well as the available, among others.

A research center director will have access to all the center's projects, but this time, aggregated by percentage of execution. Here, there are four "boxes": the red box where projects with very low execution rates are found; the yellow box with projects with low execution rates; the green box with projects with execution rates as planned; and, finally, the blue box where projects with higher execution rates are located.

Following the same logic, a director of an OU sees all the research centers in that unit. And, finally, the rector has access to the twelve Organic Units that make up UMinho. In the appendix A some images of the platform are presented.

4.1.1 System Architecture

Thus, we can observe the architecture of the platform under development in the figure 14. This development depends completely on the files sent periodically by the Financial and Patrimonial Services Unit, so this architecture starts precisely with its collection. These files are subjected to a rigorous Extract, Transform and Load (ETL) process (e.g.: file merging) so that they can later be uploaded to the database.

Once the database is loaded with the necessary data, the code developed for the creation of the platform will be able to consume the data. Finally, the data visualization platform created from the data provided can be viewed by users.

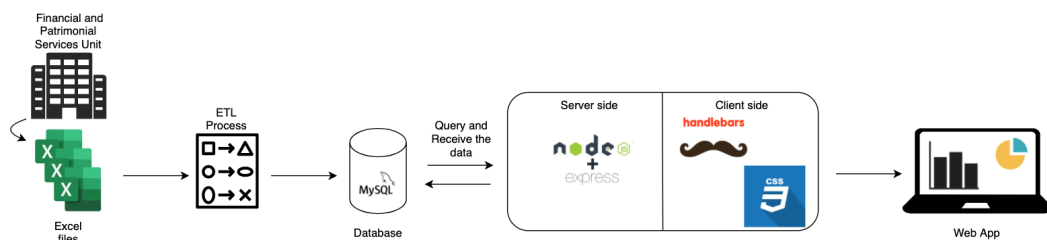


Figure 14: Platform System Architecture

4.1.2 The Dataset

As mentioned, the financial analysis of R&D projects is the core of this platform. This analysis aims, through data already collected, to draw some conclusions that can help their financial management. Also, for a correct analysis of the data, it is necessary, first of all, to understand them.

The area of project management and its financial execution proved to be quite vast, so it was necessary to study what data were in fact relevant and which were not.

For this, let's see what comprises a research project. A research project has some crucial points to understand:

- Has one or more Responsible Researchers. A Responsible Researcher is the lowest level of responsibility that exists when it comes to project management, as well as on the platform mentioned above, but also the most important. This researcher can be identified by his mechanographic number;
- It has two types of budget:
 1. Global budget;
 2. Global budget without overheads, tuition expenses, and human resources belonging to the board.

The difference between these two budgets is in the amount that can, in fact, be spent. While in the first part of it is already destined to expenses such as general expenses, tuition fees, and others, in the second these expenses are already discounted.

It is important to understand this difference so that, later, it is understood which of the two should be considered, depending on the situation;

- Despite having a global budget, this budget is divided according to the duration of the project and your spending needs. Thus, each project has a budget assigned to the date, i.e., a total amount assigned to that project that can be spent for a certain period of time;
- Finally, it has the budget available. This is the budget that really piques researchers;
- However, the amount of a budget cannot be spent anywhere. Each project has a certain number of subcategories, called rubrics, which in turn have a budget. The budget for each rubric is then a part of the overall project budget.

Rubrics are a very important part of projects, as it is through them that the money allocated to a project is conducted and used. The rubrics can be of any type according to the needs of the project (e.g.: general expenses, human resources, purchase of goods and services, etc ...).

Knowing the bases and fundamental parts of a project, it becomes easy to understand the importance of financial analysis and management of R&D projects. For the success of this analysis, it is also important to understand that there are some factors to take into account that depend a lot on the knowledge of the world that surrounds the research projects.

Even so, we know that R&D projects move a lot of money and that, if it is not well oriented to the most diverse needs of researchers and projects, can result in a huge monetary dismissal.

4.2 COVID-19 IMPACT ON R&D PROJECTS

As mentioned, the rubrics can be related to any set of expenses, functioning as groups where the money allocated to the projects can be spent.

One of the most used rubrics is related to missions and scientific dissemination. These expenses are supplemented by all expenses for demonstrating, promoting, and publicizing a project. This item generally has a large share of expenses on travel, accommodation, and food outside the area where the development of the R&D project takes place. These expenses can refer to missions in the country or abroad. In summary, the rubric "Missions/Scientific Production" refers to expenses with missions in the country and abroad directly attributable to the project.

The positive financial execution of a project depends on the execution of the rubrics intended for it. Thus, we can see that if the budget for the different rubrics is not executed, the project will automatically have poor execution.

In 2020, the world saw the emergence of a pandemic caused by the new coronavirus (Covid-19) that devastated the economy of many countries, and disrupted our way of life. The Covid-19 pandemic manifests itself as an acute respiratory disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2; formerly called 2019-nCoV). The first case of infection identified occurred on December 1, 2019 in the city of Wuhan, Hubei province, China.

Worldwide, several borders have been closed to prevent the spread of the pandemic and several confinements and states of emergency have been implemented. As a result, the population was largely restricted to their own country and, often, to their own city.

In order to understand the impact that the pandemic had on the financial execution of R&D projects, we analyzed some projects³. The selected projects have as a requirement to have a large part of their execution in 2020 and to have expenses destined for the "Missions/Scientific Production" rubric.

As explained, each project has its own budget for the rubric in question, so, to carry out a fair analysis and draw plausible conclusions, this study was done based on the percentage of budget already spent on the rubric compared to the budget allocated to it, both globally and for the year 2020⁴.

As we can see in the figures 15 and 16, the execution percentages are very low. However, it is necessary to pay a little more attention to these graphs and observe in detail what they mean.

The first graph represents a comparison between the total budget for missions and the money already spent on missions and the second graph represents a comparison between the budget for missions per year and the money already spent on missions. Although the first graph may be misleading, it is important to have it in this study, since all projects, except project 19, started in 2020. This means that for all other projects, the budget already spent is also the budget spent in the year 2020.

³ Project IDs were hidden to guarantee the necessary confidentiality requested by UMinho.

⁴ The data acquired so far date from November 2020, so it was only possible to carry out the analysis of projects with expenditures up to that point.

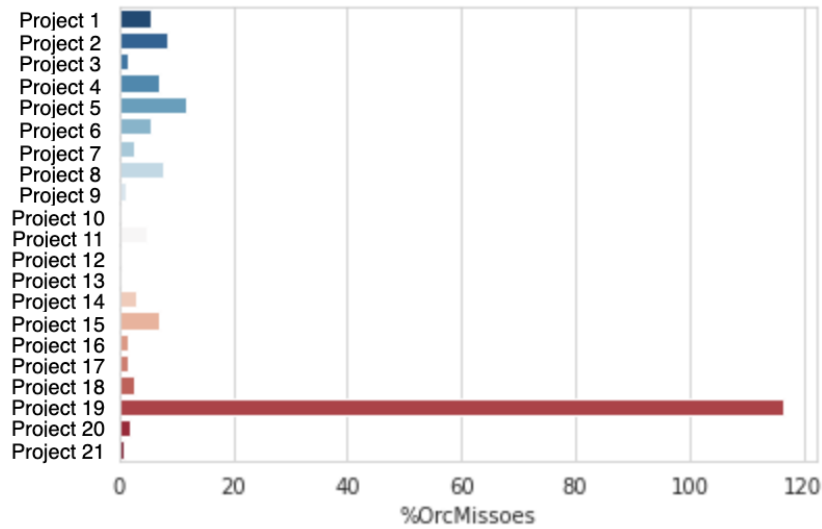


Figure 15: Total percentage of budget spent on missions

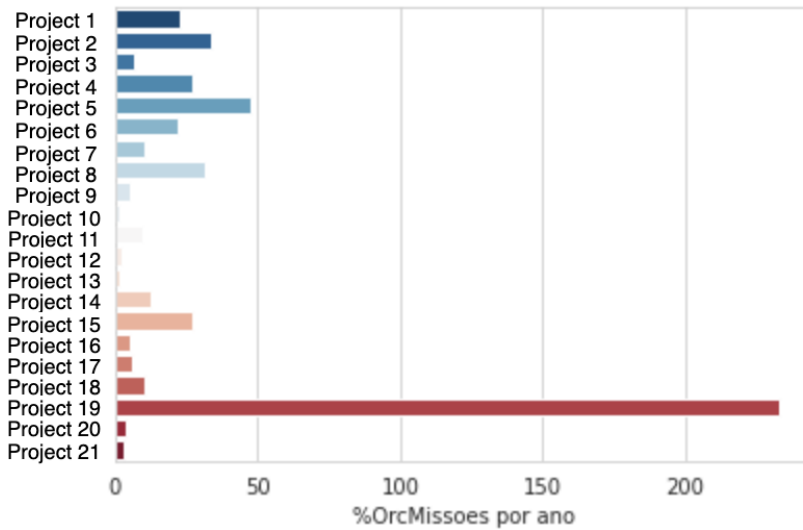


Figure 16: Percentage of budget spent on missions per year

This information becomes even more relevant when we observe that all these projects have 2 or 4 years to be executed. This means that in the first graph, since the data dates from the end of 2020, the percentages should be 25% ($\frac{1}{4}$ of 100%) or 50% ($\frac{1}{2}$ of 100%).

In the table 5 we can learn a little better about the projects under study. The penultimate column refers to money already spent on missions and the last to the total budget allocated to missions.

As we can see, both in the graphs and in the table, only project 19 is at odds with the others. However, it is easy to understand why. This project started on 01/01/2019, that is, at the beginning of the year 2019 and the only rubric it has destined is precisely the rubric of missions. Thus, this is a project with 2 years of execution and had 1 year to make use of its budget before the pandemic started, exclusively for missions.

Project	Start Date	End Date	Number of years	Budget spent on missions	Budget allocated to missions
Project 1	2020-01-01	2023-12-31	4	3 242.73 €	57 856.72 €
Project 2	2020-01-01	2023-12-31	4	1 300.00 €	15 312.00 €
Project 3	2020-01-01	2023-12-31	4	395.00 €	23 880.00 €
Project 4	2020-01-01	2023-12-31	4	1 641.78 €	24 000.00 €
Project 5	2020-01-01	2023-12-31	4	1 118.07 €	9 400.47 €
Project 6	2020-01-01	2023-12-31	4	1 754.60 €	32 000.00 €
Project 7	2020-01-01	2023-12-31	4	825.63 €	32 000.00 €
Project 8	2020-01-01	2023-12-31	4	1 969.94 €	25 000.00 €
Project 9	2020-01-01	2023-12-31	4	1 584.89 €	125 000.00 €
Project 10	2020-01-01	2023-12-31	4	622.88 €	172 800.00 €
Project 11	2020-01-15	2022-01-12	2	125.90 €	2 650.00 €
Project 12	2020-01-01	2023-12-31	4	1 284.76 €	214 738.56 €
Project 13	2020-01-01	2023-12-31	4	50.00 €	12 039.00 €
Project 14	2020-01-01	2023-12-31	4	2 230.86 €	72 000.00 €
Project 15	2020-01-01	2023-12-31	4	1 269.38 €	18 584.64 €
Project 16	2020-01-01	2023-12-31	4	590.00 €	42 953.52 €
Project 17	2020-01-01	2023-12-31	4	313.90 €	20 000.00 €
Project 18	2020-01-01	2023-12-31	4	797.45 €	30 000.00 €
Project 19	2019-01-01	2020-12-31	2	10 472.42 €	9 000.00 €
Project 20	2020-01-15	2022-01-14	2	127.86 €	7 450.00 €
Project 21	2020-01-01	2023-12-31	4	260.00 €	37 894.19 €

Table 5: The projects of this study - projects that have mission expenses

Looking at this project in more detail, we can see that all expenses incurred, which make up the 10 472.42€, were made in 2019, as we can see in table 6. So, it is safe to say that the pandemic did not affect the financial execution of this project, since during 2019 the entire project budget was spent and even exceeded. These exceedances may depend on several factors and even be purposeful, so, when they appear, they should not be seen as a danger that needs to be overcome, but as something natural and should, above all, be studied to rule out any error in the formulation of the selection.

Even so, and although we can therefore say that this project is not the type of project that should be part of this study, it is important that it be represented here to prove that without the pandemic, the financial execution of the item relating to missions would perform quite differently.

We can then say that the pandemic had a major impact on the financial execution of this rubric, as stated in [Alves et al. \(2021c\)](#). Thus, there is also a high probability that this was not the only rubric affected. This small study focused on the missions rubric, but, as we know, numerous rubrics can be attributed to projects and that, consequently, may have seen their execution severely hampered by the COVID-19 pandemic.

