|NST|TUTO
UN|VERS|TÁR|O
DE L|SBOA

Risk management in Data Science projects in Portugal

Ana Cristina Afonso Varela

Master Degree in Integrated Decision Support Systems

Supervisor:
PhD Luísa Cristina da Graça Pardal Domingues Miranda, Assistant Professor,
ISCTE - University Institute of Lisbon

November, 2021

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Risk management in Data Science projects in Portugal

Ana Cristina Afonso Varela

Master Degree in Integrated Decision Support Systems

Supervisor:
PhD Luísa Cristina da Graça Pardal Domingues Miranda, Assistant Professor,
ISCTE - University Institute of Lisbon

November, 2021

Acknowledgement

I would first like to thank my parents for their continuous support and incentive throughout my academic journey. To my siblings and friends for their constant words of motivation.

This work is the result of a team effort together with my supervisor, Professor Luísa Domingues, a special thanks to her for guiding me in this personal and academic challenge, for her words of encouragement and constructive criticism that kept me focused on this work.

A special thanks to all the data science professionals that were part of the expert board of my research study, thank you very much for your contributions.

Finally, to a special colleague, Sandra Gonçalves, thank you for your friendship, for sharing the laughs, the worries, and for being part of this academic journey with me.

Resumo

A grande popularidade dos projetos de dados tem influenciado muitas iniciativas de desenvolvimento com o intuito de melhorar a performance dos negócios e tomadas de decisões. Contudo, os projetos de Ciência de Dados carregam na sua essência um conjunto de riscos e incertezas específicos. Ter uma boa gestão de risco é um dos componentes mais cruciais de um projeto. A sua eficaz conduta aumenta as probabilidades de sucesso do projeto, contudo, é necessário compreender os componentes envolventes aos riscos. Neste contexto, foi conduzida esta investigação com o propósito de criar uma lista base dos riscos dos projetos de Ciência de Dados e seus fatores envolventes.

Esta investigação foi guiada pela abordagem de Design Science Research e o processo de recolha de dados foi conduzida através da técnica de Delphi, onde foi possível identificar e analisar os riscos, seus fatores, os cenários de falhas dos projetos e perceber o contributo das metodologias de desenvolvimento nesses projetos. O estudo permitiu a criação de um artefacto, que consiste em uma lista de riscos específicos relacionados com a gestão de dados e recomendações de boas práticas. Contudo, foi possível verificar que mais de metade dos riscos no topo das classificações são semelhantes aos riscos de outros tipos de projetos de IT. Esta investigação contribui com uma lista consolidada de 25 riscos dos projetos de Ciência de Dados com o intuído de auxiliar na diminuição das falhas dos projetos deste âmbito.

Palavras-Chave: Ciência de Dados, gestão de riscos, técnica de Delphi, riscos de Ciência de Dados, sucesso do projecto, Design Science Research (DSR)

Abstract

The increasing popularity of data projects has influenced many development initiatives aimed at improving business performance and decision-making. However, Data Science projects carry in their essence a set of specific risks and uncertainties. Good risk management is one of the most crucial components of a project. Its effective conduct increases the probability of project success, however, it is necessary to understand the environment and the components surrounding risks. In this context, this investigation was conducted to create a base list of the risks of Data Science projects and their surrounding factors.

This research was guided by the Design Science Research approach and the data collection process was conducted through the Delphi technique, where it was possible to identify and analyze the risks, their factors, the failure scenarios of the projects, and to understand the contribution of the development methodologies in these projects. The study enabled the creation of an artifact, consisting of a list of specific data management-related risks and best practice recommendations. However, it was found that more than half of the risks at the top of the rankings are similar to the risks of other types of IT projects. This research contributes a consolidated list of 25 risks of Data Science projects intending to help decrease the failures of projects in this area.

Keywords: Data Science, Risk Management, Delphi Technique, Data Science Risk, Project Success, Design Science Research (DSR)

General Index

Index of Tables

Index of Figures

Index of Charts

Glossary of Abbreviations and Acronyms

Glossary of acronyms

DS – Data Science

IT – Information Technology

IS – Information System

PMO – Project Management Office

PMI – Project Management Institute

PMBOK – Project Management Body of Knowledge

AI – Artificial Intelligence

ML – Machine Learning

Iot – Internet of Things

DSR – Design Science Research

BI – Business Intelligence

BD – Big Data

DM- Data Mining

DE - Data Engineering

Glossary of symbols

K - Cohen's coefficient of concordance (Kappa)

W - Kendall's coefficient of concordance

$\bar{P}$ - Observed proportion of agreements among all classifications

$\bar{P}_e$ - Expected value of $\bar{P}$ under "random" or "null" agreement

p_j - Porportion of all classifications in j categories

Chapter 1 – Introdution

1.1. Background of the topic

Organizations operate in a mutant technological environment that challenges their survival with constant economic, legal, competitive and technological changes. Managers must quickly to threats and opportunities and react proactively to the new circumstances they are in at every moment.

In the past years, researches in the field of risks in Information Technology (IT) projects has had many successful results (Raisinghani, 2004), by encountering factors that represent a threat to the success of projects and ways to manage those threats. According to Vieira et al. (2014), effective risk management has proved to be the key to project success or failure. “Risk management is the most important management tool a project manager can use to increase the likelihood of project success” (DIDRAGA, 2013). Kwak and Stoddard (2004) additionally state that risk management must be understood, in order to create value.

When planning a project, the awareness of the involved risks must be always present (Malaska & Seidman, 2018), however, analyzing and managing it in the proper way and with the right resources is a challenge for many organizations. Unfortunately, the scenario is not different for Data Science (DS).

The growth in the use of Data Science is driven by the emergence of Big Data (BD) and social media, the speedup in computing power, the huge reduction in the cost of computer memory, and the evolution of powerful methods for data analysis and modeling, such as deep learning (Kelleher & Tierney,2018). Manheim and Kaplan (2019) support the idea that the advances in the Data Science field along with the new age in computing, brought new dangers to social values, constitutional rights, and privacy. According to Shukla et al. (2021), Machine Learning (ML) has been successful in solving complex problems, however, it faces the challenge of combining data from multiple sources and ensuring its privacy and confidentiality. ML success depends mainly on the model's training and the amount, distribution, and variety of data. Risks related to information security and data breaches, increase by the variety, volume, and wide system infrastructure that supports BD applications (Joshi & Gupta, 2020). Towards technological achievements and data-driven processes and decisions, society started to see a promised future in human well-being, however, those improvements did not come without side effects.

Data Science has proven to help in companies’ decision-making (Tunowski, 2015), yet, it involves a variety of risks (organizational, social, financial, legal among others). Risks require to be evaluated and managed properly and ethically, in order to achieve projects success.

1.2. Motivation and relevance of the topic

It is already a consensus that the data are natural resources of the new industrial revolution (Taurion, 2013). The need to use data to transform a business (Robinson & Nolis, 2020) never have been so urgent and there are thousands of companies that became led by data, because of the amount of it and its potential to business leading. Data Science has become an indispensable ally for the development of society. The collection, analysis and storage of data have allowed companies to understand and improve their business.

Yet, according to Walker (2017) and Gartner (2017), 85% of DS projects fail and investigations over the years have listed various reasons. From inadequate data (Asay, 2017), poor communication, insufficient executive support (Taylor, 2017) over-complicating, narrow problem focus (Veeramachaneni, 2016) and a variety of other reasons. In effect, it is estimated that around 90% of available digital data is not being adequately used (Taurion, 2013). To reduce the high failure rate of Data Science projects, managers need better tools to assess and manage project risk (Wallace, Keil, & Rai, 2004). Vieira et al. (2014) claim that data management is regarding recognizing that during its lifecycle, data is subject to risks that can affect its proper use and interpretation.

The above-mentioned statistics reflect a worldwide scenario, however, there is still a lack of information regarding risks of DS projects and management strategies and tools used to respond to risks.

With the research problem identified, the main motivation of this study is to contribute by presenting the list of the risks and understanding the project environment in Portugal.

1.3. Research issues and objective

The primary objective of this research is to identify the risks in Data Science projects in Portugal. With this investigation, it is expected to understand the challenges of data projects, which are the most relevant risks and the consequences that arise from them and understand some of the reasons behind the failures of projects.

The purpose of this study is therefore to contribute to the knowledge building, capable of helping to develop comprehensive strategies, leading to a better approach for prevention and management of the risks in Data Science projects in Portugal.

Specific objectives

This investigation aims to answer the questions related to the risk profile in Data Science projects. The main purpose includes:

- Identify the risks encountered in Data Science projects in Portugal,
- Understand the risks factors and their outputs,
- Identify the development methodologies used in DS projects

- Identify the tools/ strategies teams use to manage risks
- Describe and discuss the risks, to present recommendations of best practices and approaches.

1.4. Methodological approach

A methodology is "a system of principles, practices, and procedures applied to a specific branch of knowledge" (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2008). Chris Hart (1998) refers to the research methodology as "a system of methods" to perform research within a field of study.

The present investigation is guided, using the Design Science Research methodology. Design Science Research (DSR) methodology is used to create and evaluate solutions for IT problems in the IT domain and organizations (Tagle, 2019) DSR attempts to improve the functional performance of the designed artifacts, through the continuous focus on the development process and performance evaluation of those artifacts (Hevner, Salvatore, Park, & Ram, 2004). This methodology helps to define a process model to follow for designing an IS solution.

Over the years, several DSR methodologies have been presented, however, consensus on the methodology of DSR has yet to be achieved. The lack of a generally accepted process and comprehensive and detailed methodology for DSR in IS may have contributed to this issue (Peffers, et al., 2006), (Alturki, Gable, & Bandara, 2013).

This study follows the DSR methodology presented by Peffers et al. (2008), which is a combination of six steps obtained from the evaluation of different methodologies presented by various authors (Attachment D). The steps are:

- Step 1 - Problem Identification and Motivation - Define the specific research problem and justify the value of a solution.
- Step 2 - Objective of the Solution - Infer the goal of a solution from the problem definition. T
- Step 3 - Design and Development - Create the artifactual solution. Such artifacts could be constructs, models, methods, or instantiations (Hevner, Ram, March, & Park, 2004). This activity includes determining the artifact's desired functionality and its architecture and then creating the actual artifact.
- Step 4 - Demonstration - Demonstrate the efficacy of the artifact to solve the problem.
- Step 5 - Evaluation - Observe and measure how well the artifact supports a solution to the problem. This activity involves comparing the objectives of a solution to actual observed results from the use of the artifact in the demonstration.
- Step 6 - Communication - Communicate the problem and its importance, the artifact, the rigor of its design, and its effectiveness to researchers and other relevant audiences, such as practicing professionals.

Following the described methodology, it was created the structure of steps/activities for this investigation, as presented in figure 1.

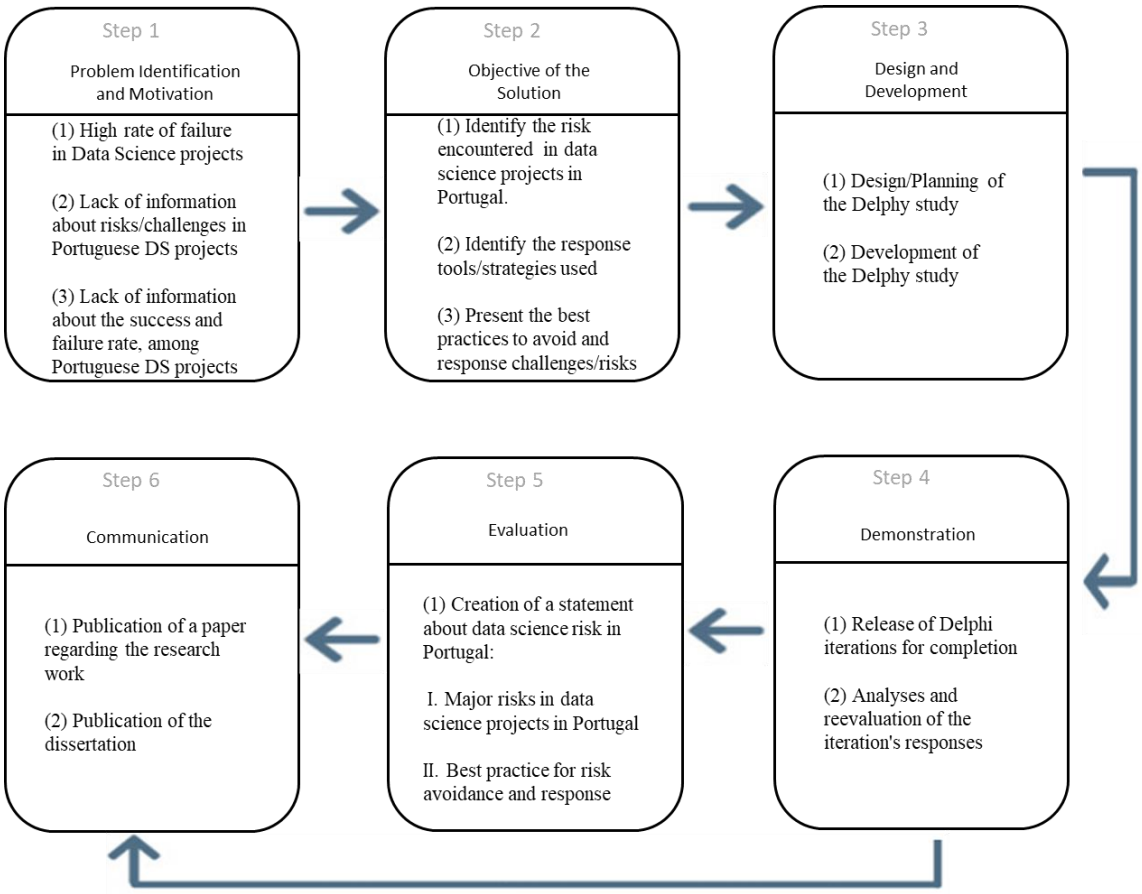

Figure 1 – DSR Methodology Steps - Adapted from (Peffers et al., 2008)

As a paradigm that creates and evaluates IT artifacts intended to solve identified problems (Hevner, Salvatore, Park, & Ram, 2004), DSR methodology provided guidelines to investigate the problem of this research methodically. The use of DSR in this research was driven by the need to understand the reasons for project failures and to understand how risks are involved in these events. To accomplish this, the first step was to understand the risk's role in those projects and how it affects the surrounding environment.

DSR is a paradigm that provides an understanding of the problem through theories, designs, and methods up to the conception of knowledge of the artifact, thus, this investigation intends to understand the problems behind the failures of DS projects by identifying the risks and other factors inherent in the projects.

For this research, the entry point is problem-centered. Supported by the research need, the process started from step 1. A more detailed description of the implemented methodology follows.

I. Problem Identification and Motivation

Over the past few decades, companies have been using data science to understand and improve their businesses. However, studies reviewed indicate that the majorities of DS projects fail. This step was supported by the literature review of the related subjects, where several online databases were selected as the literature sources such as Science Direct, Elsevier, ResearchGate, Springer-Link, IEEE, Google Academics, Semantic Scholar, and B-on.

Although the literature review provided numerous insights regarding the DS projects, it was noticed a deficit of information regarding the success and failures rates, the risks, and the challenges of the DS project in Portugal. Thus, there was a need to access the risks and challenges to create an understanding of these elements and contribute to the development of comprehensive strategies, capable of leading to a better approach for prevention and management of those risks. The problems that gave rise to the need to investigate this topic are:

- Problem 1: High rate of failure in Data Science projects
- Problem 2: Lack of information about risks/challenges in Portuguese Data Science projects
- Problem 3: Lack of information about the success and failure rate, among Portuguese DS projects

II. Objective of the Solution

The objective of this investigation is to understand the challenges of data projects, which are the most relevant risks and the consequences that arise from them and understand some of the reasons behind the failures of projects. The main purpose includes:

- Identify the risks encountered in Data Science projects in Portugal,
- Understand the risks factors and their outputs,
- Identify the development methodologies used in DS projects
- Identify the tools/ strategies teams use to manage risks
- Describe and discuss the risks, to present recommendations of best practices and approaches.

III. Design & Development

This investigation was conducted by the Delphi research approach, to gather the necessary information regarding Portuguese DS projects, the risks, the challenges, the project methodologies, the response strategies and success rates.