Although this is a small starting point for drawing conclusions about this theme, it proves to us that it is a theme that should be explored and even related to others. Thus, within the scope of this thesis, it was also proposed to

Expense Date	Expense Amount
2019-08-05	1 034.04 €
2019-08-05	350.00 €
2019-08-05	1 700.00 €
2019-08-05	850.00 €
2019-08-05	1 700.00 €
2019-08-07	1 850.00 €
2019-08-07	600.00 €
2019-08-07	350.00 €
2019-12-11	1 019.19 €
2019-12-11	1 019.19 €

10 472.42 €

Table 6: Project 19 Expenses

study the relationship between the financial execution of research projects and their scientific production. As we know, the rubric studied in this chapter is closely linked to scientific production, so, starting from this brief study, we will analyze the scientific production produced at UMinho in the next chapter.

4.3 THE RELATIONSHIP BETWEEN FINANCIAL EXECUTION IN R&D AND SCIENTIFIC PRODUCTION

R&D projects and their financial execution cannot be separated from the scientific production carried out. In other words, scientific production and research projects are inextricably linked, so analyzing one topic often involves analyzing the other. With this study we intend to answer the question: "Is there any relationship between the quantity and quality of scientific production and the financial execution of R&D?"⁵.

Each project contains a large amount of financial information. Therefore, in addition to analyzing the global perspective, it is important to analyze the expenditure per rubric in order to get detailed conclusions for the next steps of the study. We can divide sectors into 4 main groups: Equipment/Services, Missions/Scientific Production, Human Resources and Others. The latter group refers to expenses that did not fit into any of the other groups (e.g.: other expenses with project X).

Initially, we carried out a detailed analysis by research center and by organic unit, since it is very important for UMinho that this analysis takes into account the different hierarchical degrees of the university.

In order to match the financial data with the data on the scientific production of UMinho, the Scopus database API (Application Programming Interface) was used a posteriori to automate the search for scientific publications published by UMinho in the last 4 years (2017-2020). This step allowed us to quantitatively evaluate UMinho's scientific production for the years mentioned, but not to evaluate it qualitatively, as this step is used to get

⁵ Based on UMinho's case.

scientific papers grouped as needed and not to get quality measures. To perform this assessment, we used Scopus' SCImago Journal & Country Rank, in which we obtained various metrics for qualitative classification from several journals: SCImago Journal Rank (SJR), H-Index, Cites/Document and others. Then, we have merged the journals that published UMinho papers with the journals indexed in SCImago to get some classification metrics.

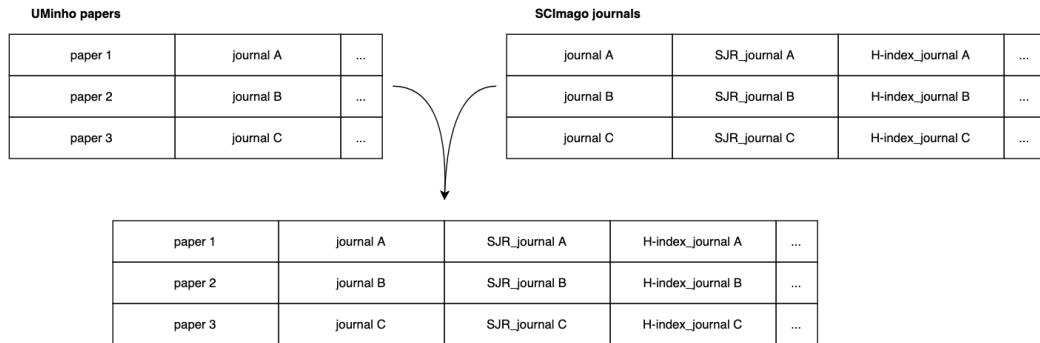


Figure 17: Merge illustration from UMinho papers with bibliometric indicators from Scopus

4.3.1 Financial Execution

In order to examine the expenditures for 2018, 2019 and 2020, the sum of all expenditures per year was made, obtaining the result shown in figure 18. As expected, spending in 2020 was lower than in 2019. However, the point of the question focuses on which rubrics the spending was, in fact, much lower. Thus, the expenditure by rubric was checked.

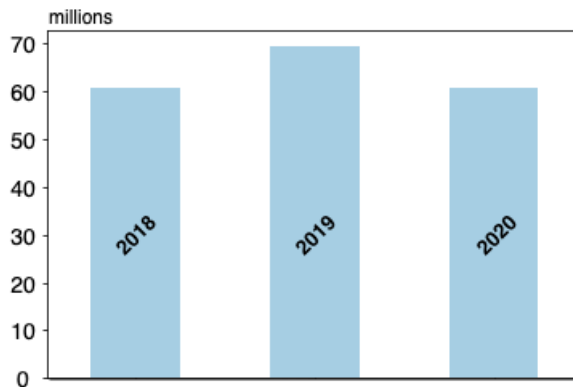


Figure 18: Sum of UMinho's expenditures per year

Although it is not the only one with low execution compared to the past two years, we can notice a significant drop in execution in terms of missions and scientific dissemination, as shown in figure 19. In fact, the financial execution of 2020 under this rubric does not even reach half of the execution of 2018 nor 2019.

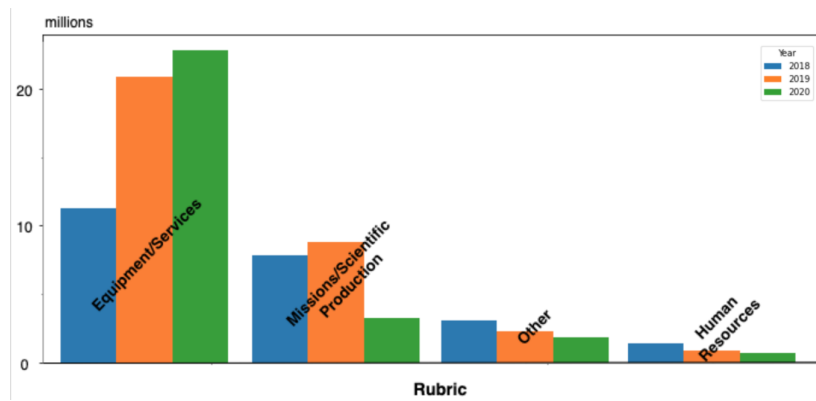


Figure 19: Expenses by rubric

For the University of Minho, the analysis of this expenditure by RC (figure 20) and by OU (figure 21) is also very important in order to analyze financial needs by area. For a better understanding of this analysis, only the 10 centers with the highest expenses are shown. For the aforementioned years, the total expenditure of the other remaining centers (3 453 267 €) does not reach half of the expenditure of these 10 centers (9 311 830 €), so their graphic representation would not be relevant. Even so, the complete analysis can be found in figure 41.

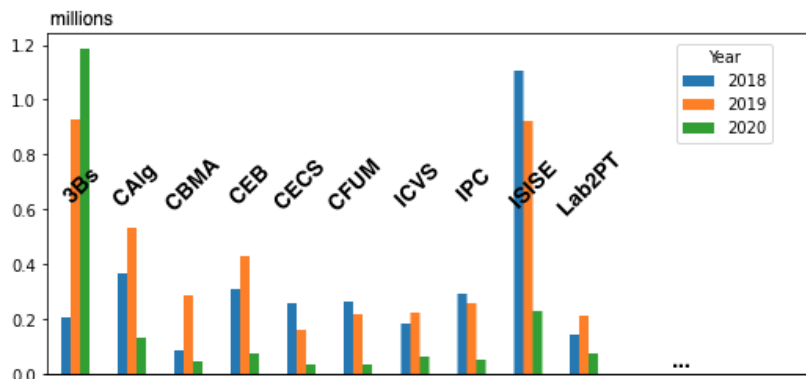


Figure 20: Expenses by Research Center

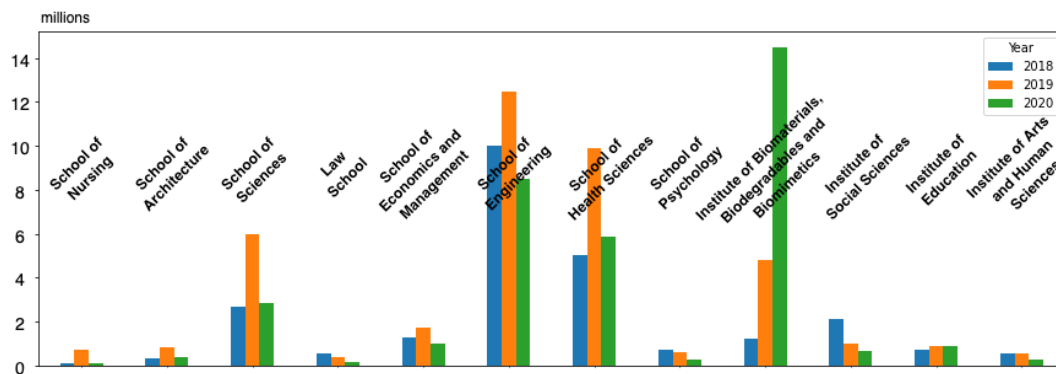


Figure 21: Expenses by Organic Unit

4.3.2 Scientific Production

After observing the financial execution, we move on to observing the scientific production. The bibliometric indicators used are well known. However, the classifications of the various entities that developed these metrics vary from year to year. Therefore, the ratings of the updated journals were collected for each year with the exception of 2020. The ratings for a given year are not published until the middle of the following year, so the ratings for 2020 are only known in June 2021. For this reason, in order to carry out the proposed assessment for the year 2020, the 2019 assessments were also used, as they are the most recent.

Year	Number of Papers	SJR	H-index	Cites/Doc.	Quartile (most freq.)
2020	3060	1,07	85,42	5,39	Q1
2019	3070	0,97	85,91	4,29	Q1
2018	2845	1,07	91,33	4,36	Q1
2017	2664	1,12	87,82	2,69	Q1

Table 7: Quantity and quality of Scientific Production at UMinho using Scopus indicators

The results of SJR, H-index and Cites/Doc. correspond to the average values.

Despite the sharp decrease in the financial execution of the "Missions/Scientific Production" rubric, we note that the qualitative assessment of UMinho's scientific production, according to some indicators, has not been influenced and even improved. Compared to 2019, SJR and Cites per Document increased and the latter even got the highest value in the last 4 years. In addition, the amount of scientific production also remained stable: in 2020 UMinho published 10 fewer articles than in 2019, which does not represent a significant decrease. In other words, the financial execution under this rubric has decreased considerably, but the quantity and quality of scientific production have not. In the next section we analyze the relationship between financial execution and scientific production.

Still regarding the quantity produced, studies were carried out to further explore this theme, grouping the scientific production produced by RC (figure 22) and by OU (figure 23). As with the financial execution analysis, only the 10 centers with the highest scientific production have been considered as the graphing of the remaining would not be relevant. Full information can be found in figure 42.

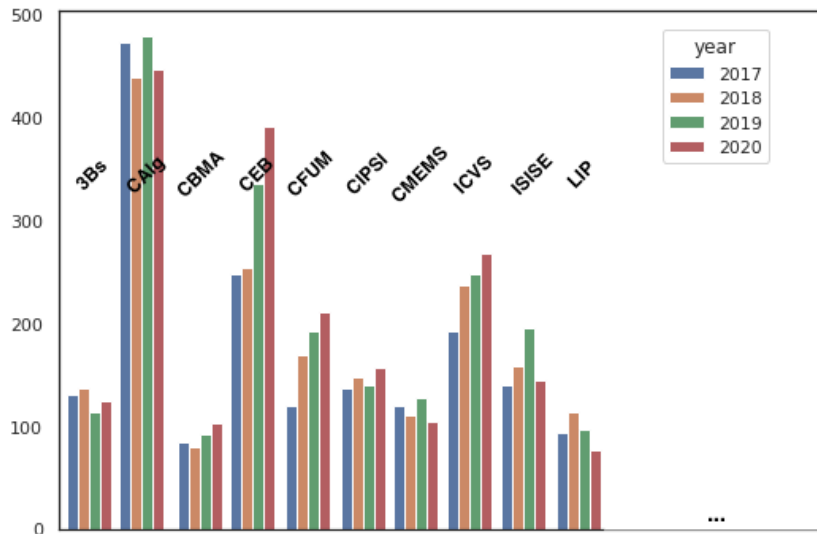


Figure 22: Amount of scientific Production by Research Centre

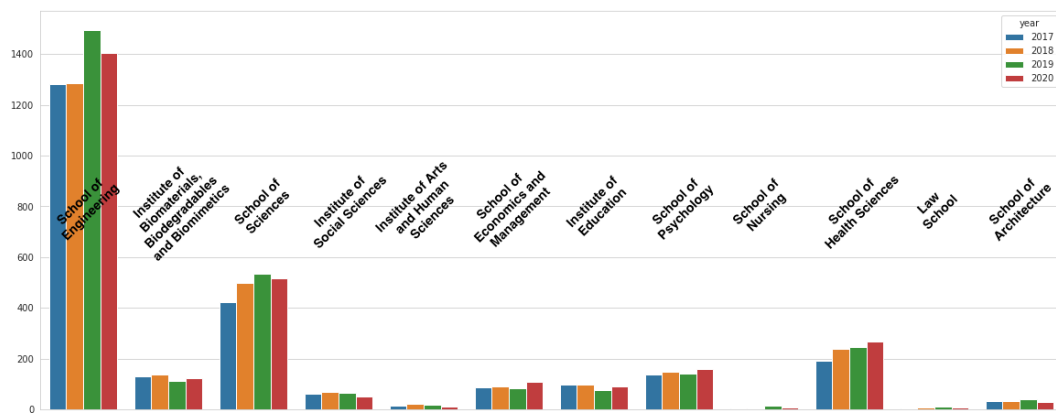


Figure 23: Amount of scientific Production by Organic Unit

UMinho’s publications in the aforementioned 4 years can also be separated by the areas defined in Scopus and see which areas receive the best classification. Table 8 lists the 5 areas that each bibliometric indicator best classifies.

SJR	H-index
<ol style="list-style-type: none"> 1. Decision Sciences 2. Biochemistry, Genetics and Molecular Biology 3. Engineering 4. Nursing 5. Energy 	<ol style="list-style-type: none"> 1. Decision Sciences 2. Dentistry 3. Computer Science 4. Chemical Engineering 5. Medicine
JIF	SNIP
<ol style="list-style-type: none"> 1. Decision Sciences 2. Veterinary 3. Biochemistry, Genetics and Molecular Biology 4. Immunology and Microbiology 5. Engineering 	<ol style="list-style-type: none"> 1. Multidisciplinary 2. Veterinary 3. Earth and Planetary Sciences 4. Biochemistry, Genetics and Molecular Biology 5. Immunology and Microbiology
Cites/Doc.	CiteScore
<ol style="list-style-type: none"> 1. Nursing 2. Veterinary 3. Mathematics 4. Earth and Planetary Sciences 5. Multidisciplinary 	<ol style="list-style-type: none"> 1. Physics and Astronomy 2. Immunology and Microbiology 3. Neuroscience 4. Biochemistry, Genetics and Molecular Biology 5. Energy
IPP	Eigenfactor
<ol style="list-style-type: none"> 1. Veterinary 2. Decision Sciences 3. Biochemistry, Genetics and Molecular Biology 4. Earth and Planetary Sciences 5. Chemical Engineering 	<ol style="list-style-type: none"> 1. Computer Science 2. Biochemistry, Genetics and Molecular Biology 3. Arts and Humanities 4. Pharmacology, Toxicology and Pharmaceutics 5. Health Professions

Table 8: Bibliometric Indicators Top 5 Areas: UMinho

Considering all those metrics, the area that appears most frequently in this top 5 is Biochemistry, Genetics, and Molecular Biology with 6 occurrences (6/40), followed by Decision Sciences and Veterinary with 4 occurrences (4/40).

4.3.3 Correlation between Financial Execution and Scientific Production

In order to draw more informed conclusions about the correlation between the financial execution of missions and scientific dissemination and scientific production, we analyzed it using the Pearson's correlation method, as in [Alves et al. \(2021a\)](#).

Looking at the results, we can say that there is almost no correlation between these two variables, as the correlation between the financial execution and scientific production and the amount of scientific production is very close to 0. Besides, we have a very negative correlation in terms of the correlation of the same financial

	UMinho's financial execution
Quantity of Scientific Production	-0,05
SJR	-0,99
IPP	-0,82
CiteScore	-0,99
Cites/Doc.	-0,93

Table 9: Correlation between financial execution and scientific production

execution with the quality of the scientific output produced. This means that for the years and metrics studied, the quality of work done has increased, even though financial execution for the rubric in question has declined sharply.

Concluding, we can say that there is not a strong correlation between the financial execution for scientific production and the amount of that production. From a qualitative point of view, however, we can find a strong negative correlation. That is, for the years taken into account, despite there being a big decrease in terms of financial execution, there is an increase in the quality of the work produced.

4.4 SCIENTIFIC PRODUCTION IN PORTUGUESE PUBLIC UNIVERSITIES

In order to further investigate the evolution of scientific production, we conducted a study to explore the topic beyond the University of Minho. This way, the analysis of scientific production in Portuguese public universities was carried out. This investigation, besides helping to understand and enrich the evolution of scientific production at UMinho, is also an opportunity to explore some bibliometric indicators and their evaluation. As mentioned earlier, not all indicators assess scientific production in the same way, leading to some scientific areas being better evaluated than others, resulting in a substantial devaluation of their scientific work.

The main question we want to answer is "which scientific areas are best classified by which bibliometric indicators?", as questioned in [Alves et al. \(2021b\)](#).

While the evaluation of scientific production in 4.3 was done only taking into account the University of Minho, at this point the evaluation is done at the national level. In other words, all papers published by all Portuguese public universities in the last 4 years (2017, 2018, 2019 and 2020) were collected. This collection allowed us to make a quantitative analysis of the scientific production of these universities for the years mentioned, but, as happened before, it did not allow us to evaluate the scientific production qualitatively. Therefore, the same merge method were used, as explained at figure 17. This means that, to carry out this evaluation, we began to use the SCImago Journal & Country Rank, powered by Scopus, where we got several indicators of qualitative evaluation, such as SCImago Journal Rank (SJR), H-index, Cites/Document. Then, we merged the journals where the papers were published with the journals indexed in SCImago to get their classification metrics.

4.4.1 Analysis of the Scientific Production of Portuguese Public Universities

Firstly, we carried out a quantitative analysis of the scientific production of Portuguese public universities for the years 2017 to 2020. There are universities with a much higher scientific output than others because they are also larger institutions with a larger scientific dimension. It is also important to note that this quantitative information was taken from the Scopus database, since it is the database chosen by UMinho to manage scientific production.

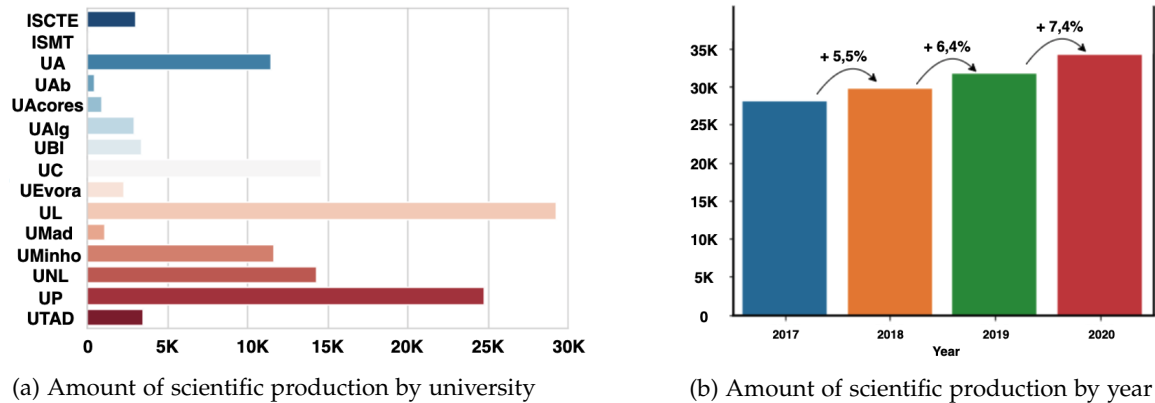


Figure 24: Amount of scientific production for the years 2017-2020

Scientific production for the considered universities increased in 2020. The greatest development took place exactly from 2019 to 2020, where there was an increase of 7,4% of scientific work produced. The most important thing from this analysis, however, is the remarkable evolution in the number of publications over the years. If we look at the papers indexed in Scopus, one of the largest and most reliable databases, we also draw this conclusion, as depicted in figure 13.

Let's see what happened in terms of the quality of this production:

Year	Amount of Papers	SJR	H-index	Cites/Doc.	IPP	SNIP	JIF	Eigenfactor	CiteScore
2020	34158	1,13	90,46	4,49	3,26	1,27	3,93	0,07	2,65
2019	31633	1,09	94,20	4,35	3,17	1,23	3,92	0,08	2,64
2018	29615	1,17	99,34	4,06	2,92	1,21	3,70	0,08	2,41
2017	27992	1,21	99,23	3,40	2,72	1,17	3,73	0,07	2,67

Table 10: Quantity and quality of Scientific Production from Portuguese Public Universities

All the values of the bibliometric indicators displayed correspond to average values. The most common quartile of the four years was also verified, resulting in the first quarter (Q1) for all years.

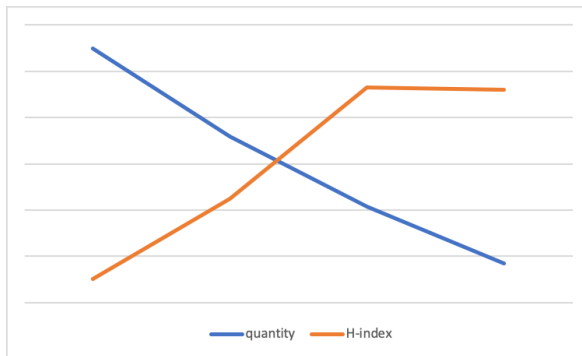
Compared to 2019, SJR and Cites/Document have increased. The latter reached the maximum value of these 4 years, while the maximum average value for SJR was in 2017 and for the H-Index in 2018. IPP, SNIP

and JIF also have the highest values in 2020. Eigenfactor peaked in 2018 and 2019, while CiteScore hit it in 2017.

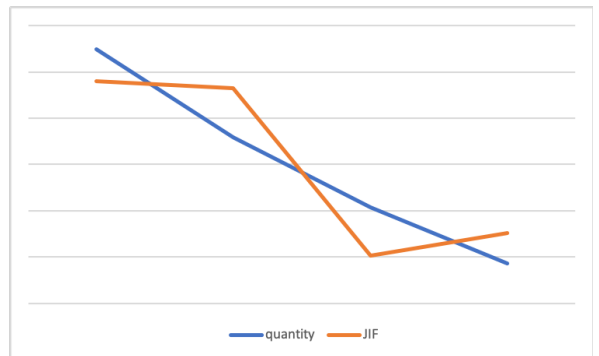
	Quantity of Scientific Production
SJR	-0,75
H-index	-0,96
Cites/Doc.	0,91
IPP	0,97
SNIP	0,99
JIF	0,87
Eigenfactor	-0,10
CiteScore	0,24

Table 11: Correlation between quantity and quality of Scientific Production

These results suggest that the correlation between the quantity and quality of scientific production depends to a large extent on the indicator used to measure quality. For example, as the number of papers increases, the SJR decreases, but the JIF also tends to increase.



(a) Correlation between the quantity and H-index



(b) Correlation between the quantity and JIF

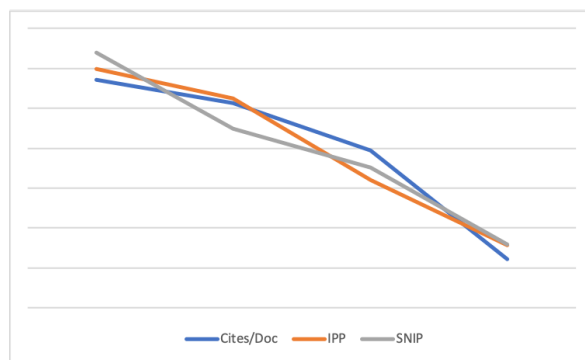
Figure 25: Correlation between the amount of papers and the indicators

Therefore, these indicators needed to be monitored more closely. When examining the correlation between the various metrics (Table 12), we found that only the Eigenfactor and CiteScore show a very weak correlation with the others, which makes sense, since already in the study correlation of the metrics with the amount of scientific production these two had a weak correlation with quantity, while the others had a strong correlation (whether positive or negative). We also conclude that the behavior of Cites/Doc., IPP and SNIP when evaluating the data in question is very similar.

	SJR	H-index	Cites/Doc	IPP	SNIP	JIF	Eigenfactor	CiteScore
SJR	1,0	0,72	-0,87	-0,89	-0,74	-0,85	-0,45	-0,16
H-index		1,0	-0,80	-0,92	-0,91	-0,94	0,26	0,48
Cites/Doc			1,0	0,97	0,95	0,77	0,31	-0,07
IPP				1,0	0,96	0,90	0,13	0,18
SNIP					1,0	0,80	0,00	0,09
JIF						1,0	-0,09	0,58
Eigenfactor							1,0	-0,64
CiteScore								1,0

Table 12: Correlation matrix of bibliometric indicators

Hence, it is also necessary to pay attention to JIF (Journal Impact Factor) as its behavior also shows something close to the behavior of the given metrics and its correlation with the amount of scientific production is also considered high. Even so, despite its differences, we cannot guarantee with such certainty that this indicator will behave like the others, since its deviations are still significant. With this observation we confirm that the evaluation of the quality of the scientific work with Cites/Document, IPP or SNIP is the same. The evaluation carried out with the JIF can also behave in the same way. However, in order to have greater certainty, a study must be carried out with more data, that is, collect data from more years and observe its behavior compared to other bibliometric indicators. Figure 26 shows the behavior of the three metrics mentioned above and the behavior of the JIF indicator compared to the average of such metrics.