The Delphi method has the potential to provide a deep understanding of current issues, because it enables a set of anonymous specialists' opinions and judgments to be gathered and organized, leading to the dense and consensual outcome on the subject (Worrell, Gangi, & Bush, 2013).

This study is structured by the following steps:

- Definition of the study purpose – (Collect, analyze and identify the Portuguese DS project risks and challenges)

- Creation of the seed list questionary (based on risks and challenges in DS projects collected from literature review)
- Definition of the number of expected rounds (iterations to obtain a consensus)
- Definition of number of items in each round
- Definition of the criteria used for keeping items at each round
- Selection of specialists board for each round (best scenario, +15 – Portuguese DS professionals, who understand the DS projects scenarios)

IV. Demonstration

This stage intends to gather the necessary information regarding the risks, to create the statement of this investigation. The Demonstration step includes:

- The release of the surveys for completion/response,
- Qualitative and quantitative analyses and reevaluation of the survey's responses (After closing each iteration)
- Definition of the statistics reported for each round (the rated result of risks/challenges of previous iterations)
- The final analyses of the study and other findings:
    i. Major risks in Data Science projects in Portugal
    ii. Risk response strategy and methodologies used
    iii. The reported rate of project's success and failure

V. Evaluation

In the Evaluation stage, it is built the statement and conclusions regarding the risks information collected in the Delphi study. It includes:

- Creation of a statement regarding the risk of Data Science in Portugal, based on the conducted study
- Best practice recommendations for responding to and avoiding risk
- Limitations of the study
- Future research proposal

VI. Communication

The research method used in this investigation provided useful information, which was shared through the:

- Publication, presentation and discussion of a paper regarding risks in Data Science projects in Portugal (CENTERIS | HCist | ProjMAN 2021)
- Research/dissertation publication

1.5. Structure and organization of the dissertation

The present study is organized into five chapters that intend to reflect the different phases of the methodology until its conclusion.

The first chapter identified as “Introduction”, presents the theme of the research, which is risk in Data Science projects in Portugal. This chapter describes the problem, the main goal and motivation for this study, and a brief description of the structure of the work.

The second chapter reflects the theoretical framework, called “Literature Review”. In this chapter, the state of the problem is investigated. The literature review is divided into two approaches. The first makes an introduction to the theme of Data Science and project management and the second makes a review of the related work/studies. This chapter presents the meaning and purpose of the Data Sciences's different areas how are they managed worldwide, some statistics regarding the success rate of those projects, and finally, the listing of the challenges and risks of Data Science projects.

The third chapter designated “Design and Development”, is dedicated to the methodology used in the process of data collection and processing as well as the methods of analysis used to create the understanding of the topic in the study. In this section, it is described the method (Delphi Study) used to gather the information needed to state this study. This chapter describes the Delphi study structure, including the purpose, participants, the platform used and date organization.

The fourth chapter “Demonstration” is reserved for the release of the Delphi study’s rounds and the analysis of the results obtained from it. This session contains the other findings of the interviews/meetings and presents structured information regarding risks in Data Science projects in Portugal.

The fifth chapter “Evaluation”, presents the main conclusions of this study as well as recommendations of the best practices for avoidance and control of the encountered risks, limitations of the study and future work proposal.

Chapter 2 –Literature Review

2.1. Data Science

Before getting into Data Science, it is important to understand what data is.

OAIS (2012) defines data as "A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen". There are three types of data, structured, unstructured and semi-structured data. Structured or organized data refers to data that is sorted into a row/column structure (Ozdemir, 2016). It is usually disciplined, predictable and repeatable (Inmon & Nesavich, 2017). Unstructured or unorganized data is the type of data that is in free form, usually text or raw audio/signals, e-mails, documents that must be parsed further to become organized (Inmon & Nesavich, 2017). Semi-structured data are data that do not reside in a relational database, however, has some organizational properties and internal markings that facilitate its analysis and categorization (Taulli, 2019). They are data organized in a heterogeneous and irregularly structured way (Bertocchi, 2016). XML and JSON files, fall into this category. It is estimated that semi-structured data represents only 5% to 10% of the overall data (Taulli, 2019) and unstructured data, represents 80 to 90% of the world's data (Ozdemir, 2016).

So, what is Data Science?

According to Kelleher et al. (2018), Data Science encompasses a set of principles, problem definitions, algorithms, and processes for extracting non-obvious and useful patterns from large data sets. Ozdemir (2016) defines it as the "art and science of acquiring knowledge through data". Data Science addresses "the nontrivial extraction process of implicit, previously unknown, and potentially useful information underlying large amounts of data" (Frawley, 1992).

Data Science is all related to how the data is obtained processed and used to acquire insights and knowledge. According to Arabnia et al. (2020), the benefits of DS include customization of service, knowledge of the target market, assertive methods of analysis, creation of focused digital strategies, agility in decision-making and others.

The concept of Data Science has a similarity to the concept of the DIKW framework. DIKW (data, information, knowledge, wisdom) hierarchy known as well as knowledge pyramid or wisdom hierarchy is a hierarchical sequence, which from data via information and knowledge ends in wisdom (Mikos, Tiwari, Yin, & Sassa, 2017). DIKW, originally IKW presented in 1934 by T. S. Eliot, is a concept based on the assumption that data is used to create information, information is used to create knowledge, and knowledge is used to possibly obtain wisdom (Bratianu, 2015).

Over the years, respecting the basic segments of the DIKW model, authors have proposed adjusts and customization of this model according to their scientific needs and interest (Mikos, Tiwari, Yin, &

Sassa, 2017). Ackoff (1989) proposed a model with the addition of Understanding between Knowledge and Wisdom, as shown in figure 2 (left).

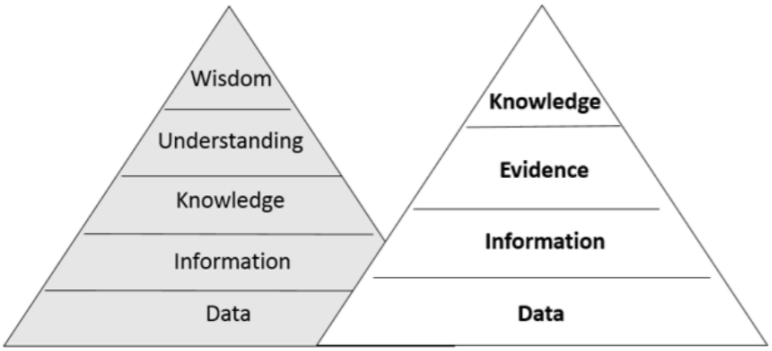

Figure 2 - Ackoff's Knowledge (DIKW) hierarchy (left) and DIEK (right) - source: (Dammann, 2018)

As seen in figure 2, Data represent the beginning of the process, it is the most basic level. Data are symbols that represent the properties of objects and events (Ackoff, 1989) and it does not have the meaning of themselves (Bellinger, Castro, & Mills, 2004). Information consists of processed data and represents its usefulness, it adds context (Jifa, 2013). Knowledge is conveyed by instructions, Understanding is conveyed by explanations, and Wisdom deals with values. Ackoff (1989) considers that wisdom involves the exercise of judgment. Bellinger et al. (2004) restructured the idea of Ackoff's DIKW model. They consider that Understanding is present in every step of the framework and supports the transition from each stage to the next, as presented in figure 3.

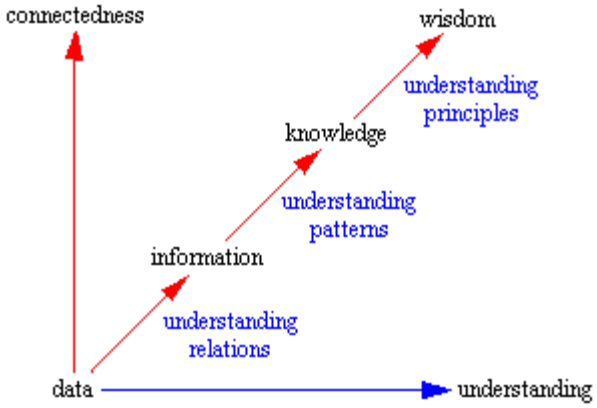

Figure 3- DIKW Model, source: (Bellinger et al., 2004)

Dammann (2018) proposed an adaptation of Ackoff's DIKW hierarchy designed to Data Science, which dropped the notion of the Wisdom stage and inserted Evidence between Information and Knowledge, as shown in figure 2 (right). He justified the removal of Wisdom, as he considers that judgment is a needed element at all levels of the hierarchy and Wisdom does not add much to the decision-making. Dammann considers that in the context of DS Evidence is used in the analysis and hypothesis-testing process, to support claims/hypotheses and decision-making and Knowledge is evidence-based belief. "Evidence is information that can be used to support a hypothesis by testing it. Thus, all evidence is information, however, not all information is evidence" (Dammann, 2018).

O'Neil and Schutt (2014) support Drew Conway's point of view regarding Data Science, which stands by the understanding of three basic areas: hacking skills, math and statistics knowledge, and substantive expertise, as is illustrated in Figure 4.

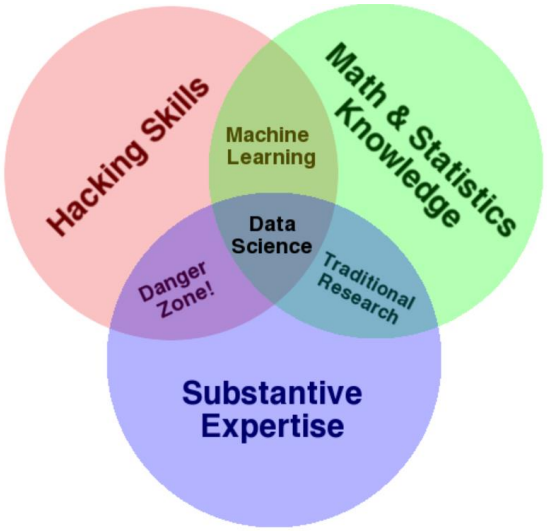

Figure 4 - Drew Conway's Venn diagram of Data Science - source: (Conway, 2010)

The diagram illustrates the key components of Data Science. According to Ozdemir (2016), those with hacking skills can program complex algorithms. Math and statistics use equations and formulas to perform analysis. They allow the evaluation of the algorithms and create a specific procedure to suit the condition. The domain knowledge/expertise allows the applications of the area concepts effectively.

Data Science is the intersection of those three areas. O'Neil and Schutt (2014) state that Data Science is not merely hacking, "its acolytes possess a practical knowledge of tools and materials, coupled with a theoretical understanding of what is possible". To obtain knowledge, it is necessary to use computer programming to reach and handle the data, comprehend the math and statistics behind the models and understand the analyses achieved according to the business area in the study.

Data Science requires various skills and involves different areas. From Big data, Data Mining, Artificial Intelligence, Deep Learning, Machine Learning, Data Analytics, Business Intelligence and everything that involves extracting knowledge from data (Robinson & Nolis, 2020).

Contemplating several fields, Data Science involves numerous types of processes, they include the collection, staging, storage, modeling of datasets, creation of reports, creation and execution of machine learning models, and so forth (Malaska & Seidman, 2018).

2.2. Project Management

A project is "a temporary endeavor undertaken to create a unique product, service, or result (PMI, 2017).

A project is composed of many features, which determine its category, complexity, area and so on.

Schwalbe (2016) and Luckey & Phillips (2006) consider scope, time and cost as the three-base constraints of a project. These limitations are sometimes introduced as the triple constraint or iron triangle, as illustrated in figure 5.

Figure 5 - Project constraints (Iron Triangle), source: (Luckey & Phillips, 2006)

The scope constraint determines what work is planned to be performed/done or what are the results expected from the developed product or service, the time constraint determines how long it takes to complete the project and who can change the projects schedule and deadlines, and the cost constraint determines the project budget and everything related to it. Luckey & Phillips (2006) state that to achieve quality in the overall project, the Iron Triangle must remain balanced.

As it can be seen, in Figure 5, to accomplish a successful project, it is required to adjust the constraints. As Schwalbe (2016) referred, projects can be large or small and involve one person or thousands of people, yet, it is demanded to accomplish those three main features by following good project management. PMI defines project management, as the application of knowledge, skills, tools, and techniques to project activities to meet project requirements (PMI, 2017).

According to a Gartner report (2017), only between 15% and 20% of DS projects are completed and around 8% of them generate value. As reported by Forbes (2020), only 15% of leading firms have deployed Artificial Intelligence (AI) capabilities into production. A Capgemini study (2014), reported that only 27% of respondents described their BD initiatives as “successful” and only 8% described them as “very successful”. Although many organizations implement management processes into their projects, not all projects are successful. Factors such as time, money, and unrealistic expectations, among many others, are capable of sabotaging a promising project if it is not properly managed (Schwalbe, 2016). The basis for any successful data project is a clear understanding of what it is tasked to build and then understanding the major items that are needed to consider to design a solid solution (Malaska & Seidman, 2018).

2.2.1. Risk in IT projects

The concept of risk can be linked to uncertainties associated with events (Ayyub, 2003). Usually, risks are associated with bad events, although some risks cause very positive impacts and create opportunities.

All projects involve risk, which means that pursuing a project without any risk is worthless. Uncertainties and risks managed in the right way can bring opportunities to the project (The Standish Group, 2015). Any data project brings with it a set of risks (Malaska & Seidman, 2018) and for decades, several frameworks and tools were created to explain the different types of IT risks, risk management strategies and measures of software projects performance. Risk management is the most important management tool a project manager can use to increase the likelihood of project success (DIDRAGA, 2013). The management approach answers the question of how to handle risks to prevent loss or damages.

Bakker et al. (2011) state that the importance of the use of risk management techniques and provide project managers with guidelines on how to apply risk management within their projects. The authors state that, in addition to an instrumental effect, risk management can contribute to project success through communicative effects. Risks assume different forms and can affect various subjects of a project, from financial to operational, functional and technical issues (Rekha & Parvathi, 2015).

Bakker et al. (2011) conclude that risk identification is considered the most influential risk management activity of all, followed by risk reporting, risk registration and risk allocation, risk analysis, and finally risk control.

2.3. Data Science project management challenges and risks

Businesses have long used data analytics to help direct their strategy to increase profits and support their decision-making processes (Chen, Chiang, & Storey, 2012). However, creating value from data requires a range of talents, from data integration and preparation to designing specialized computing/database environments, to data mining and intelligent algorithms (Gartner, 2014). The first step to conduct risks, recommended by many project management experts, is to understand them and their drivers. To improve hindsight to foresight, leaders need to better understand the types of risks is been taken, their interdependencies, and their underlying causes. Cheatham et al. (2019) state that making real progress demands a multidisciplinary approach involving leaders across the company, experts in areas ranging from legal and risk to IT, security, and analytics, and managers who can ensure vigilance at the front lines. As Data Science englobes diverse fields and types of processes, it is expected that different types of challenges arise from each field.

This chapter, therefore, aims to discuss opportunities and challenges which arise throughout the use of Data Science.

2.3.1. Big Data

The term Big Data is frequently used in corporate practice, as well as in scientific research, however, the term is often interpreted diversely. To date, the literature presents no uniformly recognized definition of Big Data (Gärtner, Hiebl, & R.W., 2018).

Taurion (2013) states that BD handles a huge amount (**volume)** of structured and unstructured data (**variety)**, inside and outside companies (social media data, for example), that needs to be validated and accurate (**veracity**) and treated at the appropriate speed (**velocity**) to create **value** for the business. Gartner (2014) defines BD as the term adopted by the market to describe problems in the management and processing of extreme information that exceeds the capacity of traditional information technologies over one or more dimensions. Gartner and Hiebl (2018) presented a definition of Big Data that combines different definitions from literature in one: "Big Data refers to the generation, storage, processing, verification and analysis of large, highly versatile and quickly growing volumes of data to create valuable information".