(a) Behavior of Cites/Doc., IPP and SNIP



(b) Comparison of JIF's behavior with the average of the others three metrics

Figure 26: Bibliometric Indicators behavior

4.4.2 Comparison of Scientific Production Indicators

We also did a more detailed analysis of the metrics to get a wider range of ratings. The purpose of this analysis was to understand which areas are better classified and with which metrics. Below, in Table 13, we see the top 5 of each bibliometric indicator.

Chemistry, Energy and Biochemistry, Genetics and Molecular Biology are the most common areas, all with 5 occurrences (5/40).

SJR	H-index
1. Energy 2. Decision Sciences 3. Engineering 4. Chemistry 5. Biochemistry, Genetics and Molecular Biology	1. Computer Science 2. Chemistry 3. Pharmacology, Toxicology and Pharmaceutics 4. Arts and Humanities 5. Nursing
JIF	SNIP
1. Veterinary 2. Engineering 3. Chemistry 4. Immunology and Microbiology 5. Biochemistry, Genetics and Molecular Biology	1. Veterinary 2. Earth and Planetary Sciences 3. Engineering 4. Dentistry 5. Energy
Cites/Doc.	CiteScore
1. Nursing 2. Veterinary 3. Energy 4. Biochemistry, Genetics and Molecular Biology 5. Earth and Planetary Sciences	1. Neuroscience 2. Immunology and Microbiology 3. Biochemistry, Genetics and Molecular Biology 4. Physics and Astronomy 5. Energy
IPP	Eigenfactor
1. Veterinary 2. Energy 3. Earth and Planetary Sciences 4. Engineering 5. Chemistry	1. Computer Science 2. Agricultural and Biological Sciences 3. Arts and Humanities 4. Biochemistry, Genetics and Molecular Biology 5. Chemistry

Table 13: Bibliometric Indicators Top 5 Areas - Portuguese Public Universities

As we can see by comparing the areas shown in Table 4 with the areas shown in Table 13, there are some areas that don't even appear in the latter:

- Business, Management and Accounting
- Mathematics
- Chemical Engineering
- Medicine
- Economics, Econometrics and Finance
- Multidisciplinary
- Environmental Science
- Psychology
- Health Professions
- Social Sciences
- Materials Science

Some authors, such as Moed (2010), have already stated that it is inappropriate to compare indicators based on the number of citations, as the way the different areas create the citations can be very different. For example, work in the field of biochemistry often has over 50 references, while work in the field of mathematics can only have 10 references. This difference explains why papers in biochemistry are much more cited than those in mathematics.

In fact, math is one of those areas that never appears in the top 5 metrics presented, while the area of biochemistry is exactly the area that comes up the most.

Additionally, we also analyzed these metrics by university, that is, we checked which universities are performing the best on each of the metrics we examined:

	SJR	H-index	JIF	SNIP	Cites/Doc.	CiteScore	IPP	Eigenfactor	Total
ISCTE				1st					1
ISMT									0
UA		2nd	1st		2nd	3rd	1st		5
UAb				2nd					1
UAcores						4th		5th	2
UAlg								1st	1
UBI						2nd			1
UC	5th								1
UEvora									0
UL	1st	1st	4th	3rd	5th		5th	4th	7
UMadeira						1st			1
UMinho	4th	5th	5th	4th	1st		3rd		6
UNL	3rd	3rd	3rd		3rd		2nd	3rd	6
UP	2nd	4th	2nd	5th	4th		4th	2nd	7
UTAD						5th			1

Table 14: Top 5 Universities per Bibliometric Indicator

The most common universities are UP (University of Porto) and UL (University of Lisbon), both with 7 occurrences, followed by UMinho (University of Minho) and UNL (NOVA University Lisbon) with 6 occurrences.

Nonetheless, the frequency rating may not be the most accurate because the metrics do not rate all parameters and may not all have the same relevance. Indeed, several authors argue that the quality assessment of scientific production should be based on multiple metrics, not just one. We believe that assigning weights is a fair way to use different metrics for scoring.

Thus, we assigned a weight to each metric (from 1 to 5) based on their correlation with the others (Table 12). The more one metric correlates with the others, the greater its weight. The metrics with the highest weight (5) had five very strong correlations (shown in dark red in Table 12): Cites/Document, IPP, and JIF. With weight 4, the metrics that had four very strong correlations: H-index and SNIP. With weight 3, those who had three very strong correlations (SJR) and with weight 1 those who did not have a very strong correlation (Eigenfactor and CiteScore).

	Weight
Cites/Doc.	5
IPP	5
JIF	5
H-index	4
SNIP	4
SJR	3
Eigenfactor	1
CiteScore	1

Table 15: Bibliometric Indicators Weights

Consequently, for each university shown in the table, its value (1 to 5) was normalized and the sum of the product was calculated using the above weights. With this approach, we got the following top 5:

1. UA (University of Aveiro)
2. UL (University of Lisbon)
3. UMinho (University of Minho)
4. UP (University of Porto)
5. UNL (NOVA University Lisbon)

We followed the same approach for the scientific areas, in which we also received a new top 5 overall:

1. Energy
2. Engineering

3. Chemistry
4. Earth and Planetary Sciences
5. Biochemistry, Genetics and Molecular Biology

These results show that evaluating scientific production quality by frequency may not be the best approach. We based the weights given by the authors of [Alves et al. \(2021b\)](#) for the various indicators on data from Portuguese papers and the correlation between the indicators. As in this work, it is recommended to weigh these weights on a case-by-case basis, considering the case of the study.

4.5 MAKING PREDICTIONS ON R&D PROJECTS

Finally, there was also an opportunity to explore the financial data provided by UMinho a little more. One of the challenges of this dissertation was the possibility of making predictions using the financial execution data of R&D projects.

To prevent money from being badly distributed, badly spent or even wasted, the help of these predictions is a very good approach. This way, we can come to a new way of analyzing the future of research projects and perhaps more effectively direct the money used in them. This would be a big step in project management because, in addition to being able to analyze the data present through the data visualization platform, we can also have a system to accurately predict the financial performance of these projects.

When we obtain this information, the financial execution of the R&D projects will undergo positive changes, and the money will have a much more guided path.

In addition, there are also several successful studies that use machine learning for similar problems, such as [Yeh and Chen \(2020\)](#); [Acion et al. \(2017\)](#); [Mazumdar et al. \(2020\)](#); [Munos et al. \(2021\)](#); [Salah et al. \(2018\)](#), among others.

4.5.1 *Expenses Forecast*

Initially, the following question was proposed: "Is it possible to predict the expenses incurred in R&D projects?". This question turned out to be much more complex than initially expected and a great deal of effort was put into trying to get a positive answer.

Firstly, the data were treated in a way that the forecast took into account only columns that, in fact, represented useful information. For example, columns such as the "subcenter" that serves as an ID for a project would not help in obtaining results, so it was quickly eliminated. Columns that could condition the final result were also eliminated, like columns that are not known a priori, since most of them are largely related to the final expenditure, which is what we want to forecast. On the other hand, the budget and duration of a project are very important when it comes to forecasting expenses. Thus, after the correct treatment of the data, several algorithms were used (e.g.: Random Forest and Gradient Boosting) to obtain positive results.

However, several problems were encountered. Although not entirely unexpected, the biggest problem encountered was the lack of data. UMinho has only electronic financial data since 2018, that is, in electronic format, the expense records of R&D projects date only since that year. In addition, it also made data available only until November 2020. This means that there are data for only 3 years (2018, 2019, and 2020), which in practice represents only 35 months of data (forecast points). But the lack of data was not the only problem since the expenses incurred in this area are very unpredictable. For example, in one month a project may have an expense of just 5€ in the purchase of office supplies, and in the next, it may have an expense of thousands of euros in travel.

Despite these obstacles, we managed to obtain very positive results. Table 16 shows the results obtained. We can say that these models may prove themselves useful for financial management and, if deepened, they will certainly be beneficial. Still, we believe that these results can be improved, so to get better and more cohesive conclusions, more data collection is essential.

To get these conclusions and assess whether it would be possible to make these predictions, regression algorithms were used, including Random Forest and Linear Regression. Those algorithms were already explored on 3.3.2.2 as examples of supervised learning. Linear regression is another very well known algorithm that is often used to predict one variable from another. Authors such as Smith et al. (2013) also stated that linear regression is by far the most common form of regression used, since it is a very simple algorithm that establishes a linear relationship between dependent and independent variables. Besides, two of the most frequently used metrics are RMSE and R^2 , so these were the metrics used to make the evaluation.

Root Mean Squared Error (RMSE) is the metric that calculates the average of the square roots of the error between values (actual) and predictions (hypotheses). This metric has a range from 0 to infinity and returns the magnitude of errors. The scores are negatively-oriented, so lower values are better. A score of 0 means that, on average, the predictions are great, that is, 100% effective.

On the other hand, R^2 or R-Squared is the proportion of variation in the outcome that is explained by the predictor variables. Another possible definition is given by the formula: $\frac{\text{total variance explained by model}}{\text{total variance}}$. Unlike RMSE, the higher the value, the better the model. 1 means that the model explains 100% of the variance of the labels. 0 means that the model doesn't understand how the labels vary.

	RMSE	R^2
Random Forest	271.60 ($\approx 6.07\%$)	0.61 (61%)
Linear Regression	309.14 ($\approx 6.50\%$)	0.34 (34%)

Table 16: RMSE and R^2 results

Since the beginning, we considered this problem a substantial challenge because of the inherent difficulty in forecasting expenses, so getting positive results was very satisfactory.

Observing them, we can see that the Random Forest model performs better, with an average error (RSME) of 6.07% and a R-squared (R^2) of 0.61 which means that the model explains 61% of the variance. The Linear Regression model has a RMSE of 6.50% and a R-squared of 0.34 (34%).

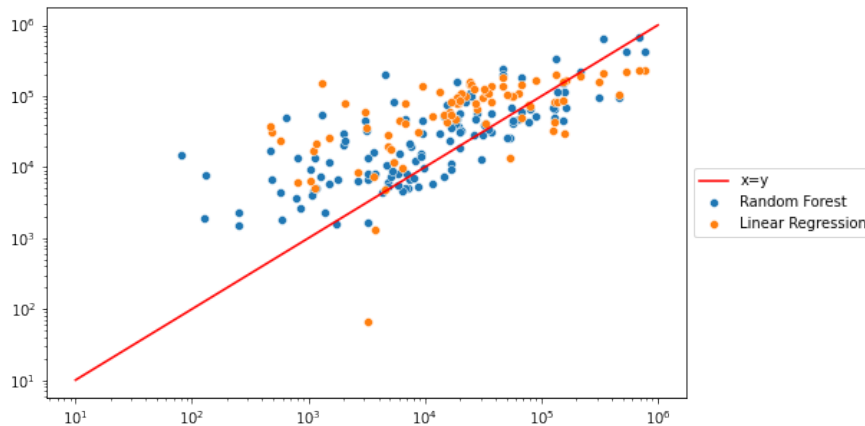


Figure 27: Graphic representation of the behavior of the two algorithms. Blue dots represent predictions made by the Random Forest algorithm and orange dots represent predictions made by the Linear Regression algorithm. The red line represents the ideal algorithm, that is, an algorithm that perfectly predicted these expenses.

Observing figure 27 we can also see that there are more blue dots crossing the red line than orange dots, which is in agreement with the conclusion made in the previous paragraph.

This result, despite responding affirmatively to the question initially proposed, alerts us to the need for these data to be analyzed and managed, continuously and attentively, by humans. Although the results correspond to our expectations, there is still a considerable margin of error that is very important for financial management and that must be considered. In turn, this need leads us to the importance of understanding the data to be processed and its visualization, leading us, once again, to highlight the importance of the main focus of this dissertation: the development of a Data Visualization platform. This platform is, in a way, the answer to the problem that was exposed here, since it aims to facilitate the analysis of these data that are so important and fluctuating in time.

It is also important to point out that these forecasts can be very useful overall, but for an isolated example the forecast is likely to be worse. Thus, despite the good results obtained, this topic should be explored more diligently to better understand the problem and get better results.

4.5.2 Success Forecast

After the expenses forecasting, the idea of trying to predict whether a project will be successful or not came up. This approach solves the problem of data irregularity since it is done through binary classification.

The success of a project is measured using a formula created by the UMinho rectory that makes use of two other very important rates and on which success depends: the time rate and the financial rate.

So, we have the time rate:

$$time\ rate = \frac{months\ already\ passed}{total\ project\ months} \times 100 \quad (1)$$

And the financial rate:

$$financial\ rate = \frac{total\ expenditure}{project\ budget} \times 100 \quad (2)$$

And the success rate is measured based on these two rates:

$$financial\ rate - time\ rate \quad (3)$$

This gives us a value between -100 and 100 and the values are grouped according to the result as follows:

- ≤ 40 : project with very low success rate;
- $> 40 \leq -10$: project with low success rate;
- $> -10 \leq 20$: project with good success rate;
- > 20 : project with very good success rate.