Big Data may offer firms and other organizations various opportunities (Hashem, et al., 2015). Especially in marketing, new analysis options arise when, for example, considering consumer behavior and interacting with customers. Besides some opportunities, different challenges and risks surrounding Big Data have additionally been explored in the literature. Data sets with many possible input variables bring essentially high risk and falsely identify an input variable as important (Graaf, 2019). Those events occur because most companies do not have a clear vision of what BD is, its potential and how to leverage this potentiality (Taurion, 2013). Davenport (2014) refers that although the new and extended options for analysis may be innovative, challenges may further present themselves, such as the automation of management tasks (Frey & Osborne, 2013). As with any business initiative, a BD project involves an element of risk. To Jurney (2014) the extensive skill set needed to build data products represents both an opportunity and a threat. A rich skilled team is able to decompose the client's problem into parts and solve it, however, as the team grows up to satisfy the project's needs, communications problems start to arise. Large team meetings are unlikely to be productive and the extensive hierarchy may interfere with the perception of the customer's most urgent problem.

According to Gartner et al. (2018), larger quantities of data do not necessarily lead to better decision-making. The risk arises when the most relevant or correct information fails to be retrieved from existing data (Gärtner, Hiebl, & R.W., 2018) and as reported by Taurion (2013), it is estimated that around 90% of available digital data is not being adequately used. At some point when the volume, variety and velocity of the data are increased, the current techniques and technologies may not be able to handle the storage and processing of the data (Jadhav D. K., 2013) and it additionally brings issues regarding the cost of its maintenance and the security providence (Raguseo, 2018). Joshi and Gupta (2020) stated that risks related to information security and data breaches, increase by the variety, volume, and wide system infrastructure that supports Big Data applications. Kshetri et al. (2017) stated that In

some cases, BD may challenge the principles of privacy on which modern laws are based, and if consumers' data are not handled ethically and correctly by organizations, issues related to civil rights violations could arise.

With the large, unlimited amounts of data available, the selection of data samples, data filtering and the subsequent profitable usage of data present considerable challenges (Tan, Zhan, Ji, Ye, & Chang, 2015), as the higher concentration of data makes a more appealing target for hackers and cybercriminals (Kshetri, Fredriksson, & Torres, 2017). On the demand of velocity, real-time data from the Internet navigations are constantly been collected, stored, and reused, and this process poses a significant privacy risk.

Another challenge is the lack of understanding of what Big Data represents and its potentials and limitations, which can create risks for the business (Taurion, 2013), as well as the lack of qualified personnel/ experts (Gärtner, Hiebl, & R.W., 2018). Hazen et al. (2014) emphasize that a core challenge in the use of Big Data lies in the verification of the reasonableness, correctness, immediacy, completeness and format of the used data. Taurion (2013) and Accenture (2014), identified some risks and challenges companies must take special attention, they are:

- Data security failure, privacy loss and uncontrolled access to confidential data. Finding a robust security mechanism for a data storage system is a challenging problem in BD (Jadhav D. K., 2013). In terms of privacy risks, it is sometimes not clear who is the owner of the data and using the data without the right legal foundation or consent may cause serious legal problems (Ernst & Young, 2014).
- Budget
- Lack of talent to implement big data - Lack of expertise and skilled professionals to run Big Data and analytics on an ongoing basis
- Integration with existing systems - Such problems can lead to unnecessarily long times-to-implementation of Big Data technologies, especially when decisions are not made in real-time or are insufficiently delegated (King, 2014).
- Enterprise not ready for Big Data - Predictive analyses can generate results that may question some business perspectives. Big Data demands knowledge of new technologies and mainly changes in the mindset of the company.
- The analytical capacity to translate data into information and knowledge is another challenge. It requires much more sophisticated training and visualization tools.

Big Data has the potential to transform economies, creating a new wave of economic productivity, yet, only makes sense if the value of the data analysis offsets the cost of its collection, storage and processing (Taurion, 2013).

2.3.2. Business Intelligence

As it became a real concept, various definitions of Business Intelligence (BI) have emerged in the academic and practitioner literature (Isik, Jones, & Sidorova, 2012). Some authors have a simple definition of BI as "a holistic and sophisticated approach to cross-organizational decision support" (Moss & Atre, 2003), others approach it as "a system comprised of both technical and organizational elements that present its users with historical information for analysis to enable effective decision making and management support, with the overall purpose of increasing organizational performance" (Isik, Jones, & Sidorova, 2012). Business intelligence covers a wide range of applications and practices for the collection, integration, analysis, and presentation of business information, with the most important goal to support organizational learning and better business decision making (Bara & Knezevic, 2013). Based on those concepts, in simple words, it can be said that Business Intelligence refers to a broad concept that is designed to support and improve decision-making based on hard data that leads to an increase in the efficiency of an organization.

BI application provides companies with the means to gather and analyze data that facilitates reporting, querying and decision-making (Raisinghani, 2004), by permitting information recovery, examination and clarification of the business need (Nofal & Yusof, 2013). Its systems have become an integral part of running a business in the twenty-first century due to the constantly increasing needs of organizations in the field of analysis, interpretation and data processing (Tunowski, 2015). Organizations have implemented BI to achieve a variety of organizational benefits, however, BI success is defined differently by different organizations, depending on the benefits expected from the BI initiative (Isik, Jones, & Sidorova, 2012).

Although the main purpose of Business Intelligence is to support better and faster business decisions (Balachandran & Prasad, 2017), it has its own challenges. The list below represents some of the challenges and risks identified in the literature. They are:

- According to CMBI (2020), the challenge for a BI project is to combine the business problem, BI tools, and data without the normal overhead associated with bespoke application development.
- Ambiguity - Hasan et al. (2015), report that some organizations tend to adopt BI with ambiguous objectives, allied with poor business and data management and limited funding and expertise, leading to a great failure.
- Training and user acceptance. It happens when the management is lacking in providing the training and support for BI solutions.
- Improper change management during the process of the BI systems implementation (Tunowski, 2015).

2.3.3. Artificial Intelligence

According to Manheim and Kaplan (2019), the main goal of AI is to filter the noise, find meaning, and act upon it, ultimately with greater precision and better outcomes than humans can achieve on their own. The finance and banking industry also make use of artificial intelligence to ensure they can monitor various activities that take place (Nadimpalli, 2017).

As stated by Cheatham et al. (2019), nearly 80 percent of executives at companies that are deployed AI told that they saw moderate value from it. The same authors additionally referred to AI as a sword of double edge, when considering that even as it generates consumer benefits and business value, it is also giving rise to a host of unwanted, and sometimes serious, consequences.

AI has encountered numerous advantages to the technologic processes, however, there are enormous concerns regarding the consequences of its failure.

Cheatham et al. (2019) presented a few challenges and risks of AI, they are:

- Lack of labeled data – this point was furthermore affirmed by (Freedman, 2017), as it says, "the data simply does not exist in the right format—or in any form at all. Or the data may be scattered throughout dozens of different systems, and difficult to work with."
- Uni-task orientation of weak artificial intelligence
- Affordability of required computational expenses
- Adversarial attacks – The shakiness of deep decisions
- Lack of transparency and interpretability.
- Data difficulties
- Technology troubles.
- Security snags.
- Models misbehaving (Algorithms failure).
- Human-machine interactions (Interaction issues).
    - Scripting errors, lapses in data management, and misjudgments in model-training data easily can compromise fairness, privacy, security, and compliance.

Freedman (2017) considers health care as one of the hottest segments of the market for machine-learning technologies. As an example, Cheatham et al. (2019) quote: "if an AI medical algorithm goes wrong, or it compromises the national security, or if an adversary feeds disinformation to a military AI system, it represents significant challenges for organizations, from reputational damage and revenue losses to the regulatory backlash, criminal investigation, and diminished public trust".

Even though companies are making huge investments to improve service quality by training the system to recognize medical pathology, the AI failures in healthcare are reported to not necessarily be associated with technology failures. Instead, with difficulties in deploying AI tools in practice (Tizhoosh & Pantanowitz, 2018), uncertainties and unplanned costs (Raguseo, 2018).

Additionally to the medical field, Tizhoosh and Pantanowitz (2018) listed other risks specific to the implementation of AI in this field. They are:

- Pervasive variability – the risk of the algorithm will not understand/ recognizing the right type of tissue.
- Non-boolean nature of diagnostic tasks - A pathology diagnosis employs several processes including cognition, understanding clinical context, perception, and empirical experience. Binary language may only be desirable in easy, obvious cases.
- Dimensionality obstacle - Downsampling images for better networking, may result in loss of crucial information. On the other hand, a deep network with a large input size can be difficult or even impossible to train.

2.3.4. Data Mining

Data Mining (DM) is the process of extracting information from large volumes of data to improve decision-making (Kelleher & Tierney, 2018). The real-world data is diversified, incomplete and noisy, and as time goes by, the type of data became more complex and the scale of data ever larger (Yang, 2010). Data in large quantities normally are inaccurate, unreliable and messy.

Even though DM is very powerful, it faces many challenges during its application. Those challenges might be related to performance, data, methods, technologies used and other elements. The data mining process succeeds when the issues are identified and fixed correctly. "Extracting value out of data is not a trivial task. One of the key elements of any such 'making sense out of data' program is the people, who must have the right skills and capabilities." (Gartner, 2014).

2.4. Risk structure

Based on related studies consulted, it was possible to collect and list the major risks encountered in DS projects. Most of the risks (e.g. lack of qualified experts on board, inadequate/lack of information, budget deficiency, communication, wrong requirements, and others), apply to all DS fields investigated in this study. Which are Big Data, Data Analytics, Business Intelligence, Artificial Intelligence (Machine learning) and Data Mining.

The risks listed for the development of this study followed the risk breakdown structures approach. A source-oriented Risk Breakdown Structure (RBS) is a hierarchical approach that recognizes that risk can be identified and assessed at several levels (Norris, Perry, & Simon, 2000). The RBS employed in this study is the one presented by (PMI, 2017), as illustrated in Attachment A. they are:

- Technical risks
    - Scope /Requirements definition
    - Estimates

    - Technical processes
    - Technology
- Management risks
    - Operations (quality) management
    - Project Management
    - Organization
    - Resourcing
    - Communication
- External Risks
    - Legislation/legal
    - External (Environmental, facilities)
    - Regulatory
    - Surrounding Environment

The table below (Table 1) represents the risks identified in the literature review and organized by the applicable category.

Table 1- Risks of Data Science Projects from literature review

Category	Risk of Data Science	Selected references
Scope	Begin with the wrong questions	(Plugar, 2018) /
	Lack of awareness regarding future changes	(Lewis, 2015)
	Focus on the wrong problem	(Plugar, 2018)/ (Hasan, Rahman, & Lahad, 2015)
Estimates	Poorly defined estimates	(Balachandran & Prasad, 2017)
Technology	Security glitch / Internal or External Attacks (cyber-attacks)	(Accenture, 2014)/ (Liulka, 2017)/ (Ulrich, Frank, & Timmermann, 2020) / (Vieira, Ferreira, Barateiro, & Borbinha, 2014)
	Data security negligence	
	Technological insufficiency/deficiency	(Malaska & Seidman, 2018)
	Digital vulnerability	(Cheatham, Javanmardian, & Samandari, 2019)
	Hardware/Software Faults and Obsolescence	(Vieira, Ferreira, Barateiro, & Borbinha, 2014)
	Theft/leakage of information	
Technical Processes	Model instability or performance degradation Model (failed Algorithms)	(Cheatham, Javanmardian, & Samandari, 2019)

	Complex projects	(The Standish Group, 2015)
	Wrong Methodologies/ tools	(Jadhav D. K., 2013)
	Interaction issues (Human vs machine)	(Cheatham, Javanmardian, & Samandari, 2019)
	Limited data storage structure (Storage overload)	(Balachandran & Prasad, 2017)
Organization	Wrecked reputation integrity	(Cheatham, Javanmardian, & Samandari, 2019)
	Management failure	(Tunowski, 2015)
	Lack of Security awareness of the employees	(Ulrich, Frank, & Timmermann, 2020)
Project management	Deadlines difficult to meet/short time	(Balachandran & Prasad, 2017)
	Negligence in risk planning	
	Complex projects	
Resoursing	Budget undervaluation	(Cheatham, Javanmardian, & Samandari, 2019)/ (Liulka, 2017)/
	Inability to reinforce funding	(Vieira, Ferreira, Barateiro, & Borbinha, 2014
	Start project without proper specialists on board	(Liulka, 2017)/ (Hasan, Rahman, & Lahad, 2015)
	Lack of talent (Lack of knowledge in data intelligence)	(Accenture, 2014)/((Ulrich, Frank, & Timmermann, 2020)
Communication	Communication issue	(Bakker, Boonstra, & Wortmann, 2011)/ (Jadhav D. K., 2013)
	Requirement misunderstanding	(Rekha & Parvathi, 2015)
	Lack of transparency and interpretability	(Tizhoosh & Pantanowitz, 2018)/ (Jadhav D. K., 2013)/(Manheim & Kaplan, 2019)
	Group think (acceptance of one's opinion without any real conviction regarding it)	(Lewis, 2015)
Operations (Quality)	Loss of Authenticity of Data	(Vieira, Ferreira, Barateiro, & Borbinha, 2014)
	Gap in Decision making	(Bara & Knezevic, 2013)
	Improper Analytics	(Liulka, 2017)/ (Waterman & Bruening, 2014)/ (Jadhav D. K., 2013)
	Lack of concern regarding data quality	(Lewis, 2015)/ (Tizhoosh & Pantanowitz, 2018) / Vieira et al. (2014)
	Human Operational Errors	(Vieira, Ferreira, Barateiro, & Borbinha, 2014)

	Corruption of collected data/ failure in data collection	(Waterman & Bruening, 2014)/ (Malaska & Seidman, 2018)
	Flaws in data entry	(Waterman & Bruening, 2014)
	Inadequate data/ missing of information	(Ulrich, Frank, & Timmermann, 2020)
	Complex Data	
	Unexpected/ incorrect or misleading results	(Jadhav D. K., 2013) / (Waterman & Bruening, 2014)
	Model Instability/Degradation (Algorithm Failure)	
	Very complex models	
	Use of inappropriate tools/techniques for data collection/processing	
Legislation /Regulatory	Unintended illegality	(Lewis, 2015)/(Jadhav D. K., 2013)
	Data privacy/ Confidentiality violation	(Liulka, 2017) / (Waterman & Bruening, 2014) / (Manheim & Kaplan, 2019)
External	Natural Disaster	(Vieira, Ferreira, Barateiro, & Borbinha, 2014)
	Accidents / physical loss (e.g. fire, hardware theft)	
Surrounding Environment	Political instability	(Cheatham, Javanmardian, & Samandari, 2019)
	Economic and Social Disruption	(Osoba & Welser, 2017)
	Lack of human employment	(Nadimpalli, 2017)/
	Diminished Resilience (loss of human skills due to automation)	(Osoba & Welser, 2017)

As found through the literature review, DS projects involve many risks and challenges. The authors defend the usefulness and advantages that Data Science technologies have brought to the automation of processes, improvement of business, and decision-making, however, they point out the challenges inherent to the applicability of these technologies and processes. Many of the risks identified are transcendent to the entire IT area, such as functional, communication, legal, economic/social, and financial risks. However, Data Science projects face some other risks specific to the area.

As mentioned by Ozdemir (2016), much of the data being studied in these projects are unstructured, in this regard, having a competent team to ensure the understanding, labeling, processing, and quality

of the data is a challenge. This challenge constitutes the risk of poor quality data and information processed by the models, negatively influencing the business understanding and decision-making.

Chapter 3 – Design and Development

3.1. Research method

To Pardal & Correia (1995) methodology is the "(...) research guiding body which, following a system of norms, makes the selection and articulation of techniques possible, to be able to develop the process of empirical verification".

Research methods are specific procedures for collecting and analyzing data. There is a variety of approaches capable of supporting research data collection. Nevertheless, each research has its particularity and there is a need to select suitable methods. To obtain the intended data for this investigation, which consists of information regarding risks of DS projects in Portugal, it was conducted a Delphi study. The Delphi approach is a technique used to obtain the most stable consensus from a group of experts on a given subject. This technique seemed convenient for this study because it can provide a useful discussion on the subject, through a set of individual opinions capable of highlighting the real issues of DS projects in Portugal.

This section consists of four parts, which is structured by the following concepts:

- The introduction to the Delphi study – Method definition, feature and advantages,
- The Delphi method objective - considering points relating to the question of consensus,
- The Design and Planning of the study undertaken - Which details the information background, questionaries development, the method to choose the participants, the time and the platform used for conducting the study surveys,
- The Development and preparation of the study – Which includes the participants' selection, rounds and discussion topics preparation, and the schedule definition.

3.2. Delphi Study

“Delphi may be characterized as a method for structuring a group communication process. Therefore, the process is effective in allowing a group of individuals, as a whole, to handle a complex problem” (Linstone & Turoff, 2002).

The Delphi method originated in the early 1950s at the RAND Corporation (Dalkey & Helmer, 1963) and the name came from the Greek mythology “Delphic” oracle, who could predict future events (Buckley, 1995). The objective was to develop a technique to obtain the most reliable consensus of a group of specialists/experts and since its introduction, researchers have developed variations of the method (Okoli & Pawlowski, 2004). Forecasting and issue identification/prioritization represent one type of application of the method. Most of the Delphi efforts during the first decade were for pure forecasting, including both short- and long-range forecasts (Okoli & Pawlowski, 2004).

The method requires knowledgeable and specialist contributors individually and anonymously responding to questions and submitting the results to a central coordinator (Grisham, 2009), (Cooper,

Gallegos, & Granof, 1995). The use of Delphi provides an opportunity to invite a specialist to comment on and discuss the questions asked and to supply reasons for their responses. A specialist panel in a Delphi study is a group of individuals deemed experts by either objective (e.g., job title, work experience, organizational affiliation) or subjective (Worrell, Gangi, & Bush, 2013). Those who seek to utilize Delphi usually recognize a need to structure a group communication process to obtain a useful result for their objective (Linstone & Turoff, 2002). This technique involves repeated individual questioning by interview or survey and avoids confrontation of the specialists with one another (Dalkey & Helmer, 1963). Buckley (1995) defends that, one clear use of the Delphi technique is when the issue under investigation benefit greatly from subjective judgments on a collective basis.

3.3. Objective (Consensus question)

Data Science has been used and needed more than ever (Provost & Fawcett, 2013) and to ensure quality business decisions it must provide information based on reliable, fresh and precise data and knowledge in the right time (Gartner, 2014). In light of these reasons, most companies develop Data Science projects to obtain the best out of their data. On the other hand, these projects tend to be very complex, with a high level of risk, and the need to involve a big number of actors with various skills (Sfaxi & Aissa, 2020). To achieve the expected result, it is a need to find, remove or reduce the risks, which threaten the project's success (Norris, Perry, & Simon, 2000).

This research aims to identify the risks in Data Science projects in Portugal, and the lack of literature information regarding Portuguese DS projects conducted to the development of this Delphi study. The primary purpose of this Delphi study is to answer the question, "What are the risks of Data Science projects in Portugal?". Additionally, this study aims to identify the risks factors and their impacts, the strategies and tools companies/teams use to respond and avoid risks, as well as the development methodology to better understand the outcome of the Portuguese projects and contribute to the awareness regarding the risks in this field, through the Delphi study and the research outcome.

3.4. Design and Planning

3.4.1. Rounds

The study was conducted in three phases to collect information regarding the risks, their factors and outputs, with the perspective of obtaining a consensus on their occurrence among the different areas of DS. Brooks (1979) identified consensus as "a gathering of individual evaluations around a median response, with minimal divergence". The general guideline is 3 rounds before panelist fatigue becomes an issue (Okoli & Pawlowski, 2004) and can considerer a strong consensus reached when Cohen's coefficient of concordance indicated $K >= 0.7$ (Fleiss, 1971) or Kendall's $W >= 0.7$ (Schmidt, Lyytinen, Keil, & Cule, 2001).

The figure below (Figure 6), presents the structure designed to conduct the rounds with the experts. This study was conducted in three phases. In each, the expert was presented with a list of risks for identification/classification in the form of a survey/questionnaire, and after the closure of each of these phases, an analysis of the responses and preparation for the next phase was conducted, based on the results obtained. Except for the third phase where the analysis of the results was the closing point of the Delphi study.

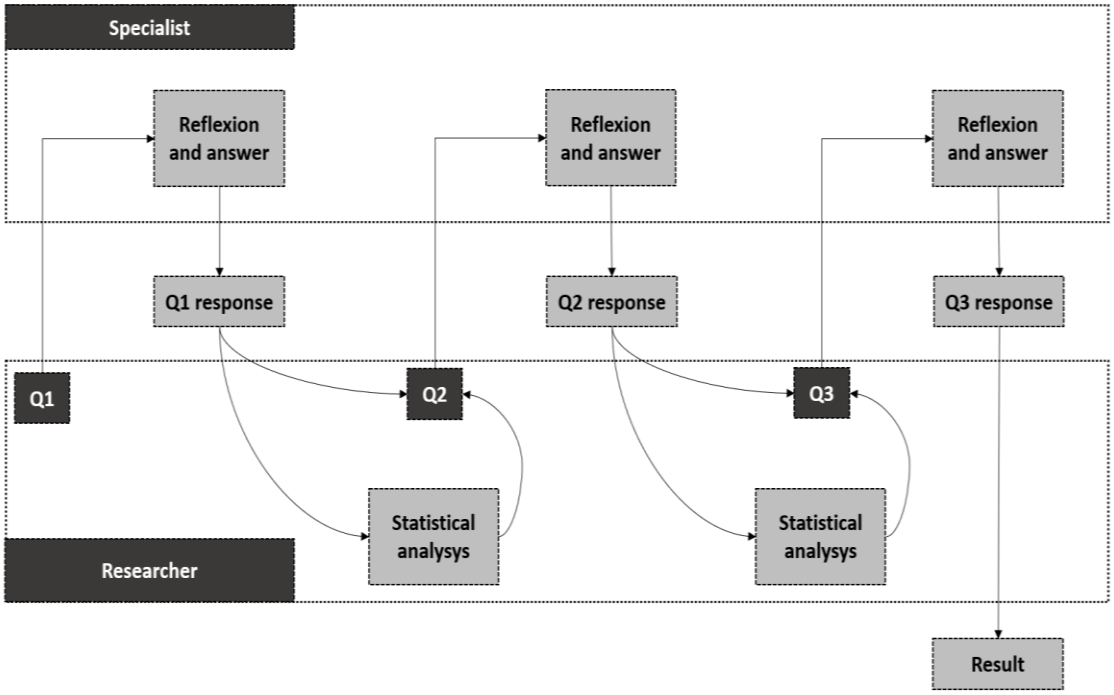

Figure 6 -Implementation of the Delphi method with three rounds – Adapted from (Marques & Freitas, 2018)

3.4.2. Platforms

The platform chosen to conduct the Delphi survey was Google Forms. The reason behind the choice is that it speeds up the response time between surveys and is more organized than emails, for example. Later, for the second and third rounds, the survey was conducted using an Excel file due to the complexity of the structured data to be evaluated. The third round was also conducted in the form of interviews through Zoom meetings.

3.5. Development

3.5.1. Rounds and discussion topics preparation

The first stage of the study, which started with the review of related studies regarding risks and challenges in DS projects, could be classified here as the "literature brainstorming" session. Information was collected regarding the past and current risks in DS projects. The related works revisions were based mostly on identifying the risks and challenges and understanding the company's position towards it. Therefore, this study intends to obtain the same answers, however, in the Portuguese market instead.

The first survey was structured with the risks collected from the literature review, as presented in Table 1 in the previous chapter. In this round, the specialists accepted and denied the risks/challenges in Portuguese DS projects and open questions were presented as well, with the aim of collecting other relevant risks/challenges, not included in the first list. The second survey is structured with all the risks accepted from the first round, plus the other collected risks. In the second round, the participants classify the risks by the business area, their frequency/ probability, impact and level of impact. The third iteration resumes the 25 most identified as frequent risks, ordered by the previous statistic rate. Where, the specialists agree or disagree with the list order and classify them by the order and impact they consider it as well as provide qualitative justification behind their rankings. This iteration additionally includes questions related to the response strategies and tools to manage risks, thus it can be possible to obtain the opinion/classification from every participant regarding the development methodology and its role in the outcome of the projects.

3.5.2. Participants selection

According to Osborne et al. (Osborne, Collins, Ratcliffe, Millar, & Duschl, 2002) commonly, the minimum number for a Delphi panel is 10 with a reduction in error and improved reliability with increased group size. Several studies (Schmidt, Lyytinen, Keil, & Cule, 2001), (Kasi, Keil, Mathiassen, & Pedersen, 2008), (Haes & Grembergen, 2009), (Okoli & Pawlowski, 2004) revealed that studies tend to utilize between 10 and 30 specialist participants. According to Worrell et al. (Worrell, Gangi, & Bush, 2013), besides de usual interval of specialists, a specialist panel as small as four is appropriate if panelists demonstrate a deep understanding of the subject matter or the focal topic requires a unique set of conditions where only a few panelists can contribute towards a solution.

It is important to recall that a fundamental part of the Delphi method is the complete anonymity of all panel members (Worrell, Gangi, & Bush, 2013). The definition of the panelists/specialists was inspired by (Okoli & Pawlowski, 2004) five steps proposed structure:

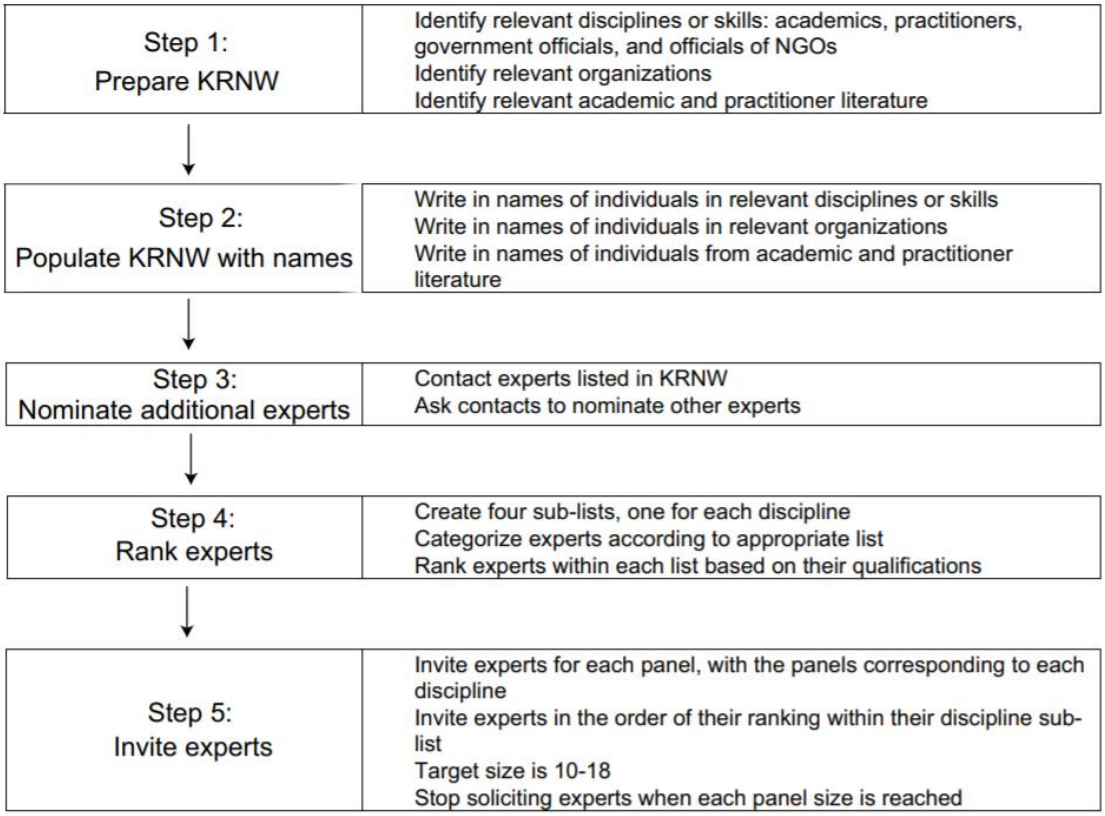

Figure 7 - Procedure for selecting panelists – Source: (Okoli & Pawlowski, 2004)

Figure 7 represents the experts' selection procedure proposed by Okole and Pawlowshi (2004), composed of five steps to create a methodic and organized panelist definition. The steps are:

- Step 1 - Prepare a knowledge resource nomination worksheet (KRNW). The purpose of the KRNW is to help categorize the experts before identifying them. This step defines the skills, expertise and profile intended for the panelist board.
- Step 2 - Populate the KRNW with names. This procedure is used to populate the categories with actual names of potential experts for the Delphi study.
- Step 3 - First-round contacts – In this step the identified experts in contacted and invited to the study and asked to nominate other experts.
- Step 4 - Ranking experts by qualifications - At this step, it compares the qualifications of those on the large list of experts and ranks them in priority for the invitation to the study
- Step 5 - Inviting experts to the study - Based on the rankings, a panel for each of the categories is created. Thus, it is performed the invitation for each panelist and explained the subject of the study and the procedures required.

Based on Okoli and Pawlowshi's proposal, the expert panelist for this study was defined into four stages. "Step 3" proposed by the authors was not applied in the selection. The selection procedure was the following:

Step 1: Preparate KRNW:

This study relied on a group of professionals with valuable knowledge regarding the risks in Data Science projects in Portugal. The first step was to set the intended expertise and profile features for the study. They were:

- Data scientist
- Data Science project manager
- Data Science specialist/ expert
- Big Data specialist/expert
- Machine learning specialist
- Data Science Team manager
- Data analyst specialist/ expert
- Data mining specialist/ expert
- Artificial Intelligence specialist/ expert
- Data engineer

Step 2: Populate list with names:

The process of populating the list involved the selection of the various Data Science professionals in Portugal. After the KRNW was defined, the next iterative procedure was to populate the categories with actual names of potential specialists for the Delphi study. Over the past years, DS project offers have been growing in the Portuguese market and what a better place to find specialists than on the biggest job search/offer and professional sharing platform. The LinkedIn platform was used as a great ally to find the specialists needed. The specialist acknowledgment was provided, by LinkedIn searches.

To find the intended professionals it was applied as search keywords the skills or job profile defined in the first step. For each Data Science field, it was listed as many names as possible into the appropriate categories/areas and sent a connection request on LinkedIn.

Step 3: Rank the experts:

At this step, it was compared the qualifications of those on the list and ranked them in priority for the invitation to the study. First, it was created the three sub-lists: The specialists (in Big Data, Artificial intelligence/Machine learning, Business Analytics, Business Intelligence and Data Mining), the seniors (people with a career in DS rounding 4 years and above) and the consultants (people with some years of experience in DS). The focus of this study is the participation of the specialists and seniors based on their qualifications and years of experience.

Step 4 - Invite experts:

At this stage, the select experts were contacted, provided a brief description of the Delphi study, and explained the goal of the study.

Based on the rankings, was sent 48 invitations for the study. Each one of the professionals on the list was contacted and explained the subject of the study and the procedures required for it, including the commitment required the schedule and privacy terms. From the list, 16 accepted to take part in the study, four declined for personal/professional reasons and 28 did not respond to the invitation, which was reinforced twice.

With 16 confirmed participants, it was created the first panel. Following the recommendations from Delphi literature, the panelist was composed. Within each panel, the goal was to have the most varied backgrounds possible, thus it can be possible to perform analyses under differences in perspectives between respondents.

3.5.3.Timeline

Delphi study could take 45 days to 5 months (Okoli & Pawlowski, 2004). This study, it was planned four to five months. Taking into consideration the scenario where the panelists agree with a commitment within 3 weeks after the first invite and the surveys are responded to within 2/3 weeks.

The detailed contact (invitation) with the experts was first made on February 17th of 2021 through LinkedIn messages.

- The first round was released on March 3rd , 2021 and closed on March 19th , 2021.
- The first round was released on April 27th, 2021 and closed on May 21st, 2021.
- The first round was released on June 16th, 2021 and closed on August 31st, 2021.

Chapter 4 – Demonstration

4.1. Release of the Delphi study and analyses of the results

This session describes the process of the Delphi study conduct. Below, are detailed the conducted rounds' processes and participation, and presents the analyses of the results of this empirical study.

4.1.1. First round

In the first step of this process, the specialists received a message on LinkedIn with the link to the survey.