These results are based on the time range where the project is located and not just on the budget execution of each one, that is, these rates consider the duration of the project and what time it is in. Therefore, during the execution of the project, we can monitor its status and determine if it is going the right way. A project that ends with a success rate greater than -10 is a successful project. Otherwise, it is an unsuccessful project.

The binary classification process (successful or unsuccessful) was done using only projects that were already finished, since considering projects still in progress could influence the results. In addition, we used projects that started before 2018 for training, while projects that started at least in 2018 were used for testing.

This process has given us very positive results regarding the possibility of predicting the success of a project. We obtained an AUC of 83% and an accuracy of 79%, which represents a very satisfactory value. This means that we can predict whether or not a project will be successful with close to 80% certainty, which we consider very important and promising, as it prepares the responsible entities for eventual problems and expenses without profit.

	False	True	Error	Rate
False	40.0	9.0	0.1837	(9.0/49.0)
True	9.0	28.0	0.2432	(9.0/37.0)
Total	49.0	37.0	0.2093	(18.0/86.0)

Table 17: Confusion Matrix (Act/Pred)

These results were acquired through a *stackensemble* algorithm from H2O, which is a way of combining the answers of several models. In other words, and according to H2O documentation, *is a supervised ensemble machine learning algorithm that finds the optimal combination of a collection of prediction algorithms using a process called stacking.*

Although the results are good, they lead us to believe that there is still a great margin of progression, that is, we believe that getting more data (more projects) these results will be even better.

CONCLUSIONS AND FUTURE WORK

“If at first the idea is not absurd, then there is no hope for it.”

Albert Einstein

This dissertation had as its main objective the development of a system capable of responding to the needs of the most varied research bodies, namely the Responsible Researchers, Directors of Research Centers, Presidents of Organic Units, UMinho Rectory and Administrative Staff allied to these bodies.

The development of this system not only represents a new approach for UMinho with regard to the financial management of R&D projects but also brings the possibility of using this resource as a source for monitoring these same projects and their investments. Besides, it also made it possible to explore other peripheral themes as the scientific production at UMinho and Portugal, its relationship with financial execution, and the use of ML to make forecasts.

5.1 CONCLUSIONS

The developed Data Visualization platform was a success. After several tests and presentations to the various organic units and the rectory of the University of Minho, the university allowed this platform to be implemented in the internal structure of UMinho. This was a very important step towards improving the platform since it was being fed with data from excel files and whose processing depended on human work. With this permission, together with the Information and Communications Systems Services Unit (USSIC), the prototype is already being adapted so that data loading is done automatically. We believe that this platform will bring many advantages for the financial management of projects and, with its insertion in the internal IT structure of UMinho, its adaptability will be guaranteed.

Regarding the work carried out beyond the platform, we can also say that it was quite positive. Starting with the brief study of the impact of the Covid-19 pandemic, we were able to draw promising conclusions that led us to new questions, also addressed. In short, we can say that the pandemic had a very negative impact on the project's financial execution, especially concerning the missions and scientific production sector. However, the quantity and quality of scientific production were not negatively affected. As a result, we conclude that there is no

strong correlation between financial execution for scientific production and the quantity of this production. Additionally, the 5 best-classified areas for each bibliometric indicator considered were also studied, with Biochemistry, Genetics and Molecular Biology being the area that appears most often.

Expanding the focus area of this work and exploring the scientific production at a national level, we also obtained very interesting results. Starting with bibliometric indicators, we can say that these can be used as quality measures, but their use depends on the area of study, as they have different ways of evaluating the work carried out. We also know, based on data from the Scopus database and the indicators studied, that the Cites/Document, IPP, and SNIP have very similar behavior concerning the evaluation of scientific work, so we believe that the use of the three might be redundant. The JIF also behaves similarly. But, to draw conclusions, as a way of future work, we intend to increase the dataset and collect more data to study this nuance, since its behavior line ends up clashing a little compared to the other indicators. Finally, it is also defended that the most correct way to evaluate scientific work is not through the frequency per indicator, but through the allocation of weights for each indicator.

Regarding the realization of forecasts, we also got good news. Two different approaches were taken on this topic: expense and success forecasting. Concerning the forecast of expenses, we got good results. With the Random Forest model we were able to get an error of 6% and a R-squared of 0.61 (61%). Although these are not perfect results, we have few data expenses and these are very irregular and extremely unpredictable, so we consider these promising results. Even so, there is still a large margin for progression. However, this issue underscores the importance of the DV platform mentioned above. Since these data are so irregular, humans must handle their analysis. Thus, this problem reiterates the need to have a system that facilitates the understanding and visualization of data for human consumption. Also, when it comes to predicting the success of a project, we got even better results. We can say whether or not a project will be successful with approximately 80% certainty.

Despite the positive results achieved, it is also important to highlight some obstacles that were encountered throughout its development. Starting with the development of the platform, we can say that the biggest obstacle was data processing. Since this data was sent in excel files and its processing was done manually, obtaining consumable data was very time-consuming. When it comes to the study of scientific production, both at UMinho and at a national level, the biggest challenge was the lack of data. This was also one of the main problems in forecasting project expenses. In addition, the financial execution of research projects is a sensitive issue, so it was only possible to access financial data from UMinho, making it impossible to carry out a more comprehensive analysis of the topic.

5.2 SCIENTIFIC PUBLICATIONS

The work developed within the scope of this thesis was disseminated through the participation in three scientific events.

- **5th Theory and Applications in the Knowledge Economy Conference**

The TAKE 2021 had as main topic The Knowledge Economy in the Covid-19 Era and took place virtually on July 7th-9th, 2021. TAKE is an international scientific conference devoted to the multidisciplinary study of the knowledge economy. In particular, it intends to analyze the relation and the gap between theories and practices in the knowledge economy of the 21st century.

The theme introduced at this conference was the impact of Covid-19 on RD projects.

The Impact of Covid-19 on Research and Development Projects. Inês Alves, University of Minho, Informatics Department, ALGORITMI Center, Braga, Portugal; Cesar Analide, University of Minho, Informatics Department, ALGORITMI Center, Braga, Portugal; and Filipe Vaz, University of Minho, Physics Department, Physics Center, Guimarães, Portugal.

- **18th International Conference on Distributed Computing and Artificial Intelligence**

The 18th International Conference on Distributed Computing and Artificial Intelligence (DCAI) 2021 is an annual forum that will bring together ideas, projects, lessons and others, associated with distributed computing and artificial intelligence, and their application in different areas. DCAI 2021 will be held in Salamanca, Spain within PAAMS'21 in 6th-8th October, 2021.

In this meeting, the work presented is related to the relationship between financial execution in R&D and scientific production.

The Relationship Between Financial Execution in R&D and Scientific Production. Inês Alves, University of Minho, Informatics Department, ALGORITMI Center, Braga, Portugal; Cesar Analide, University of Minho, Informatics Department, ALGORITMI Center, Braga, Portugal; and Filipe Vaz, University of Minho, Physics Department, Physics Center, Guimarães, Portugal.

- **12th International Symposium on Ambient Intelligence**

ISAmI is the International Symposium on Ambient Intelligence, aiming to bring together researchers from various disciplines that constitute the scientific field of Ambient Intelligence to present and discuss the latest results, new ideas, projects and lessons learned. This will also be held in Salamanca, Spain on 6th-8th October, 2021.

Here, the work presented is about Scientific Production in Portuguese Public Universities.

Scientific Production in Portuguese Public Universities. Inês Alves, University of Minho, Informatics Department, ALGORITMI Center, Braga, Portugal; Cesar Analide, University of Minho, Informatics Department, ALGORITMI Center, Braga, Portugal; and Filipe Vaz, University of Minho, Physics Department, Physics Center, Guimarães, Portugal.

5.3 FUTURE WORK

Finished this dissertation, we have a very positive balance of the work developed. As future work, it is intended to continue to develop the created platform, this time migrating the work done to a responsible entity, in this case to the Centro de Computação Gráfica (CCG), a company that works together with UMinho. The migration of this platform requires the adaptation of the technologies used to those used in the company and the continuation of the last level development. On the other hand, this new approach means that we are no longer dependent on the periodic sending of excel files and the platform is now fed by data from the UMinho database, reducing the time spent on manual aggregation and reducing the risk of errors.

Another important point is to obtain more data, both on scientific production and projects and their financial information. With more data we can study the behavior of the various bibliometric indicators, deepening the study carried out and obtaining safer answers to the behavior of the JIF referred to in 4.4.1, and also improve forecast results using more projects. Furthermore, as this data increases, we can also continue to monitor the relationship between the financial execution and the scientific production.

BIBLIOGRAPHY

- Laura Acion, Diana Kelmansky, Mark van der Laan, Ethan Sahker, DeShauna Jones, and Stephan Arndt. Use of a machine learning framework to predict substance use disorder treatment success. *PloS one*, 12(4): e0175383, 2017.
- Eduardo Vicente Albertin and Daniel Capaldo Amaral. Contexto da parceria como qualificador da gestão de projetos universidade-empresa. *Production*, 20(2):224–236, 2010.
- Inês Alves, Cesar Analide, and Filipe Vaz. The relationship between financial execution in r&d and scientific production. In *Distributed Computing and Artificial Intelligence, Special Sessions 18th International Conference*, 2021a. (To be published).
- Inês Alves, Cesar Analide, and Filipe Vaz. Ambient intelligence – software and applications –, 12th international symposium on ambient intelligence. In *Distributed Computing and Artificial Intelligence, Special Sessions 18th International Conference*, 2021b. (To be published).
- Inês Alves, Cesar Analide, and Filipe Vaz. Covid-19 impact on r&d projects. In *Proceedings of the 5th Theory and Applications in the Knowledge Economy*, 2021c. (To be published).
- Rob Ashmore, Radu Calinescu, and Colin Paterson. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. 2019.
- Mariette Awad and Rahul Khanna. *Machine Learning*, pages 1–18. Apress, Berkeley, CA, 2015. ISBN 978-1-4302-5990-9. doi: 10.1007/978-1-4302-5990-9_1. URL https://doi.org/10.1007/978-1-4302-5990-9_1.
- Ana Azevedo and Manuel Filipe Santos. Kdd, semma and crisp-dm: A parallel overview. pages 182–185, 01 2008.
- D. Baier, R. Decker, and L. Schmidt-Thieme. *Data Analysis and Decision Support*. Springer, 2005.
- Judit Bar-Ilan. Which h-index?—a comparison of wos, scopus and google scholar. *Scientometrics*, 74(2):257–271, 2008.
- TA Barnes, IR Pashby, and AM Gibbons. Managing collaborative r&d projects development of a practical management tool. *International Journal of Project Management*, 24(5):395–404, 2006.
- Pedro Belli. *Is economic analysis of projects still useful?* The World Bank, 1999.

- Drew Bentley. Business intelligence and analytics. *Internet, link: <https://www.pdfdrive.com/business-intelligence-and-analytics-e56416503.html>*, 2017.
- J Martin Bland and Douglas G Altman. Bayesians and frequentists. *BMJ*, 317(7166):1151–1160, 1998. ISSN 0959-8138. doi: 10.1136/bmj.317.7166.1151. URL <https://www.bmj.com/content/317/7166/1151.1>.
- Longbing Cao. Data science: A comprehensive overview. *ACM Comput. Surv.*, 2017. doi: 10.1145/3076253.
- P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. R. Shearer, and R. Wirth. *Crisp-dm 1.0: Step-by-step data mining guide*. 2000.
- Nitesh Chawla. Discovering knowledge in data: An introduction to data mining. *Briefings in Bioinformatics*, 6, 01 2005. doi: 10.1093/bib/6.4.411.
- Lisiane Closs, Gabriela Cardozo Ferreira, Alessandra Freitas Soria, Claudio Hoffmann Sampaio, and Marcelo Perin. Organizational factors that affect the university-industry technology transfer processes of a private university. *Journal of technology management & innovation*, 7(1):104–117, 2012.
- Terry Cooke-Davies. The “real” success factors on projects. *International journal of project management*, 20(3): 185–190, 2002.
- Carlo Dell’Aquila, Francesco Di Tria, Ezio Lefons, and Filippo Tangorra. Business intelligence applications for university decision makers. *WSEAS Transactions on Computers*, 7(7):1010–1019, 2008.
- Rahul C. Deo. Machine learning in medicine. *Circulation*, 132 (20):1920–1930, 2015. doi: 10.1161/circulationaha.115.001593.
- Konstantin Didur. *Machine learning in finance: Why, what how*, 2018. URL <https://towardsdatascience.com/machine-learning-in-finance-why-what-how-d524a2357b56>.
- Matthew Dixon and Igor Halperin. The four horsemen of machine learning in finance. *SSRN Electronic Journal*, 01 2019. doi: 10.2139/ssrn.3453564.
- Yusnier Reyes Dixson, Lissette Nuñez Maturel, et al. La inteligencia de negocio como apoyo a la toma de decisiones en el ámbito académico (business intelligence as decision support system in academic environment). *GECONTEC: Revista Internacional de Gestión del Conocimiento y la Tecnología*, 3(2):63–73, 2015.
- Brent M Drake and Aaron Walz. Evolving business intelligence and data analytics in higher education. *New Directions for Institutional Research*, 2018(178):39–52, 2018.
- Sophie Emerson, Ruairí Kennedy, Luke O’Shea, and John O’Brien. Trends and applications of machine learning in quantitative finance. *SSRN Electronic Journal*, 2019.

- Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, Mar. 1996a. doi: 10.1609/aimag.v17i3.1230. URL <https://ojs.aaai.org/index.php/aimagazine/article/view/1230>.
- Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. Knowledge discovery and data mining: Towards a unifying framework. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 82–88. AAAI Press, 1996b.
- Tony Feghali, Imad Zbib, and Sophia Hallal. A web-based decision support tool for academic advising. *Journal of Educational Technology & Society*, 14(1):82–94, 2011.
- Peter Flach. *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, 1st edition, 2012. ISBN 978-1-107-42222-3.
- Gary Gensler and Lily Bailey. Deep learning and financial stability. *SSRN Electronic Journal*, 2020. doi: 10.2139/ssrn.3723132.
- Anna P Goldstein and Michael Kearney. Know when to fold 'em: An empirical description of risk management in public research funding. *Research Policy*, 49(1):103873, 2020.
- Aurlien Gron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc., 2017. ISBN 978-1-4919-6229-9. doi: 10.1007/978-1-4302-5990-9_1.
- Vicente Guerrero-Bote, Felipe Zapico-Alonso, María Espinosa-Calvo, Rocío Gómez-Crisóstomo, and Félix de Moya-Anegón. Import-export of knowledge between scientific subject categories: The iceberg hypothesis. *Scientometrics*, 71(3):423–441, 2007.
- Vicente P Guerrero-Bote and Félix Moya-Anegón. A further step forward in measuring journals' scientific prestige: The sjr2 indicator. *Journal of informetrics*, 6(4):674–688, 2012.
- Chloe Gui and Victoria Chan. Machine learning in medicine. *University of Western Ontario Medical Journal*, 86(2):76–78, Dec. 2017. doi: 10.5206/uwomj.v86i2.2060. URL <https://ojs.lib.uwo.ca/index.php/uwomj/article/view/2060>.
- Magnus Gulbrandsen and Jens-Christian Smeby. Industry funding and university professors' research performance. *Research Policy*, 34:932–950, 08 2005. doi: 10.1016/j.respol.2005.05.004.
- DENNIS Guster and CHRISTOPHER G Brown. The application of business intelligence to higher education: Technical and managerial perspectives. *Journal of Information Technology Management*, 23(2):42–62, 2012.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., 2001. ISBN 978-0-387-84858-7. doi: 10.1007/978-0-387-84858-7.

- Hemantha SB Herath and Chan S Park. Economic analysis of r&d projects: an options approach. *The Engineering Economist*, 44(1):1–35, 1999.
- Laura Igual, Santi Segu, Jordi Vitri, Eloi Puertas, Petia Radeva, Oriol Pujol, Sergio Escalera, Francesc Dant, and Llus Garrido. *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications*. Springer Publishing Company, 1st edition, 2017. ISBN 1863-7310. doi: 10.1007/978-3-319-50017-1.
- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013. ISBN 978-1-4614-7137-0. URL <https://faculty.marshall.usc.edu/gareth-james/ISL/>.
- Sean Kandel, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank van Ham, Nathalie Henry Riche, Chris Weaver, Bongshin Lee, Dominique Brodbeck, and Paolo Buono. Research directions in data wrangling: Visualizations and transformations for usable and credible data. 2011. doi: 10.1177/1473871611415994.
- Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, and Ioanna Chouvarda. Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, 15:104 – 116, 2017. ISSN 2001-0370. doi: <https://doi.org/10.1016/j.csbj.2016.12.005>. URL <http://www.sciencedirect.com/science/article/pii/S2001037016300733>.
- Vahid Khatibi, Abbas Keramati, and Gholam Ali Montazer. A business intelligence approach to monitoring and trend analysis of national r&d indicators. *Engineering Management Journal*, 29(4):244–257, 2017.
- Sirawit Kleesuwan, Somsak Mitatha, Preecha P Yupapin, and Bunjong Piyatamrong. Business intelligence in thailand’s higher educational resources management. *Procedia-Social and Behavioral Sciences*, 2(1):84–87, 2010.
- Sotiris B Kotsiantis. Use of machine learning techniques for educational proposes: a decision support system for forecasting students’ grades. *Artificial Intelligence Review*, 37(4):331–344, 2012.
- Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 2014. doi: 10.1016/j.csbj.2014.11.005.
- Abhaya V Kulkarni, Brittany Aziz, Iffat Shams, and Jason W Busse. Comparisons of citations in web of science, scopus, and google scholar for articles published in general medical journals. *Jama*, 302(10):1092–1096, 2009.
- Bárbara Lancho-Barrantes, Vicente Guerrero-Bote, and Félix Moya-Anegón. The iceberg hypothesis revisited. *Scientometrics*, 85(2):443–461, 2010a.
- Bárbara S Lancho-Barrantes, Vicente P Guerrero-Bote, and Félix Moya-Anegón. What lies behind the averages and significance of citation indicators in different disciplines? *Journal of Information Science*, 36(3):371–382, 2010b.

- Vincent Larivière, Éric Archambault, Yves Gingras, and Étienne Vignola-Gagné. The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology*, 57(8):997–1004, 2006.
- Zulaikha Lateef. *A Complete Guide To Math And Statistics For Data Science*, 2020. URL <https://www.edureka.co/blog/math-and-statistics-for-data-science/>.
- Michael Levine-Clark and Esther L Gil. A comparative citation analysis of web of science, scopus, and google scholar. *Journal of Business & Finance Librarianship*, 14(1):32–46, 2008.
- Erping Long, Haotian Lin, Zhenzhen Liu, Xiaohang Wu, Liming Wang, Jiewei Jiang, Yingying An, Zhuoling Lin, Xiaoyan Li, Jingjing Chen, Jing Li, Qianzhong Cao, Dongni Wang, Xiyang Liu, Weirong Chen, and Yizhi Liu. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nature Biomedical Engineering*, 1:0024, 01 2017. doi: 10.1038/s41551-016-0024.
- Jonas Lundberg. Lifting the crown—citation z-score. *Journal of informetrics*, 1(2):145–154, 2007.
- Amir Mahmood, Faisal Asghar, and Bushra Naoreen. “success factors on research projects at university” an exploratory study. *Procedia-Social and Behavioral Sciences*, 116:2779–2783, 2014.
- Svetlana Mansmann and Marc H Scholl. Decision support system for managing educational capacity utilization. *IEEE Transactions on Education*, 50(2):143–150, 2007.
- Gonzalo Mariscal, Oscar Marbán, and Covadonga Fernández. A survey of data mining and knowledge discovery process models and methodologies. *Knowledge Eng. Review*, 25:137–166, 06 2010. doi: 10.1017/S0269888910000032.
- M. Markus. Datification, organizational strategy, and is research: What’s the score? *The Journal of Strategic Information Systems*, 26:233–241, 09 2017. doi: 10.1016/j.jsis.2017.08.003.
- Madhu Mazumdar, Jung-Yi Joyce Lin, Wei Zhang, Lihua Li, Mark Liu, Kavita Dharmarajan, Mark Sanderson, Luis Isola, and Liangyuan Hu. Comparison of statistical and machine learning models for healthcare cost data: a simulation study motivated by oncology care model (ocm) data. *BMC health services research*, 20(1):1–12, 2020.
- C. McCue. Data mining and predictive analysis: Intelligence gathering and crime analysis: Second edition. *Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis: Second Edition*, pages 1–393, 01 2015.
- Ignasi Merediz-Solà and Aurelio F Bariviera. A bibliometric analysis of bitcoin scientific production. *Research in International Business and Finance*, 50:294–305, 2019.
- William K Michener and James W Brunt. *Ecological data: Design, management and processing*. John Wiley & Sons, 2009.

- Tom M. Mitchell. *Machine Learning*. McGraw Hill, 2017. <http://www.cs.cmu.edu/~tom/mlbook/keyIdeas.pdf>.
- Henk F Moed. The source-normalized impact per paper (snip) is a valid and sophisticated indicator of journal citation impact. *arXiv preprint arXiv:1005.4906*, 2010.
- Bernie Monegain. *IOM sounds alarm on diagnostic errors*, 2015. URL <https://www.healthcareitnews.com/news/iom-sounds-alarm-diagnostic-errors>.
- Philippe Mongeon and Adèle Paul-Hus. The journal coverage of web of science and scopus: a comparative analysis. *Scientometrics*, 106(1):213–228, 2016.
- Francisco G Montoya, Alfredo Alcayde, Raúl Baños, and Francisco Manzano-Agugliaro. A fast method for identifying worldwide scientific collaborations using the scopus database. *Telematics and Informatics*, 35(1): 168–185, 2018.
- Bernard Munos, Jan Niederreiter, and Massimo Riccaboni. Improving the prediction of clinical success using machine learning. *medRxiv*, 2021.
- Ruth Murray-Webster and Darren Dalcher. *APM Body of Knowledge*. Association for Project Management, 7th edition, May 2019. ISBN 9781903494820.
- Glenn J. Myatt and Wayne P. Johnson. *Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications*. Wiley Publishing, 2009. ISBN 978-0-470-22280-5. doi: 10.5555/1516229.
- DS Nagesh and Sam Thomas. Success factors of public funded r&d projects. *Current science*, pages 357–363, 2015.
- Richard Van Noorden. Global scientific output doubles every nine years, 2014. URL <http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html>. (Accessed on 05 April 2021).
- Ziad Obermeyer and Ezekiel J. Emanuel. Predicting the future — big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, 375 (13):1216–1219, 2016. doi: 10.1056/NEJMp1606181.
- OECD. Gross domestic spending on rd (indicator), 2021. URL <https://data.oecd.org/rd/gross-domestic-spending-on-r-d.htm#indicator-chart>.
- Devin Pickell. *Qualitative vs Quantitative Data – What’s the Difference?*, 2019. URL <https://learn.g2.com/qualitative-vs-quantitative-data>.
- PMI. *A Guide to the Project Management Body of Knowledge (PMBOK Guide), 4th Edition*. Project Management Institute, 2008.

- PORDATA. Despesas em actividades de investigação e desenvolvimento (id): total e por sector de execução, 2021. URL [https://www.pordata.pt/Portugal/Despesas+em+actividades+de+investiga%C3%A7%C3%A3o+e+desenvolvimento+\(I+D\)+total+e+por+sector+de+execu%C3%A7%C3%A3o-1106](https://www.pordata.pt/Portugal/Despesas+em+actividades+de+investiga%C3%A7%C3%A3o+e+desenvolvimento+(I+D)+total+e+por+sector+de+execu%C3%A7%C3%A3o-1106). (Accessed on 20 March 2021).
- Fabio Vinicius Primak. *Decisões com B.I. (Business Intelligence)*. 2008.
- Sadeghi Ramin and Alireza Sarraf Shirazi. Comparison between impact factor, scimago journal rank indicator and eigenfactor score of nuclear medicine journals. *Nuclear Medicine Review*, 15(2):132–136, 2012.
- Bruce Ratner. *Statistical and Machine-Learning Data Mining*. Taylor Francis Group, 3rd edition, 2017.
- Michal Rogalewicz and Robert Sika. Methodologies of knowledge discovery from data and data mining methods in mechanical engineering. *Management and Production Engineering Review*, 7, 09 2016. doi: 10.1515/mper-2016-0040.
- Catarina Erika Saito and Álvaro Guillermo Rojas Lezana. Fatores de sucesso de projetos universidade-empresa: um quadro atualizado para gestão de projetos. 2015.
- Mohammed Salah, Khaled Altalla, Ahmed Salah, and Samy S Abu-Naser. Predicting medical expenses using artificial neural network. *International Journal of Engineering and Information Systems (IJEAIS)*, 2(20): 11–17, 2018.
- Negin Salimi. Quality assessment of scientific outputs using the bwm. *Scientometrics*, 112(1):195–213, 2017.
- Ruhul Sarker, Hussein Abbass, and C. Newton. *Introducing Data Mining and Knowledge Discovery*. 01 2000. doi: 10.4018/9781930708266.ch001.
- SAS Institute. Enterprise miner - semma. 2006. URL <https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jn8bbj1a2.htm&docsetVersion=14.3>.
- R. Sathya and Annamma Abraham. Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, 2(2), 2013. doi: 10.14569/IJARAI.2013.020206. URL <http://dx.doi.org/10.14569/IJARAI.2013.020206>.
- Per O Seglen. Why the impact factor of journals should not be used for evaluating research. *Bmj*, 314(7079): 497, 1997.
- Umair Shafique and Haseeb Qaiser. A comparative study of data mining process models (kdd, crisp-dm and semma). *International Journal of Innovation and Scientific Research*, 12:2351–8014, 11 2014.
- Darius Singpurwalla. *A Handbook of Statistics: An Overview of Statistical Methods*. Bookboon, 1st edition, 2013.
- Paul F Smith, Siva Ganesh, and Ping Liu. A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of neuroscience methods*, 220(1):85–91, 2013.