In the introduction of the survey, there was a text that oriented them as to the purpose of the round and the completion of the survey. The experts were orientated that, based on the list of risks presented and their professional experiences, they should identify the risks they have already encountered in the projects they have participated in and suggest risks not present in the list. The survey link was distributed on March 3rd, 2021 and during the two weeks of execution, it had 13 responses. The responses were anonymous, however, the participants were asked to identify their area of business in Data Science. The profiles were:

- Data Mining specialist: 1
- Business Analyst specialist: 2
- Business Intelligence specialist: 5
- Data scientist: 1
- Data Engineer: 2
- Data Science project manager:1
- Machine Learning specialist :1

The first round was structured with a list of 41 risks of Data Science projects identified during the literature review. In the survey, a set of risks grouped by category was presented. The specialist role was to validate the risks that he/she has already identified in a Data Science project or that is considered relevant and to present new suggestions for risks and their categorization. The main goal of the first round was to undertake a brainstorming round to identify risk. The risk list was structured following (PMI, 2017) Risk Breakdown Structure. The risks categories selected were:

- Functional scope
- Operations (quality) management
- Project management
- Technology
- Estimates
- Technical processes

- Communication
- Organization
- Resourcing
- Surrounding environment
- Legislation
- Regulatory
- External (Environmental, facilities)

Results

In the first round, within the 41 listed risks and 13 participants, some risks got a remarkable identification rate compared to the others.

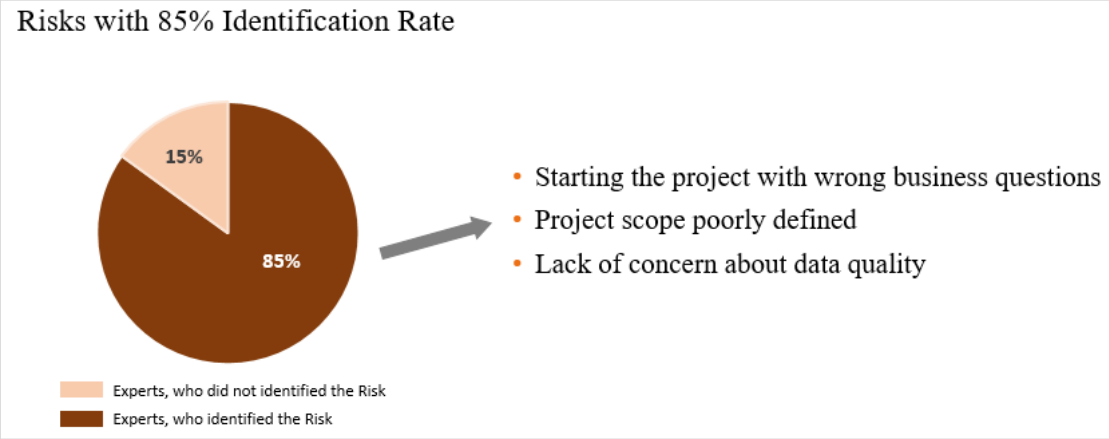

Chart 1- Risk identification rate – first round (1)

As shown in chart 1, around 85% of the experts identified risks such as “starting the project with wrong questions”, “project scope poorly outlined with the client”, and “lack of concern regarding data quality”.

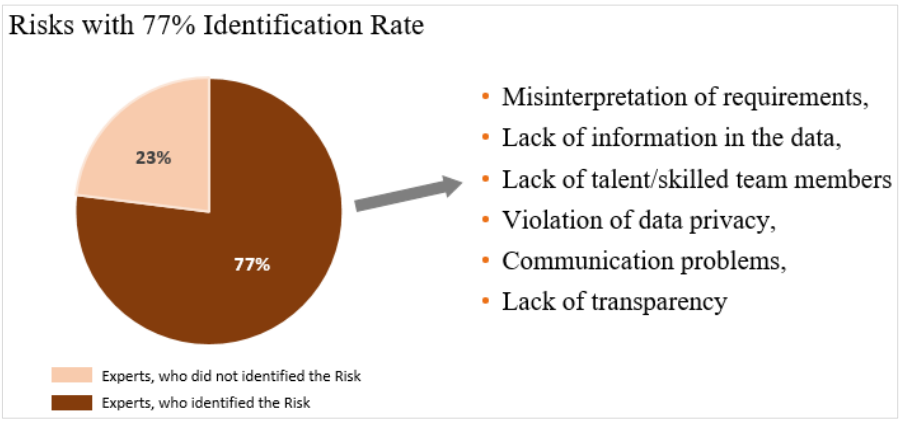

Chart 2- Risks identification rate - first round (2)

Chart 2 shows that 77% identified risks such as “misinterpretation of requirements”, “lack of information in the data”, “poor analysis/validation of model inputs/outputs”, “lack of talent (lack of knowledge and training in data intelligence)”, “violation of data privacy”, “communication problems”, and “lack of transparency”. Such risks are mostly of a functional and project management nature.

Figure 8 illustrates the risks that reached an identification rate between 53% and 61% among the experts.

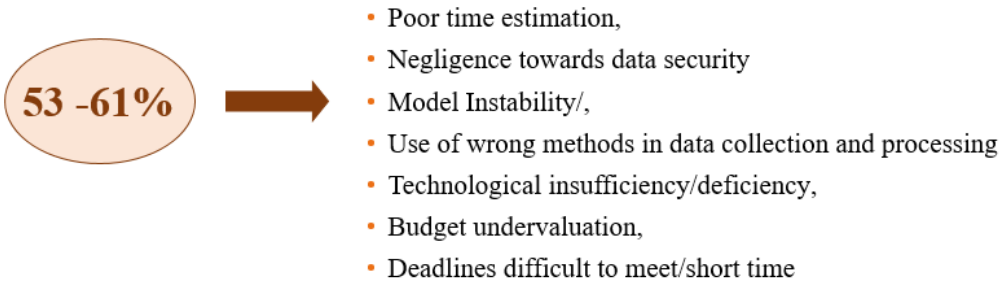

Figure 8 - Risks identification rate - first round (3)

While behaviors related to “negligence towards data security” were often identified (53% identification rate), risks related to “lack of security awareness”, “cyber-attacks” and “theft or loss of information” were rarely pointed out, reaching only around 10% identification rate. Implying that companies and employees are aware of security measures, nonetheless, those measure are not always applied and this security neglect may explain the high identification rate of “Data privacy breaches” and “Loss of confidentiality of data” (around 77%). Risks related to political, economic and social issues were rarely pointed out (around 23% of the experts identified them), as illustrated in Figure 9.

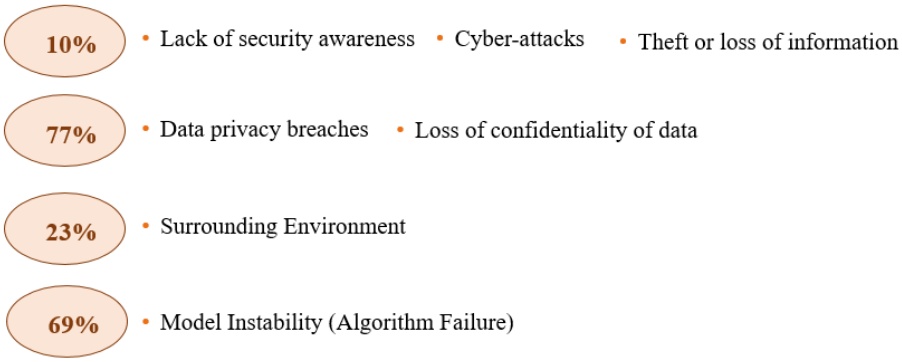

Figure 9- Risks identification rate - first round (4)

The dynamics defined for this first round also allowed the experts to suggest factors and risks considered relevant, that were not presented in the list initially presented, thus creating the brainstorming. Within the structured categories, there were suggestions in almost all of them. They are:

- Functional scope risks (“poor management of customer expectations” and “unclear or lack of documentation (technical and functional)”),
- Project management risks (“lack of clarity regarding which information sources to use”, “poor management of the team skills and high team volatility”),
- Operational management risks (“concentrating on model performance rather than usability” and “lack of knowledge of data segmentation practices”)
- Technical risks (“limited data processor capacity (memory, processor)”)
- Quality risks (“lack of a quality management system on input data (lack of validation on input formatting)” and “poor quality testing by of client”)
- Communication risks (“customer unavailability to meet”, “lack of alignment between the project manager, team and client” and “lack of model explanation for a non-technical language”)
- Technological risks (“lack of clarity regarding the expected growth/Upgrade of the source systems”, “poor software choice/selection”, “systems incompatibility (migration problems)” and “introduction with a new technology”)
- Organizational risks (“lack of information system structure support”).

The following table (Table 2) is structured with the list of risks and the obtained identification for each category. The suggestions are highlighted in grey.

Table 2 - Risks identification obtained in the first round

Category	Risk of Data Science	Identifications
Scope	Begin with the wrong questions	11
	Lack of awareness regarding future changes	6
	Focus on the wrong problem	6
	Confliting goals	
	Executive sponsor’s egos	
Estimates	Poorly defined estimates	7
Technology	Security glitch / Internal or External Attacks (cyber-attacks)	1
	Data security negligence	7
	Technological insufficiency/deficiency	7
	Digital vulnerability	2

	Hardware/Software Faults and Obsolescence	6
	Theft/leakage of information	3
	Lack of clarity regarding the expected growth/Upgrade of the source systems	
	Poor software choice/selection	
	Introduction with a new technology	
Technical Processes	Model instability or performance degradation Model (failed Algorithms)	9
	Wrong Methodologies/ tools	8
	Limited data storage structure (Storage overload)	4
	Interaction issues (Human vs machine)	
	Systems incompatibility (migration problems)	
Organization	Lack of Security awareness of the employees	1
	Negligence in risk planning	5
	Deadlines difficult to meet/short time	8
	Lack of Information System structure support	
	Complex projects	
Resoursing	Budget undervaluation	7
	Inability to reinforce funding	5
	Start project without proper specialists on board	8
	Lack of talent (Lack of knowledge in data intelligence)	10
Communication	Communication issue	10
	Requirement misunderstanding	10
	Lack of transparency and interpretability	10
	Group think (acceptance of one's opinion without any real conviction regarding it)	6
	Customer unavailability to meet	
	Lack of alignment between project manager, team and client	
	Lack of model explanation for a non-technical language	
Operations (Quality)	Very complex models	6
	Use of inappropriate tools/techniques for data collection/processing	8

	Lack of concern regarding data quality	11
	Human Operational Errors	4
	Complex Data	5
	Model Instability/Degradation (Algorithm Failure)	9
	Unexpected/ incorrect or misleading results	
	Inadequate data/ missing of information	
	Limited data processor capacity (memory, processor)	
	Corruption of collected data/ failure in data collection	
	Loss of Authenticity of Data	
	Improper Analytics	
	Flaws in data entry	
	Lack of a quality management system on input data (lack of validation on input formatting)	
	Concentrating on model performance rather than usability	
	Poor quality testing by of client	
	Lack of knowledge of data segmentation pratices	
Legislation /Regulatory	Unintended illegality	4
	Data privacy/ Confidentiality violation	10
External	Natural Disaster	1
	Accidents / physical loss (e.g. fire, hardware theft)	6
Surrounding Environment	Political instability	3
	Economic and Social Disruption	3
	Lack of human employment	1
	Diminished Resilience (loss of human skills due to automation)	3

4.1.2. Second round

The second round involved classifying all the risks identified in the first round by probability, impact, outcome and level of impact. The survey was structured in excel, with an introduction sheet, business area identification sheet and a classification sheet – where the risks were listed.

For this round, it was sent more than 14 invitations (four of them were the reinforcement of first-round invitation), six specialists accepted the invitation, one declined and the seven did not respond. From the first group (16 specialists), just seven of them participated in the second round and from the new set, just four participated. With that, the second round had 11 participants only. The responses were distributed by the following backgrounds:

- Data mining specialist: 1
- Business Analytics specialist: 1
- Business Intelligence specialist: 4
- Data Engineer:1
- Big Data Specialist: 1
- Machine Learning specialist :1
- Data Analytics specialist: 2

The scale was structured in a range from 1 to 5, where the minimum value of 1 is reflected in a very low frequency and impact, whose damage causes minimal conditioning of the work and whose resolutions are easy, with very low delays and overheads, and which do not cause many setbacks to development or deliveries. The probability and impact level options were:

Table 3 - Probability/frequency description (Pooley & Hogarth, 2015)

Scale	Frequency (%)	Classification	Description
1	0 - 10	Very Low	Rare
2	11 - 30	Low	Sporadically
3	31 - 50	Medium	Occasionally
4	51 - 70	High	Often/ Regularly
5	71 - 99	Very High	Consistently/ routinely

Table 4 - Impact description (Pooley & Hogarth, 2015) (Ayyub, 2003)

Scale	Classification	Description (Effect)

1	Very Low	Harmless (Insignificant)
2	Low	Minor (Easily managed)
3	Medium	Moderate (Manageable, yet, demanding)
4	High	Major (Critical, potential for losses and delays)
5	Very High	Catastrophic (Very critical and high loss, potentially fatal)

The impacts type options were:

- Integrity / Credibility
- Availability (service, infrastructure, human...)
- Security (public, infrastructure, business...)
- Quality of Service
- Cost
- Credibility/ Notoriety loss
- Team Productivity and/or Motivation
- Legal process (Lawsuits)

In addition, the oucome options were:

- Opportunity (competitive advance (Chen, Chiang, & Storey, 2012)
- Threat

Results

Based on the objective of the second iteration, it was collected 11 responses. After the analysis, was possible to identify the most frequent risks, across DS fields, different impact levels and the outcomes classified by the specialists. Most of them were classified as a threat, the most identified categories were:

- Project functional scope
- Communication
- Project Management

In addition, the most identified impacts were:

- Project scope
- Cost
- Time
- Service quality

Considering the different backgrounds, it was expected that some risks would occur more frequently in some fields than in others. The structure of the second survey aims to have risk ratings for each area of Data Science that the expert has worked. However, this was not achieved. This round's survey had a more complex structure than the first one. The list contained more risks and the rating extended to each area where the expert has experience. Even though some experts identified more than one area they had experience in, they only rated one of them or repeated the same ratings for all areas.

This phase had some adversities. First, the number of participants was quite small compared to what was expected, limiting the set of opinions. This lack of participation conditioned the construction of a new panelist, to prevent the study from stopping. Another difficulty was the irregularity in the risk classifications. Some possibly random classifications were noted. In this sense, the exploration and analysis of the results of this round were limited.

To improve the rating of risks for the next round, a more robust and shorter version of the risk list was created. As some risks were in a similar context, the list was reconfigured to update/clarify the description of each risk and the scenarios in which it could occur. Based on this assumption, a list of the 25 most frequent risks in each DS field under analysis was created for this iteration, and an overall top 25 was created from that list. It was considered frequent, the risks with the probability level classification "High" or "very high". The list reconfiguration is represented in Table 5.