- Jeffrey Stanton. *An Introduction to Data Science*. Syracuse University, 3rd edition, 2013.
- Xiaogang Su, Xin Yan, and Chih-Ling Tsai. Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(3):275–294, 2012.
- Pornpimol Sugandhavanija, Sukruedee Sukchai, Nipon Ketjoy, and Sakol Klongboonjit. Determination of effective university–industry joint research for photovoltaic technology transfer (uijrptt) in thailand. *Renewable Energy*, 36(2):600–607, 2011.
- David Taniar. *Data Mining and Knowledge Discovery Technologies*. IGI Publishing - Imprint of: IGI Global, Hershey, PA, 2008. ISBN 978-1-59904-961-8. doi: 10.4018/978-1-59904-960-1.
- Edward Tsang, Paul Yung, and Jin Li. Eddie-automation, a decision support tool for financial forecasting. *Decision Support Systems*, 37(4):559–565, 2004.
- Donatella Ugolini, Monica Neri, Alfredo Cesario, Stefano Bonassi, Daniele Milazzo, Luca Bennati, Luisa Maria Lapenna, and Patrizio Pasqualetti. Scientific production in cancer rehabilitation grows higher: a bibliometric analysis. *Supportive Care in Cancer*, 20(8):1629–1638, 2012.
- University of Minho. *Research and Innovation*. UMinho Editora, 2018. ISBN 978-989-8974-18-1. doi: 10.21814/uminho.ed.16.
- David van Dyk, Montse Fuentes, Michael Jordan, Michael Newton, Bonnie Ray, Duncan Temple Lang, and Hadley Wickham. ASA statement on the role of statistics in data science. 2015. URL <https://magazine.amstat.org/blog/2015/10/01/asa-statement-on-the-role-of-statistics-in-data-science/>.
- Ritvik Voleti. Data wrangling- a goliath of data industry. 2020. ISSN 2278-0181.
- Shijun Wang and Ronald M. Summers. Machine learning and radiology. *Medical Image Analysis*, 16(5):933 – 951, 2012. ISSN 1361-8415. doi: <https://doi.org/10.1016/j.media.2012.02.005>. URL <http://www.sciencedirect.com/science/article/pii/S1361841512000333>.
- Stephen C Ward and Chris B Chapman. Risk-management perspective on the project lifecycle. *International Journal of Project Management*, 13(3):145–149, 1995.
- Claus Weihs and Katja Ickstadt. Data science: the impact of statistics. *International Journal of Data Science and Analytics* 6, pages 189–194, 2018.
- Jason Wrong. Linear regression explained. 2020. URL <https://towardsdatascience.com/linear-regression-explained-1b36f97b7572>.
- Jen-Yin Yeh and Chi-Hua Chen. A machine learning approach to predict the success of crowdfunding fintech project. *Journal of Enterprise Information Management*, 2020.

- Nur Ain Zulkefli, Suraya Miskon, Haslina Hashim, Rose Alinda Alias, Norris Syed Abdullah, Norasnita Ahmad, Nazmona Mat Ali, and Mohd Aizaini Maarof. A business intelligence framework for higher education institutions. *ARPJ. Eng. Appl. Sci*, 10(23):18070–18077, 2015.
- Belén Álvarez Bornstein and María Bordons. Is funding related to higher research impact? exploring its relationship and the mediating role of collaboration in several disciplines. *Journal of Informetrics*, 15(1):101102, 2021. ISSN 1751-1577. doi: <https://doi.org/10.1016/j.joi.2020.101102>. URL <https://www.sciencedirect.com/science/article/pii/S1751157720301309>.

APPENDIX

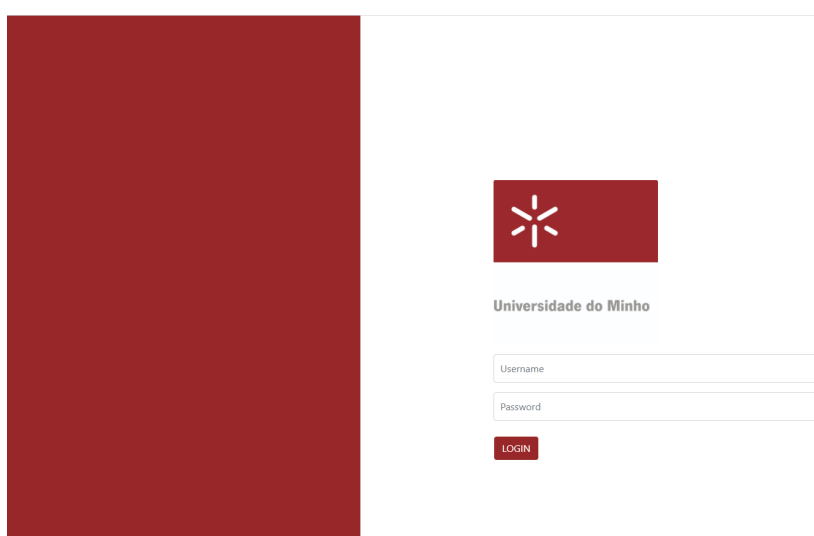


Figure 28: Platform Login Screen

A.1 LEVEL 1: RESEARCHER

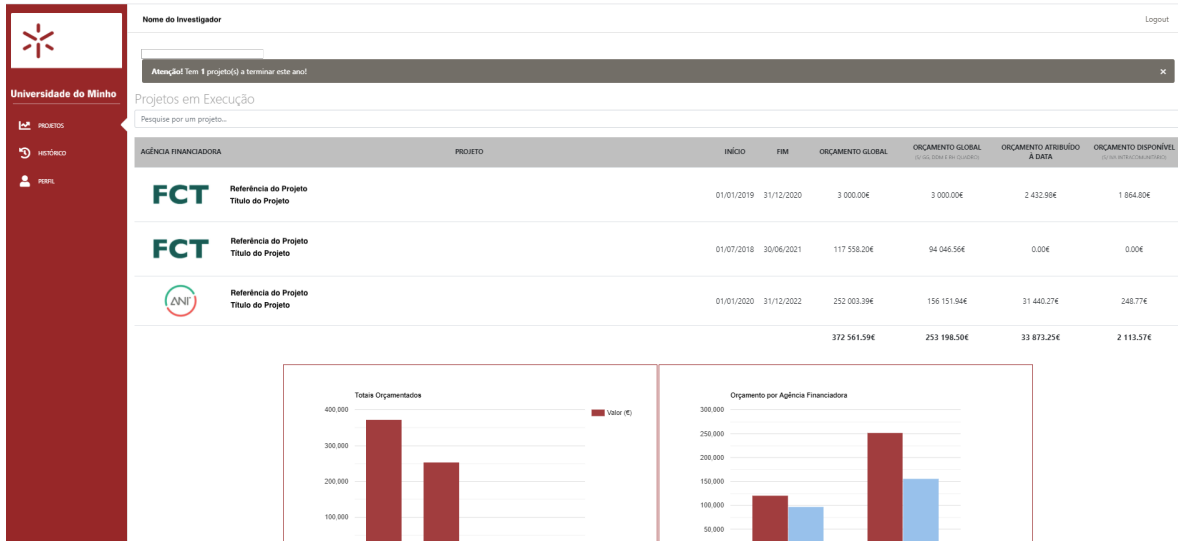


Figure 29: Researcher Main Screen

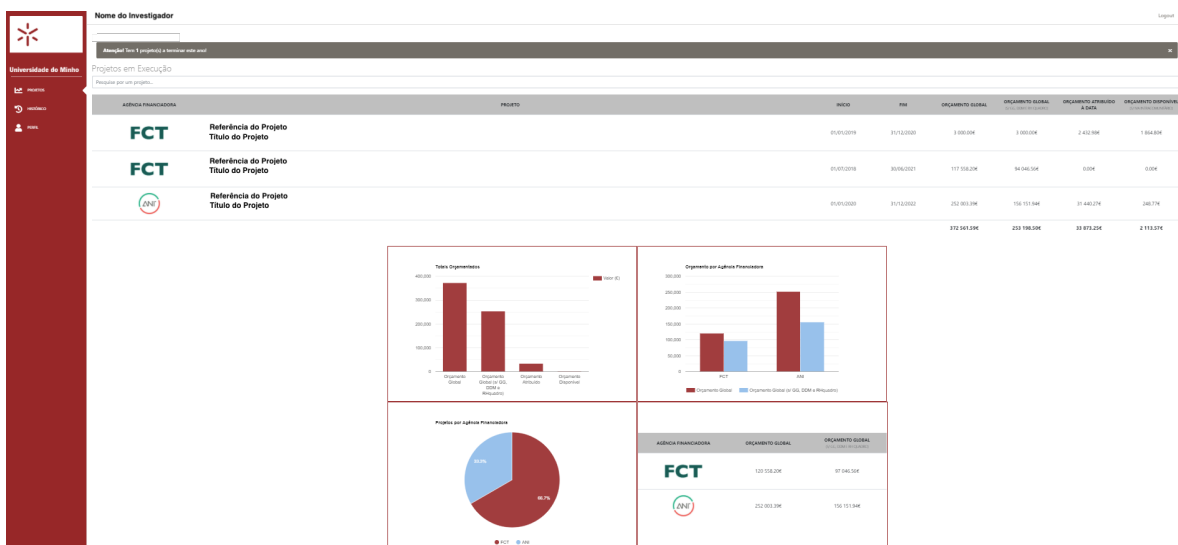


Figure 30: Researcher Main Screen (continued)

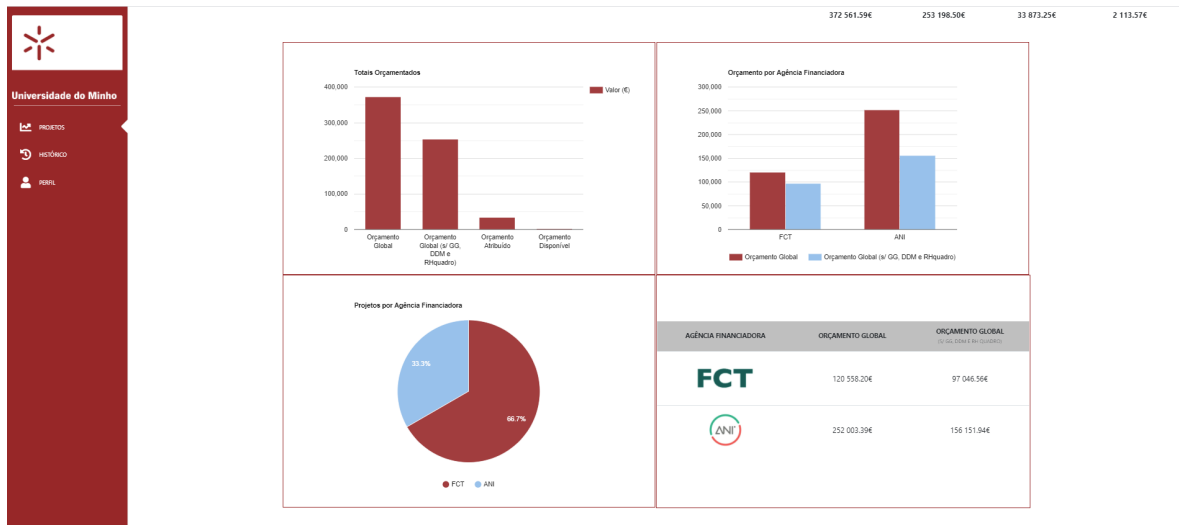


Figure 31: Researcher Main Screen (continued)