Table 5 - Top 25 most identified risks as "frequent" in the second round

	Risks	**Description**
1	Project scope poorly defined	It refers to scenarios where the project is started with the wrong questions/focus. Often, people focus more on how interesting a project is and not on how much money/time it saves the company. This can sometimes cause the project questions and answers to be wrongly planned
2	Lack of documentation	It relates to scenarios where there is a scarcity of (technical/functional) documentation and its lack of clarity when it exists. For it is necessary that there is an explanation of how the models work and how the conclusion was reached
3	Lack of transparency and understanding of the project (communication)	Refers to scenarios in which there are communication problems. For example, lack of alignment of ideas between the project manager, team and client, lack of interest or critical sense on project issues, misinterpretation of requirements, or refusal to ask for help in time
4	Little analysis/validation of model inputs and outputs model (data quality negligence)	Refers to scenarios where there is a lack of verification of the quality and format of data (input and output) by the team responsible for the project or client; Lack of concern regarding data quality

5	Negligence in risk planning	Refers to the lack of emphasis given to planning and identifying possible risks to the project
6	Poorly defined time estimates	Refers to underestimated deadlines that are difficult to meet
7	Data system failures	Refers to scenarios related to migration problems, system incompatibility, or data growth not assumed by the target system
8	Complex data - faulty data	Relates to scenarios where there is little margin for manipulation
9	Lack of awareness regarding future changes	Refers to the scenarios in which project planning and development is not carried out aiming at the constant changes and needs of the market
10	Poor skills management in the team	Refers to scenarios in which team members are responsible for tasks outside their area of expertise, or when the inputs of some members are not taken into consideration
11	Budget underestimation	Refers to the scenarios in which the budget planned for the project, is not able to meet the project's requirements
12	Inability to strengthen budget	Refers to the impossibility of increasing the project's financial resources
13	Negligence towards data security	It relates to the lack of training and awareness of employees regarding data security and digital vulnerability security and digital vulnerability
14	Human operational errors in project development	Refers to errors made by the project team during the development of the project (e.g: use of improper methods of data processing and analysis)
15	Conflict of objectives among stakeholders	Refers to the scenario where project stakeholders have diverging opinions and goals
16	Very complex models	It is related to the scenarios where there is difficulty in the usability of the models and a lack of explanation of them. Or when it becomes difficult for whoever continues with the maintenance to maintain the algorithm.
17	Introduction/application of a new technology	It relates to the development scenario of the project, with technology never before used by the team members or by the organization
18	Complex projects	Refers to projects with a high rate of complexity and demand
19	Theft/leakage of information	Refers to unauthorized access to data
20	Poor team management	Refers to scenarios that include, high member volatility and the lack of adequate project specialists (scientists with limited business knowledge)
21	Poor management of customer expectations	This refers to cases in which the client is not informed of the limitations of the work to be developed. By carelessness or lack of experience of the manager/leader

22	Instability/degradation of the model or software (algorithm failure)	Refers to scenarios where models are made available without having solid stability or scenarios where the models stop working well after some time
23	Cyber attacks	It relates to unauthorized access under a computer system, intending to alter, disable, destroying or steal information
24	Lack of knowledge of data segmentation/organization practices	It relates to the good practices of the data organization phases (cleaning, labeling, storage)
25	Client unavailability	It relates to cases where the client delays in responding to the team or has difficulties to meet

4.1.3. Third round

The third round summarised the 25 risks with higher quotations at the level of probability/frequency in the second round. In this round, interviews/meetings were held with the experts for the last classification of the risk scenarios by the same measurement used in the previous round and addressed the other aspects related to project management, team and risks in organizations.

Based on the adversities encountered in the last round, the change in the approach to the interviews was thought of as a way, to deliver and gather information in a more efficient manner.

It was expected, together with the experts, to rank the top 25 risk scenarios of the second round, by their probability/frequency, impact level, the types of impacts (consequences), the strategies/tools and types) of response (mitigate/control, transfer, avoid, accept, escalate) used to manage those risks.

In this round, apart from the ratings, it would also be possible to talk regarding the challenges and risks in Data Science projects.

Taking into consideration the participation in the previous rounds, in this phase, 43 more experts were invited to the study. Where 13 were the reinforcements of previous invitations, 15 were new invitations and 15 were invitations to participants from the previous rounds.

Regarding the answers, 18 experts did not reply, three rejected and five experts accepted the invitations, however, did not participate. In this round, the panelist was composed of 21 participants. The expertise profiles are the following:

- Data mining specialist: 1
- Business Intelligence specialist: 6
- Data Science specialist: 7
- Data Engineer specialist: 3
- Machine Learning specialist:2
- Data Analyst specialist: 2

The experts who participated in this round were not all the same ones who participated in previous rounds and their backgrounds differed. The meetings were previously scheduled with the experts and lasted between 30 mins up to two hours and on average, they lasted 45 mins. A detailed description of the conduction of the interviews can be consulted in Appendix C.

The years of experience ranged from one to 17 years, giving an average of six years, as seen in the figure below.

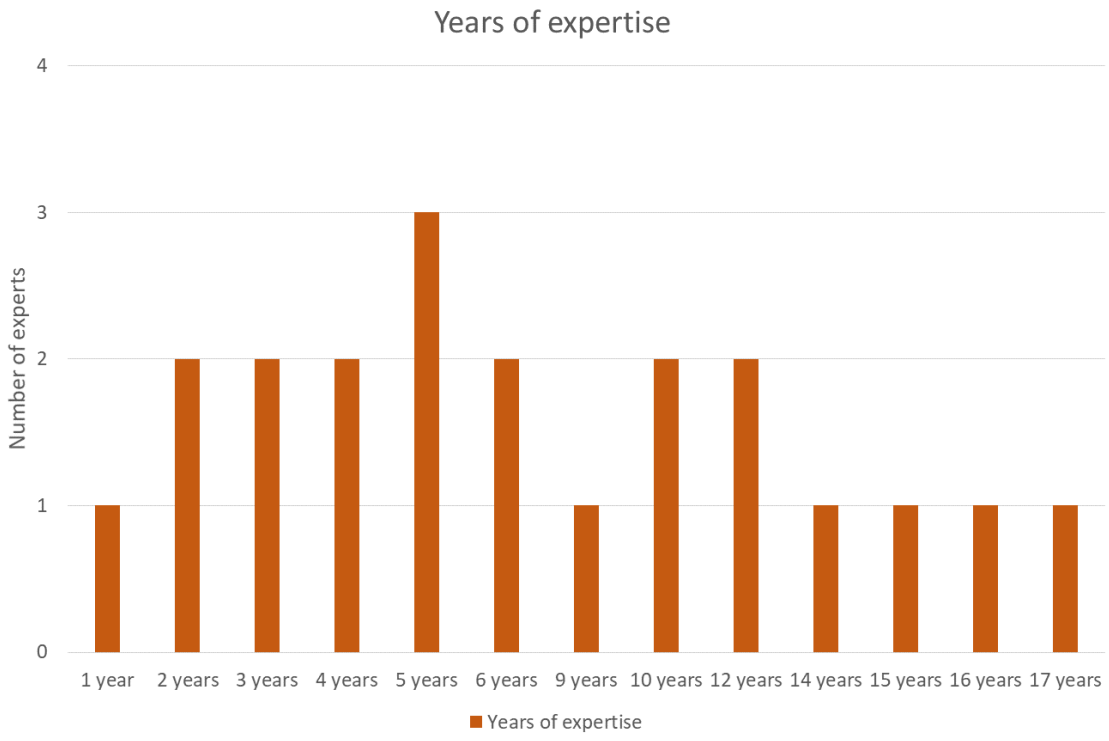

Chart 3 - Distribution of the experts' years of experience

By the end of this study only six experts, participated in all three rounds.

Results

With a slightly different format from the second round, the last round was carried out in the form of meetings and survey responses. In this phase, the specialists classified the 25 risks, in terms of severity (impact) and their frequency and most impacted areas. With this classification, it was possible to make a ranking of the risks considered most frequent and those considered most serious in a Data Science project.

Frequency

By comparing the opinions from the second round versus the third round, it was noticed that the experts changed the ratings regarding half of the listed risks. This factor might have been influenced by the change of approach from file filling to interview or/and by the reformulation of the risks list. In this

third round, there was a better explanation of the risks and that may have influenced more conscious ratings. The results are as follows:

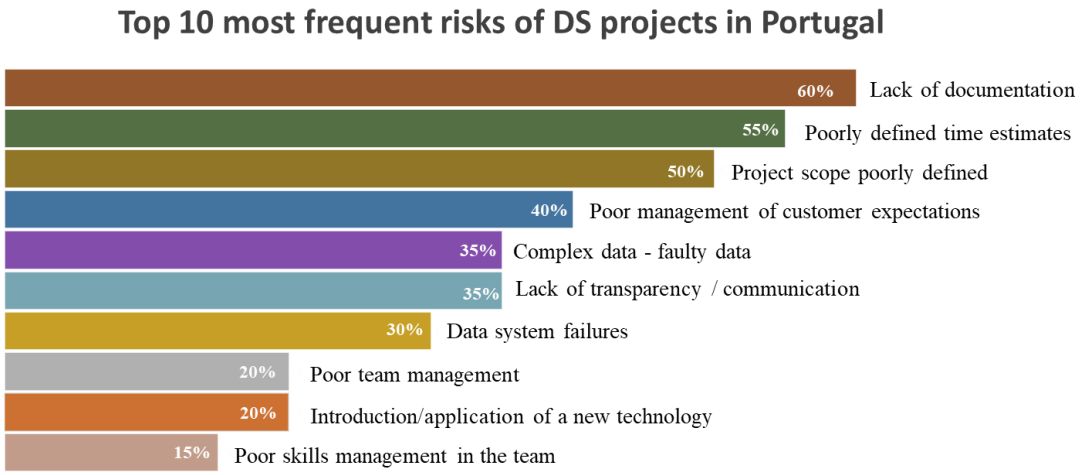

Chart 4 - Top 10 most frequent risks of DS projects in Portugal

The presented chart (4) highlights that around 55% of the specialist considered “poorly defined time estimate” a risk with high to very high frequency. This risk had the highest ranking in terms of frequency and was more distinguished for its sensitivity to the dynamic factors of a project. In other words, factors such as "poorly defined scope", "team inexperience" or "lack of appropriate talents", even "difficulties in accessing data" and "data quality" were identified as the drivers of this risk. The main consequence is the delay in deliveries and the cost associated with the delay. Other risks classified as frequent as the poor time estimate, was the “wrong or bad scope definition” and “lack of documentation”. These risks come from the influence of the client and the project management. As for the “project scope poorly defined” of the project, with a 50% high-frequency rating, this results from the client's lack of sensitivity in realizing their real problem and which business questions they want to answer with that project. With the impact ranging from project delay to project cancellation, this is a very serious risk and a very important factor in triggering other risks, especially the delivery of a project with no business value as well as the impact of "additional costs", "team demotivation" and more.

Regarding the "lack of documents", both functional and technical, there is still a deficit of awareness of the importance of recording decisions and processes related to developments. Ranked second on the frequency list, 95% of the experts reported that this risk occurs on a scale from occasionally to routinely. Other risks, such as "lack of communication and transparency", “poor management of customer expectations” were among the most frequently mentioned (65% - 85%) in a range of medium to very high frequency, in the different areas. The communication issue was pointed out as always being paired with some other risk. Lack or failure of communication influences the misalignment of scope, tasks and responsibilities around the entire project.

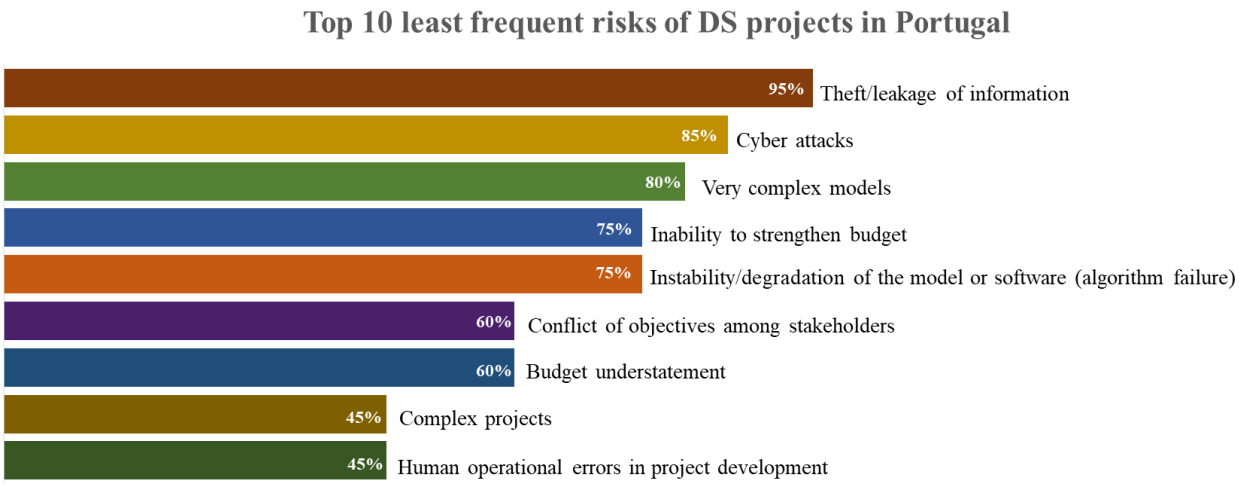

Chart 5- Top 10 least frequent risks of DS projects in Portugal

As shown in Chart 5, 95% of the experts consider "Theft or leakage of information" as a low to very low frequent risk, although the risk related to negligence regarding data security was rated with a frequency of 50% on a scale from medium to very high. Risks related to the budget understatement and inability to reinforce the budget were rated as 60 -75% of low to very low frequency. It was determined that projects with complex models do not occur as frequently (low frequency of 80%), compared to projects of complex scope (considered occasionally to frequently 55% of the time).

The remaining risk frequency rank can be consulted in Table 6. The risks are listed in order of decreasing frequency on a scale from five to one. This listing was prepared with the help of the mode calculation among the ranks.

Table 6 - Risks frequency results

Rank	**Risk**	**Statistical values**	**Frequency Rating Percentage**				
		Mode	**5- Very High**	**4- High**	**3- Medium**	**2- Low**	**1- Very Low**
1	Poorly defined time estimates	5	30%	25%	30%	10%	5%
2	Insufficient / lack of documentation	4	15%	35%	15%	20%	15%
3	Project scope poorly defined	4	25%	35%	35%	5%	0%
4	Poor management of customer expectations	3	20%	20%	45%	10%	5%

5	Complex data - faulty data	3	20%	15%	30%	20%	15%
6	Lack of transparency and understanding of the project (lack of communication)	3	15%	20%	30%	20%	15%
7	Data system failures	3	10%	20%	40%	15%	15%
8	Poor team management	3	10%	10%	45%	30%	5%
9	Introduction/application of a new technology	3	10%	10%	35%	25%	20%
10	Poor skills management in the team	3	5%	10%	40%	30%	15%
11	Lack of knowledge of data segmentation/organization practices	2	20%	15%	10%	30%	25%
12	Negligence in risk planning	2	15%	15%	20%	40%	10%
13	Lack of awareness regarding future changes	2	10%	25%	25%	30%	10%
14	Negligence towards data security	2	10%	15%	25%	25%	25%
15	Human operational errors in project development	2	5%	20%	30%	40%	5%
16	Little analysis/validation of model inputs and outputs model (data quality negligence)	2	5%	15%	35%	35%	10%
17	Budget understatement	2	5%	15%	20%	30%	30%
18	Complex projects	2	0%	30%	25%	35%	10%
19	Instability/degradation of the model or software (algorithm failure)	2	0%	15%	10%	45%	30%
20	Very complex models	2	0%	10%	10%	50%	30%
21	Client unavailability	1	20%	15%	25%	15%	25%
22	Conflict of objectives among stakeholders	1	5%	20%	15%	20%	40%

23	Inability to strengthen budget	1	5%	10%	10%	25%	50%
24	Cyber attacks	1	5%	5%	5%	20%	65%
25	Theft/leakage of information	1	0%	0%	5%	10%	85%

Impact

Regarding the impacts of the risks, the risk considered most serious, with a rating of 90% (critical and very critical), was the "Project scope poorly defined". The experts consider that a poorly defined scope causes serious deterioration to a project, from delays, additional costs and even cancellation. As one of the very frequently identified risks, it is safe to consider that this is a major problem for DS projects and should be avoided as much as possible.

Risks related to data security were also considered to be among the most serious. "Information leakage" (80%), "cyber attacks" and "negligence regarding data security" were rated with high severity by about 65% to 80% of the experts. These risks are characterized by the most serious consequences for companies, jeopardizing the privacy and confidentiality of data, the availability of systems, or even the loss of software.

Another risk of high severity is the “poor management of customer expectation”. This risk was considered severe by 75% of experts. One of the most discussed concerns, links to client awareness of project conditions and data limitations, as this undermines client satisfaction and confidence regarding new projects. The above classification distributions are presented in chart 6.

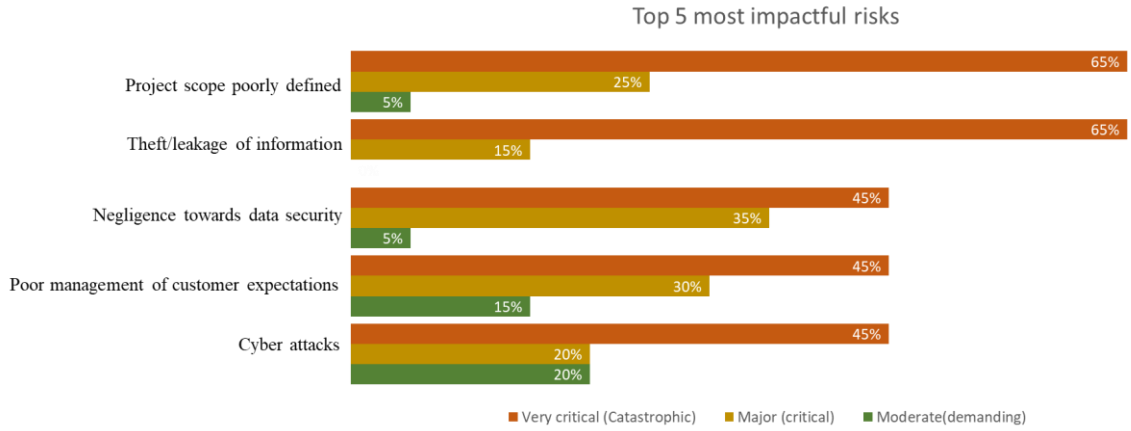

Chart 6 – Top 5 most impactful risks

Concerning impact areas, time and cost are associated with more than half of the risks. Impacts on service/product quality were strongly associated with risks such as "poor team skills management", "poor model analysis", "operational errors", "conflicting objectives among stakeholders" and "poorly defined scope". Legal processes (Lawsuits) were linked to the risks related to data security and customer expectations. The team’s credibility was associated with risks such as “Poorly defined time estimates”,

“negligence towards data security”, “lack of communication and transparency” and “poor management of customer expectations”.

The 10 most frequent risks share the following impact areas:

- Time
- Cost
- Team productivity/motivation
- Service quality

The remaining impact classifications can be consulted in Attachment B.

Concordance value

In this last round besides the mode of the ratings, the coefficient of concordance analysis was also conducted to ascertain the agreement for the frequency of the risks. It was calculated the Kappa coefficient of concordance, proposed by Cohen in (Cohen J. , 1960). Cohen's (Cohen J. , 1960) Kappa is defined in equation 1.

$$K = \frac{(\bar{P} - \bar{P}_e)}{(1 - \bar{P}_e)} \quad (1)$$

Where, $\bar{P}$ denotes the observed proportion of agreements among all classifications and $\bar{P}_e$, denotes the expected value of $\bar{P}$ under "random" or "null" agreement. (Gross, 1986), or the proportion of units for which agreement is expected by chance (Cohen J. , 1960). The Kappa coefficient measures the degree of agreement between two raters using multiple categories in classifying the same group of subjects. However, the coefficient was extended by Fleiss (Fleiss, 1971), to the case of multiple raters. If all specialists agree on the risks classifications, K is one (Cohen J. , 1960). The more varied the rating, the closer K gets to zero.

To develop a measure of agreement of this study, the value of Kappa was obtained through the following equation (Fleiss, 1971):

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{k} Pi \quad (2)$$

$$P_i = \frac{1}{n(n-1)} \left(\sum_{j=i}^{k} {n_{ij}}^2 - n\right) \quad (3)$$

$$\bar{P}_e = \sum_{j=1}^{k} {p_j}^2 \quad (4)$$

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \quad (5)$$

Where, N= 25 represents the total number of risks, n=20 the number of ratings per risks and k=5 the number of valued scale into classifications were made.

The obtained value of $\bar{P}$ was 0.25, it means that if a risk is selected at random and classified by a randomly selected specialist and the same risk was also classified by a second randomly selected specialist, the second specialist would agree with the first over 25% of the time. Using Cohen's "K" equation that calculats the normalized measure of overall agreement or the proportion of agreement after the chance agreement is removed from consideration, the final value of Kappa was 0.04 that represents 4% of agreement.

As the number of items classified and raters was significant, the concordance value was also computed for each Data Science field present in this round. The results obtained were:

Table 7- Agreement's value for each Data Science field

Field	Specialist	$\bar{P}$	$\bar{P}_e$	K
BI	5	0.23	0.218	0.017
DS	7	0.26	0.21	0.063
DE	3	0.24	0.251	-0.015
DA	2	0.12	0.123	-0.004
ML	2	0.32	0.299	0.03
DM	1	-	-	-

By the end of this study, it is acceptable to interpret a reasonable level of agreement regarding the risks' frequency (25%).

4.2. Analyses of the interviews

In this section, a more detailed analysis is made of the results obtained in the last round of the study. This section also presents the results of the analysis of the interviews conducted. This analysis contains the experts' opinions regarding project challenges, development methodologies and success and failure rates of the projects that have already participated.

4.2.1. Others findings

Regarding the risks planning, 70% of the experts said that there was inexisting or little risk planning in the projects. The big focus of management in planning is only right at the beginning, and it should not be that way. The other 30% said that the risks are planned throughout the project development, as their assumptions. Therefore, if they occur the team knows what procedure should be taken.

The experts agreed that there are risks that are conducive to creating other risks and these deserve attention as soon as they are detected. However, most of the time is to try to solve them as soon as they appear and if possible, in an internal context. If it is something that the relationship itself with the client provided (lack of communication, client delay, lack of access), this issue is shared with the client to be solved. Sometimes, the time and cost impact to the project is evaluated, and to see if it is worth addressing soon. If it is not eliminated or mitigated, it is controlled through supervision and documentation, thus the team is always aware of the probabilities of aggravation and its impacts. Regarding the development methodologies used, Agile was the most mentioned, Waterfall was also mentioned and it was furthermore mentioned that in some cases the project was carried out without any specific methodology.

Regarding the success and failure rates, the statistic was relative to what the expert believes is a success. Project success at the DE level does not always constitute success at the BI level. Because models do not always respond to the customer's real problems. There is the success of a project at the technical and delivery level and success at the level of decision making and usability. A successful project is not exclusively the project delivery within the planned parameters scope, time and cost, it is the project delivery of a project with quality and capable of answering the business questions. Success goes beyond what is delivered, it is necessary that the product responds to the company's needs and generates value.

Considering this, the “success” definition discussed was: “a project delivery, within the time, scope and cost planned”. The ratings ranged from zero to 70% within the parameters defined above. Moreover, experts emphasized that these were small or closed projects. As projects with open scope, it is always readjusted throughout the project. As already mentioned in the previous points, a good part of the identified risks had a different influence on the project deadline. Despite delays, experts consider that the projects were successfully delivered, as they responded to the defined scope.

Regarding failed projects, 60% of experts reported that they had projects canceled. The reasons ranged from, technology or scope no longer made sense, cost, time (projet taking too long to complete), complexity, customer lack of budget (poorly planned), lack of necessary resources, customer lack of interest and others. Regarding the biggest challenges with data itself, two major challenges stated were:

- Data governance and availability - Knowing who owns that data, who is responsible for correcting it in case it goes wrong and having the commitment of these people, that is, having

all the involved parties committed to working for the data and the project is the most difficult. The larger the company, the greater is the bureaucracy and the greater is the segregation of data. The accesses require a lot of bureaucratic and time-consuming treatment. In this regard, the project is conditioned to the delay right from the beginning.

- Quality of data - Deliver maximum value without adding too much “junk” as data brings a lot of information from different contexts.

Even though the agreement level was quite low, the study provided useful opinions regarding the exiting risks of DS projects, the consequences associated with those risks, the methodologies used and their role, and a brief understanding of how companies and teams deal with risks.

The remaining content of the interviews, containing the expert’s personal opinions regarding each risk is described in Attachment C.

Chapter 5 – Conclusions and recommendations

5.1. Main conclusions (Key findings of the study)

The present work had as a proposal the study of the risks of Data Science projects, focusing on the identification and analysis of these risks in projects in Portugal, to understand the most relevant ones and the consequences that arise from them and understand some of the reasons behind the failures of projects.

To support this research, a study was conducted using the Delphi technique, where it was possible in the first phase to identify and collect the risks present in the DS projects in Portugal. Later, it was possible, based on the list raised, to classify those risks in the different areas of DS application, to collect information about the frequencies, impacts and other components of these risks.

Considering the characteristics of the study participants and the objectives under study, the sampling process chosen was a very small sample of what is real in the market, not being possible to generalize the results of this study. The target population of the Delphi study consisted of 30 people throughout the 3 phases of the study. The Delphi technique was very useful for this qualitative research. It helped explore and identify the nature and fundamental elements of Data Science projects risks, creating the basis for this statement.

Although the result of the consensus of the frequency of risks was only 25%, the objective of this study was achieved, because it was possible to acknowledge that:

- The main dangers and major factor of negative events in Data Science projects is the lack and failure of communication between the parties involved in the project.
- The biggest risks are wrong estimates, wrong scope/business questions, poor project transparency and lack of documentation.
- Regarding the data, the biggest risks are the lack of access and translation of data and its poor quality.
- The identified risks were mostly considered as threats to the project. The most identified impacts were cost, time, quality of service and team motivation.
- Agile was the development methodology most identified in data projects in Portugal. It was also found that some projects do not follow a concrete methodology during their development. Which causes disorganization and an unsound segmentation of the solution requirements, causing many constraints to appear during project development.
- No risk management tools were identified during this investigation and the risk control strategies are unsound. Mostly, it includes postponing the elimination or control of the risk until it is no longer possible.

- There is a need for risk planning during project planning. Therefore, the team can avoid some of them.
- This research allowed the knowledge that data projects have a high failure rate in Portugal. The exact percentage is still unknown, but reasons such as costs, delays, changes in business approaches and lack of project value for business are among the causes. These causes are among the main impacts of the risks identified in DS projects in Portugal.
- Organizations are required to invest in training their employees, both in better understanding Data Science and its components and improving the team's ability to understand the business of the data they are handling.
- Risks have a circulatory and generative cycle around each other. Good management can reduce the impacts of some risks, enabling the non-emergence of other possible hazards and risks in a project.

The experts consulted during the study consider that this topic is little talked about and should be discussed more often, at the beginning and during the projects. They considered that this study was a starting point for the awareness of risks in Data Science projects and a way to learn about the different types of risks and how to avoid them.

This study was also presented as a paper in the international project management conference "CENTERIS | HCist | ProjMAN 2021", where was presented and discussed the results of the Delphi study. The audience considered it a very meaningful topic, as the conference also included other presentations regarding DS projects and this topic complemented the discussion.

It is considered that this research has managed to bring to the surface some of the problems inherent in Data Science projects in Portugal. It is hoped that the result of this study will be able to shed light on the factors that cause poor project outcomes and improve the results of current and future work and projects, whether small or large solutions.

It is also considered that the objectives of this research were achieved and that in the future it will be possible to have greater participation of specialists in this IT area, therefore, a broader perspective of the scenarios can be analyzed.

For future work, it is recommended to have a solid base of classification items to achieve the desired consensus and prevent study structure disorder.

Finally, it is hoped that this research was the first of many others to come, concerning the risks and scenarios of Data Science projects in Portugal.

5.2. Recommendations of best practices for risk avoidance and control

After the conclusion drawn from this study, it is clear that data managers, teams, and corporations should better analyze the outcome of their projects, identify what is wrong with Data Science initiatives,

and learn how to avoid the same mistakes and ensure success and best practices. Based on the risks and risk factors presented in this study, a list of best practices and initiatives has been developed to help companies improve the outcome of projects.

One of the first steps in a project is business understanding and then understanding the data, and its sources (Butka & Ivančáková, 2020). With this knowledge, there is a better understanding of the customer's problem and only then, the manager and the development team are able to plan the best solution for the customer. Presenting most of the business to the engineering team greatly reduces the chances of the scope not being consistent with the customer's problem and improves all decisions within the team throughout the development process. Ex: which data best responds to the customer's forecast or query, which tool best fits the process, which departments should be involved in the project, among other decisions. The validation of results together with the business (The Standish Group, Factors of Success (FOS2015), 2015), to have the conviction that the project is answering or will answer the client's business questions is crucial. According to (Bakker, Boonstra, & Wortmann, 2011), the most important roles in risk management and project performance are the communication and collaboration between stakeholders.

Another good practice is to prioritize the demands of the project, especially when the project schedule starts to lag. It is to deliver value at each sprint, thus, the customer can see what has been done. Making the business is closer to the technical possibilities of the solution. Other good practices also have been reported in other research (The Standish Group, CHAOS Manifesto, 2013), (Malaska & Seidman, 2018), that satisfy the answer for the challenges and risks presented in this study. They are:

- Executive management support
- Systems optimization
- Skilled resources (provide continuous training)
- Project management expertise
- Agile process
- Clear business objectives
- Emotional maturity
- Execution
- Tools and infrastructure
- Data Authorization, Encryption and Auditing

Traditional Agile processes were intended to reduce uncertainty and answer questions related to the actual needs of customers and how software can be delivered reliably and continuously. Carter and Hurst (2019) stated that in Data Science, the questions are related to the meaning of the data and what solutions the team can provide based on that data. The authors stated that in Data Science, Agile principles could

be applied to solve problems and reduce the uncertainties associated with data. For this, there is a need for a modern Agile approach that includes mixed talent.

Finally, a recommendation from Malaska and Seidamn in (2018) which says, that balancing technology, team and requirements risks and setting up the right methodology, creates an environment for success. Using development principles and strategies for managing and mitigating risk, setting realistic expectations provides a guide to developing a project of success.

By the end of this investigation, it is considered that, to successfully, execute data projects that deliver value to the business through data meaning, a set of qualified skills, processes and good practices must be adopted. The team is supposed to identify the risks associated with each project under evaluation or development and create a base plan to avoid and resolve them to achieve project success.

5.3. Limitations of the study

One of the main limitations of this study was being able to maintain the participation of the experts in all three phases of the study. Since each phase had a stipulated time interval, it was not possible to wait for the response or participation for too long. Therefore, the panelist was updated whenever necessary so as not to compromise the process. For this reason, the collection of information was in itself very challenging and the inconsistency of the participations did not allow the conduct of the Delphi method in its standard structure, which requires a fixed panelist.

The listing and rankings were done based on a business approach. A major contribution to this study would be the participation of experts with experiences outside the corporate bubble.

Another factor that is a limitation was the lack of membership of experts with a management profile in the DS area. It is considered that these are individuals with a more general and framed view of many risks that engineers and data analysts are not aware of. The lack of representativeness regarding the specialty of the experts represented a limitation as well.

5.4. Future research proposals

Considering the fact that, it was not possible to gather a consensus on the frequency of risks, a continuation of the Delphi process, with two more rounds, would be very useful for the investigation of this topic, as it will bring greater certainty regarding the occurrence of risks in DS projects in Portugal.

It would be very useful for the area of Data Science to conduct this study for each field of practice.

Bibliographical references

Accenture. (2014, April). *Accenture Big Success with Big Data Survey*. Accenture.

Ackoff, R. L. (1989). From Data to Wisdom. *Journal of Applies Systems Analysis, 16*, 3-9.

Alturki, A., Gable, G. G., & Bandara, W. (2013). THE DESIGN SCIENCE RESEARCH ROADMAP: IN PROGRESS EVALUATION. *Association for Information Systems (AIS)*.

Arabnia, H. R., Daimi, K., Stahlbock, R., Soviany, C., Heilig, L., & Brussau, K. (2020). *Principles of Data Science*. USA; Belgium; Germany;: Springer.

Asay, M. (2017, July 12). *3 ways to massively fail with machine learning (and one key to success).* Retrieved from Tech Republic: https://www.techrepublic.com/article/3-ways-to-massively-fail-with-machine-learning-and-one-key-to-success/.

Ayyub, B. M. (2003). *Risk Analysis in Engineering and Economics.* Boca Raton / London / New York / Washington, D.C.: Chapman & Hall/CRC.

Bakker, K. d., Boonstra, A., & Wortmann, H. (2011). Risk managements' communicative effects influencing IT project success. *International Journal of Project Management*, 14.

Balachandran, B. M., & Prasad, S. (2017). Challenges and Benefits of Deploying Big Data Analytics in the Cloud for Business Intelligence. *Procedia Computer Science, 112*, 1112–1122. doi:10.1016/j.procs.2017.08.138

Bara, D., & Knežević, N. (2013). The impact Impact of Right-Time Business Intelligence on Organizational Behavior. *Interdisciplinary Management Research Ix*, 27-42.