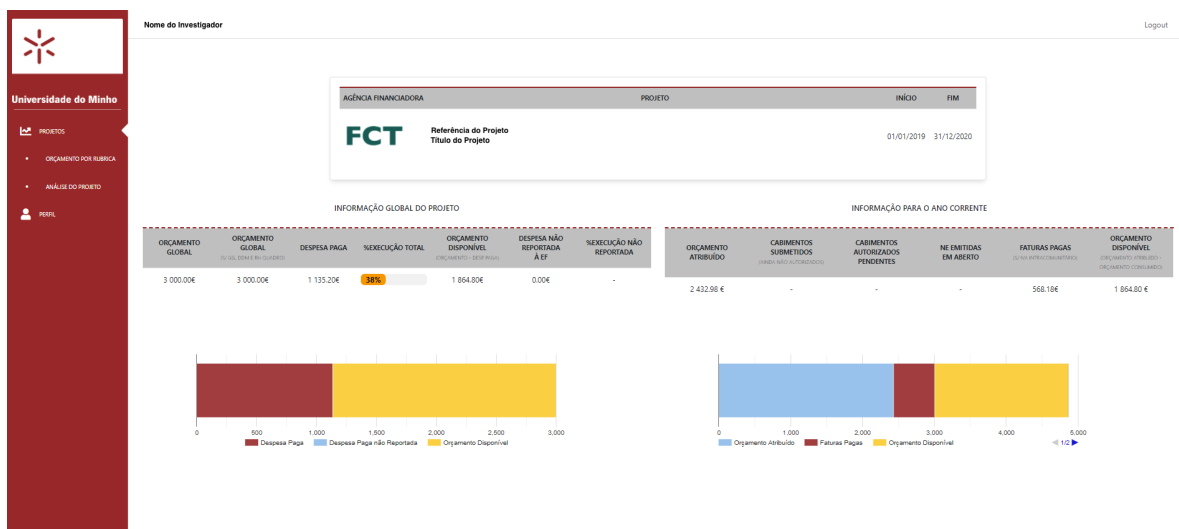


Figure 32: Main screen of a specific project

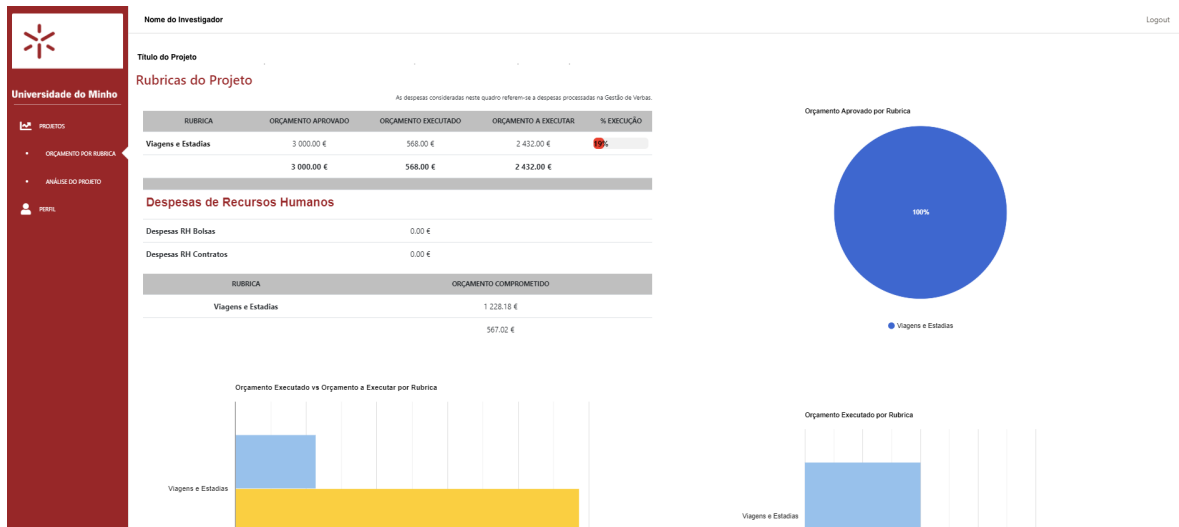


Figure 33: Screen with expenses by rubric of a specific project

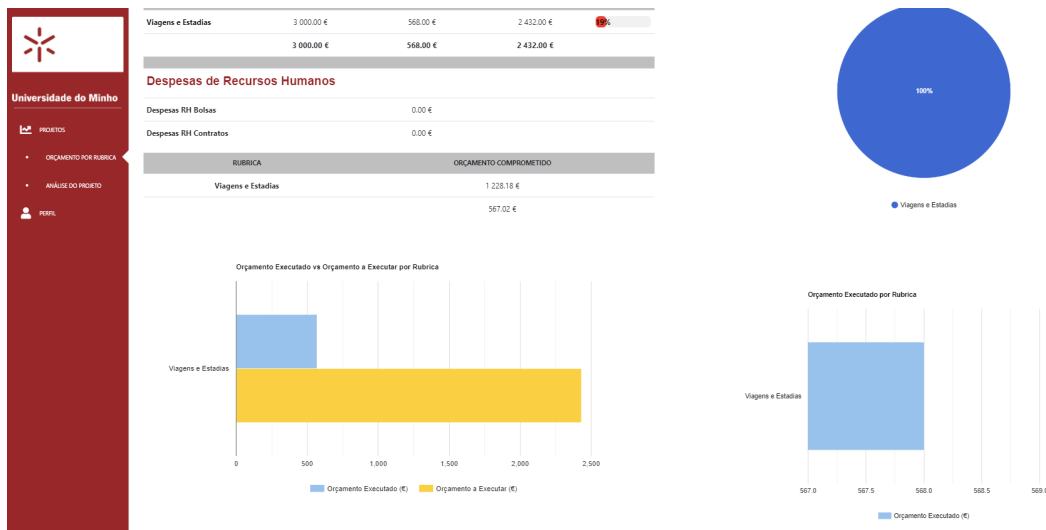


Figure 34: Screen with expenses by rubric of a specific project (continued)

Nome do Investigador Logout

Título do Projeto
Viagens e Estadias

CABIMENTO	DESCRIÇÃO	VALOR
CAB.2019.25525		
ENCERRADO	Avião: descrição da despesa	550,00 €
24/10/2019		
CAB.2019.25525		
ENCERRADO	Avião: descrição da despesa	550,00 €
24/10/2019		
CAB.2020.4494		
ENCERRADO	Alimentação: descrição da despesa	128,18 €
07/04/2020		

Figure 35: Screen with the expenses of a specific rubric

Nome do Investigador Logout

Título do Projeto
Viagens e Estadias

CABIMENTO	DESCRIÇÃO	VALOR
CAB.2019.25525		
ENCERRADO	Avião: Joel Nuno Pinto Borges; Lyon de 25/10/2019 a 30/10/2019 (Ref. AC: 2019-10189); Atividades de Investigação no âmbito da Cooperação Transnacional - Programa Pessoa - Acordo entre Portugal e a Fra	550,00 €
24/10/2019		
CAB.2019.25525		
ENCERRADO	Avião: Diogo Emanuel Carvalho Costa; Lyon de 25/10/2019 a 30/10/2019 (Ref. AC: 2019-10189); Atividades de Investigação no âmbito da Cooperação Transnacional - Programa Pessoa - Acordo entre Portugal e	550,00 €
24/10/2019		
CAB.2020.4494		
ENCERRADO	Alimentação; Diogo Emanuel Carvalho Costa; Lyon de 25/10/2019 a 30/10/2019 para a realização de atividades de Investigação no âmbito da Cooperação Transnacional - Programa Pessoa - Acordo entre Portug	128,18 €
07/04/2020		

Figure 36: Screen with the expenses of a specific rubric (continued)

Nome do Investigador Logout

Título do Projeto
Pedidos de Pagamento

Valor de Pedidos de Pagamento já validados pela Entidade Financiadora 0,00 €

Valor de Pedidos de Pagamento pendentes de análise na Entidade Financiadora 0,00 €

Receita Arrecadada da Entidade Financiadora 3 000,00 €

Total de Despesa Paga 1 135,20 €

TOTAL COMPROMETIDO NO ANO	TOTAL CABIMENTADO NO ANO	DESPESA PAGUA NO ANO	CABIMENTOS POR COMPROMETER	COMPROMISSOS POR PAGAR
568,18 €	568,18 €	568,18 €	0,00 €	0,00 €

Figure 37: Screen with payment requests and payment summary

A.2 LEVEL 2: RESEARCH CENTER

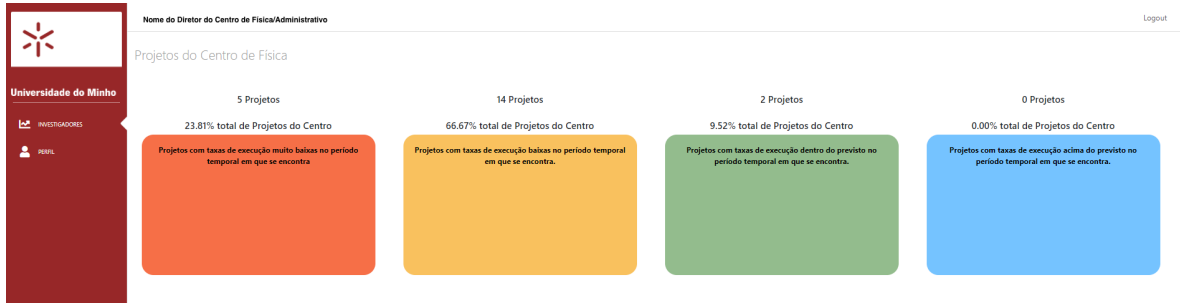


Figure 38: Research Center Main Screen

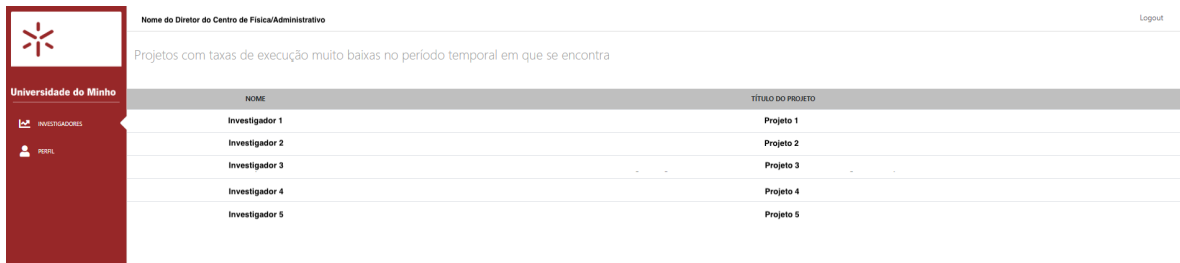


Figure 39: Screen with researchers' names and project titles with very low execution rate

A.3 LEVEL 3: ORGANIC UNIT



Figure 40: Organic Unit Main Screen

DETAILS OF RESULTS

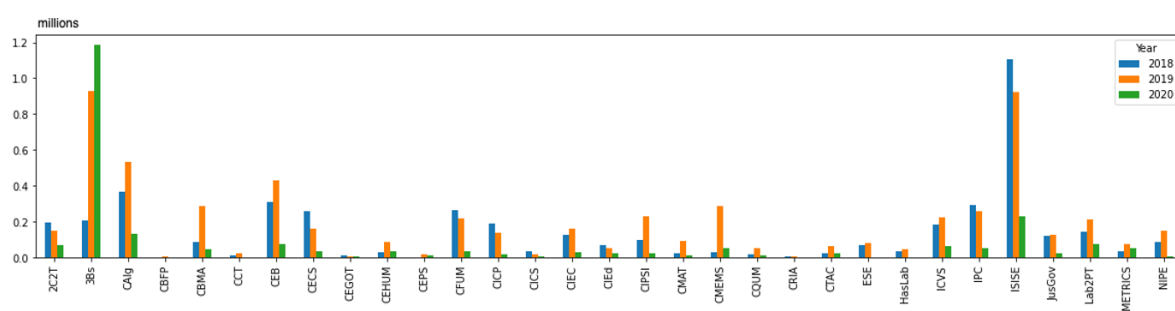


Figure 41: Expenses by Research Center: Complete Information

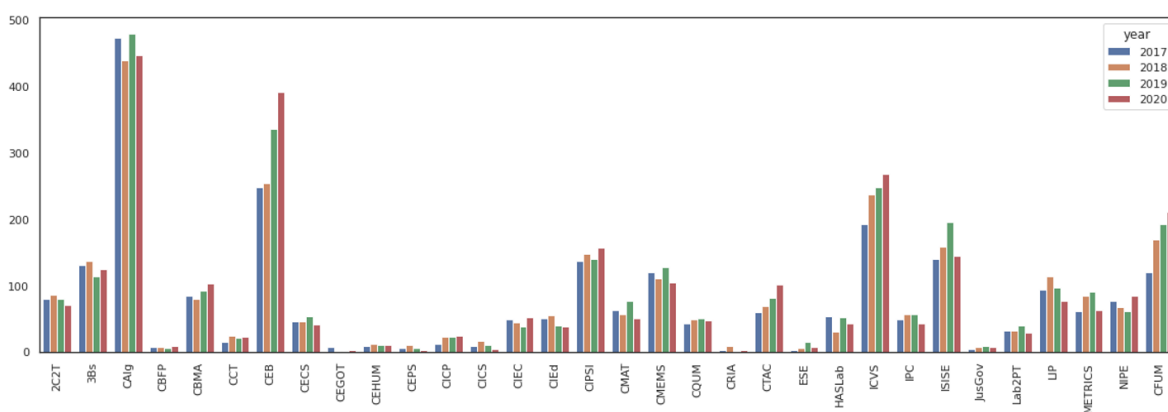


Figure 42: Amount of Scientific Production by Research Center: Complete Information