Bellinger, G., Castro, D., & Mills, A. (2004). *Data, Information, Knowledge and Wisdom.* Retrieved from Systems Thinking: https://www.systems-thinking.org/dikw/dikw.htm

Bertocchi, D. (2016). *Dos dados aos formatos: a construção de narrativas no jornalismo digital.* Brazil: Editora Appris Ltd.

Bratianu, C. (2015). *Organizational Knowledge Dynamics: Managing Knowledge Creation, Acquisition, Sharing, and Transformation.* USA: IGI Global.

Brooks, K. W. (1979). Delphi techniques: Expanding applications. *North Central Association Quarterly*.

Buckley, C. (1995). Delphi: a methodology for preferences more than predictions. *Library Management*, 16-19.

Butka, P., & Ivančáková, J. (2020). Methodologies for Knowledge Discovery Processes in Context of AstroGeoInformatics. *Astrogeoinformatics*, 1-20.

Capgemini. (2014). *Cracking the Data Conundrum: How Successful Companies Make Big Data Operational.*

Carter, E., & Hurst, M. (2019). *Agile Machine Learning: Effective Machine Learning Inspired by the Agile.* USA: Apress.

Chapman, C., & Ward, S. (2003). *Project Risk Management: Processes, Techniques and Insights.* Chichester,: John Wiley & Sons, Ltd.

Cheatham, B., Javanmardian, K., & Samandari, H. (2019). Confronting the risks of artificial intelligence. *McKinsey Quarterly*, 1-9.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, 1165-1188.

Chowdhury, A. A., & Arefeen, S. (2011). Software Risk Management: Importance and Pratice.

CMBI. (2020). *BUSINESS INTELLIGENCE PROJECT MANAGEMENT.* Retrieved from CMBI: http://www.cmbi.com.au/6000_BIProjectManagement.html

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20(1)*, 37–46. doi:10.1177/001316446002000104

Cohen, L., & Holliday, M. (1996). *Practical Statistics for Students: An Introductory Text.* London: Paul Chapman Publishing Ltd.

Conway, D. (2010, September 30). *THE DATA SCIENCE VENN DIAGRAM.* Retrieved from drewconway.com: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Cooper, W. W., Gallegos, A., & Granof, M. H. (1995). A Delphi study of goals and evaluation criteria of state and privately owned Latin American airlines. *Socio-Economic Planning Sciences, 29*, 273–285. doi:10.1016/0038-0121(95)00017-8

Dalkey, N., & Helmer, O. (1963). AN EXPERIMENTAL APPLICATION OF THE DELPHI METHOD TO THE USE OF EXPERTS. *Management Science*, 458-467.

Dammann, O. (2018). Data, Information, Evidence, and Knowledge:A Proposal for Health Informatics and Data Science. *Online Journal of Public Health Informatics, 10(3), 10(3):e224*. doi:10.5210/ojphi.v10i3.9631

Davenport, T. H. (2014). *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities.* Boston: Harvard Business Review Press.

DIDRAGA, O. (2013). The Role and the Effects of Risk Management in IT Projects Success. *Informatica Economică, 17*, 13. doi:10.12948/issn14531305/17.1.2013.08

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin,, 76(5)*, 378–382. doi:10.1037/h0031619

Frawley, W. (1992). Knowledge Discovery in Databases: An Overview. Relational Data Mining. *AI Magazine* , 28-47.

Freedman, D. H. (2017, June 27). *A Reality Check for IBM's AI Ambitions*. Retrieved from MIT Technology Review: https://www.technologyreview.com/2017/06/27/4462/a-reality-check-for-ibms-ai-ambitions/

Frei, L. (2019, March 27). *Why The Data Science Venn Diagram Is Misleading*. Retrieved from Towards data science: https://towardsdatascience.com/why-the-data-science-venn-diagram-is-misleading-16751f852063

Frey, C. B., & Osborne, d. M. (2013). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting & Social Change*, 284-280.

Gaikema, M., Donkersloot, M., Johnson, J., & Mulder, H. (2019). Increase the success of Governmental IT-projects. *Research Gate*, 97-105.

Gartner. (2014, October 21). Egham: Gartner Press Releases. Retrieved from Gartner Identifies the top 10 strategic technologies trends for 2015.

Gartner. (2017, 7 june). *Critical capabilities for data science and machine learning platforms.* Retrieved from Gartner Research: https://www.gartner.com/en/documents/3740317-critical-capabilities-for-data-science-platforms

Gärtner, B., Hiebl, & R.W., M. (2018). Issues with Big Data. *The Routledge Companion to Accounting Information Systems*.

Graaf, R. d. (2019). *Managing Your Data Science Projects: Learn Salesmanship, Presentation, and Maintenance of Completed Models 1st ed. Edition.* Australia: Apress.

Grisham, T. (2009). The Delphi technique: a method for testing complex and multifaceted topics. *International Journal of Managing Projects in Business,* 112-130.

Gross, S. T. (1986). The Kappa Coefficient of Agreement for Multiple Observers When the Number of Subjects is small. *Biometrics, 42(4),* 883-893. doi:10.2307/2530702

Haes, S. D., & Grembergen, W. V. (2009). An exploratory study into IT governance implementations and its impact on business/IT alignment. *Information Systems Management*, 123-137.

Hart, C. (1998). *Doing a Literature Review: Releasing the Social Science Research Imagination.* London: Sage Publications.

Hasan, N. A., Rahman, A. A., & Lahad, N. A. (2015). ISSUES AND CHALLENGES IN BUSINESS INTELLIGENCE CASE STUDIES. *Jurnal Teknologi*, 171-178.

Hashem, I. A., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, 98-115.

Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *Internacional Journal of Production Economics*, 72-80.

Hevner, A. R., Ram, S., March, S., & Park, J. (2004). Design research in information systems research. *MIS Quarterly, 28*, 75–105.

Hevner, A. R., Salvatore, T. M., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 75-105.

Hobbs, B., & Aubry, M. (2010). *Project Management Office (PMO): A Quest for Understanding.* Pennsylvania: Project Management Institute Inc.

Inmon, W. H., & Nesavich, A. (2017). *Tapping into Unstructured Data: Integrating Unstructured Data and Textual Analytics Into Business Intellince.* USA: Pearson Education, Inc.

Isik, O., Jones, M. C., & Sidorova, A. (2012). Business intelligence success: The roles of BI capabilities and decision environments. *Information & Management*, 13-23.

J.H, R., & R, P. (2015). Survey on Software Project Risks and Big Data Analytics. *Procedia Computer Science*, 295 – 300.

Jadhav, D. K. (2013). Big Data: The New Challenges in Data Mining. *International Journal of Innovative Research in Computer Science & Technology*, 2347-5552.

Jadhav, D. K. (2013). Big Data: The New Challenges in Data Mining. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST)*, 2347–5552.

Jifa, G. (2013). Jifa, Gu (2013). Data, Information, Knowledge, Wisdom and Meta-Synthesis of Wisdom-Comment on Wisdom Global and Wisdom Cities . *Procedia Computer Science, 17*, 713–719. doi:10.1016/j.procs.2013.05.092

Joshi, R. C., & Gupta, B. B. (2020). *Security, Privacy, and Forensics Issues in Big Data.* USA: IGI Global.

Jurney, R. (2014). *Agile Data Science: Building Data Analytics Applications with Hadoop.* USA: O'Reilly Media, Inc.

Kasi, V., Keil, M., Mathiassen, L., & Pedersen, K. (2008). The post mortem paradox: a Delphi study of IT specialist perceptions. *European Journal of Information Systems*, 67-78.

Kelleher, J. D., & Tierney, B. (2018). *Data science.* Cambridge, Massachusetts: MIT Press.

King, S. (2014). *Barriers to the Implementation of Big Data.* Wiesbaden: Springer.

Kshetri, N., Fredriksson, T., & Torres, D. C. (2017). *Big Data and Cloud Computing for Development: Lessons from Key Industries and Economies in the Global South.* New York: Routledge.

Kwak, Y., & Stoddard, J. (2004). Project risk management: lessons learned from software development environment. *Technovation, 24(11)*, 0–920. doi:10.1016/s0166-4972(03)00033-6

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33(1)*, 159-174. doi:10.2307/2529310

Legendre, P. (2005). Species Associations: The Kendall Coefficient. *Journal of Agricultural, Biological, and Environmental Statistics*, 226–245.

Lewis, M. (2015, October 16). *Risk Management in Analytics Projects*. Retrieved from Sciente Consulting: https://www.scienteconsulting.com/blogs/823/risk-management-analytics-projects/

Linstone, H. A., & Turoff, M. (2002). *The Delphi Method: Techniques and Applications.* New Jersey.

Liulka, V. (2017, November 14). *RISKS IN DATA SCIENCE PROJECT*. Retrieved from Linkedin: https://www.linkedin.com/pulse/risks-data-science-project-vladimir-liulka/

Love, P. E., Irani, Z., Standing, C., Lin, C., & Burn, J. M. (2005). The enigma of evaluation: benefits, costs and risks of IT in Australian small–medium-sized enterprises. *Information & Management*, 947–964.

Luckey, T., & Phillips, J. (2006). *Software Project Management For Dummies.* Canada: Wiley Publishing, Inc.

Malaska, T., & Seidman, J. (2018). *Foundations for Architecting Data Solutions.* Sebastopol,: O'Reilly Media, Inc.

Manheim, K., & Kaplan, L. (2019). Artificial Intelligence: Risks to Privacy and Democracy. *21 Yale J.L. & Tech*, 106.

Marques, J. B., & Freitas, D. d. (2018). Método DELPHI: caracterização e potencialidades na pesquisa em Educação. *Pro-Posições*. Retrieved from https://doi.org/10.1590/1980-6248-2015-0140

Mikos, M., Tiwari, B., Yin, Y., & Sassa, K. (2017). *Advancing Culture of Living with Landslides: Volume 2 Advances in Landslide Science.* Slovenia, USA, Indonesia, China, japan: Springer. doi:10.1007/978-3-319-53498-5

Moslem, S., Ghorbanzadeh, O., Blaschke, T., & Duleba, S. (2019). Analysing Stakeholder Consensus for a Sustainable Transport Development Decision by the Fuzzy AHP and Interval AHP. *sustainability*, 3271–.

Moss, L. T., & Atre, S. (2003). *Business Intelligence Roadmap: The Complete Project Lifecycle for Decision- Support application.* Boston: Pearson Education Inc.

Nadimpalli, M. (2017). Artificial Intelligence Risks and Benefits. *International Journal of Innovative Research in Science,Engineering and Technology*, 4.

Napoleão, B. M. (2019, Junho 26). *Matriz de Riscos (Matriz de Probabilidade e Impacto)*. Retrieved from Ferramentas da qualidade: https://ferramentasdaqualidade.org/matriz-de-riscos-matriz-de-probabilidade-e-impacto/

Nofal, M. I., & Yusof, Z. M. (2013). Integration of Business Intelligence and Enterprise Resource Planning within Organizations. *Procedia Technology*, 658-665.

Norris, C., Perry, J., & Simon, P. (2000). *Project Risk Analysis and Management Guide.* Buckinghamshire: APM Publishing Ltd.

OAIS. (2012). *The Reference Model for an Open Archival Information System (OAIS).* Washington: Magenta Book. Retrieved from OAIS Reference Model (ISO 14721): https://public.ccsds.org/pubs/650x0m2.pdf

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: an example, design considerations and applications. *Information & Management*, 15-29. doi:10.1016/j.im.2003.11.002

O'Neil, C., & Schutt, R. (2014). *Doing Data Science: Straight Talk from the Frontline.* USA: O'Reilly Media, Inc.

Osborne, J., Collins, S., Ratcliffe, M., Millar, R., & Duschl, R. (2002). What "Ideas-about-Science" should be taught in school science? A Delphi study of the expert community. *Journal of Research in science teaching*, 692-720.

Osoba, O. A., & Welser, W. (2017). The Risks of Artificial Intelligence to Security and the Future of Work. *RAND Corporation*, 23. doi:10.7249/pe237

Ozdemir, S. (2016). *Principles of data science.* Birmingham - Mumbai: Packt Publishing.

Pardal, L., & Correia, E. (1995). Métodos e Técnicas de Investigação Social. Porto: Areal Editores.

Peffers, K., Tuunanen, T., Gengler, C. E., Rossi, M., Hui, W., Virtanen, V., & Bragge, J. (2006). THE DESIGN SCIENCE RESEARCH PROCESS: A MODEL FOR PRODUCING AND PRESENTING INFORMATION SYSTEMS RESEARCH. *1st International Conference, DESRIST 2006 Proceedings*, 83-106.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2008). A Design Science Research Methodology. *Journal of Management Information Systems*, 53.

Pereira, M. N. (2012). *Aplicação da Metodologia de Registo de Riscos a um Empreendimento em Construção.* Lisboa.

Plugar. (2018, March 23). *4 reasons for failure in Data Science projects and how to avoid them*. Retrieved from Plugar Data & Intelligence: https://www.plugar.com.br/4-falhas-data-science-como-evitar/

PMI. (2017). *A Guide to the Project Management Body of Knowledge (PMBOK® Guide) Sixth Edition.* Pensilvânia: Project Management Institute, Inc.

Pooley, T., & Hogarth, R. (2015). *Risk Bandits: Rescuing Risk Management from Tokenism.* Bloomington: Balboa Press.

Press, G. (2020, January 13). *AI Stats News: Only 14.6% Of Firms Have Deployed AI Capabilities In Production*. Retrieved from Forbes: https://www.forbes.com/sites/gilpress/2020/01/13/ai-stats-news-only-146-of-firms-have-deployed-ai-capabilities-in-production/?sh=4685374c2650

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data Analytics Thinking.* Sebastopol: O'Reilly Media, Inc.

Quivy, R., & Campenhoudt, L. V. (1998). Manual de Investigação em Ciências Sociais.

Raguseo, E. (2018). Big data technologies: An empirical investigation on their adoption, benefits and risks for companies. *International Journal of Information Management*, 187-195.

Raisinghani, D. M. (2004). *Business Intelligence in the Digital Economy: Opportunities, Limitations and Risks.* Dallas: Idea Group Inc.

Rekha, J. H., & Parvathi, R. (2015). Survey on software project risks and big data analytics. *Procedia Computer Science*, 295–300. doi:10.1016/j.procs.2015.04.045

Robinson, E., & Nolis, J. (2020). *Build a Career in Data Science.* New York: Manning Publictions Co.

Schmidt, R., Lyytinen, K., Keil, M., & Cule, P. (2001). Identifying Software Project Risks An International Delphi Study. *Journal of Management Information System*, 5-36.

Schwalbe, K. (2016). *Information Technology Project Management, Eighth Edition* . Boston: Cengage Learning.

Sfaxi, L., & Aissa, M. M. (2020). DECIDE: An Agile event-and-data driven design methodology for decisional Big Data projects. *Data & Knowledge Engineering*.

Shukla, S., Kureethara, J. V., Han, D. S., Unal, A., & Mishra, D. K. (2021). *Data Science and Security: Proceedings of IDSCS 2021.* India, Korea, USA: Springer Nature Singapore Pte Ltd.

Stommel, M., & Dontje, K. J. (2014). *Statistics for Advanced Practice Nurses and Health Professionals.* New York: Springer Publishing Company.

Tagle, B. (2019). *A Design Science Research Methodology: Guide for graduate students.* Augusta.

Tan, K. H., Zhan, Y., Ji, G., Ye, F., & Chang, C. (2015). Harvesting Big Data to enhance supply chain innovation capabilities. Analytic infrastructure based on deduction graph. *International Journal of Production Economics*, 223-233.

Taulli, T. (2019). *Introdução à Inteligência Artificial: Uma abordagem não técnica.* São Paulo: Apress Novatec.

Taurion, C. (2013). *Big Data.* Rio de Janeiro: Brasport Livros e Multimídia Ltda.

Taylor, B. (2017, October 16). *Why Most AI Projects Fail.* Retrieved from DataRobot: https://www.datarobot.com/blog/why-most-ai-projects-fail./

The Standish Group. (2013). *CHAOS Manifesto.*

The Standish Group. (2015). *Factors of Success (FOS2015).* Boston: The Standish Group.

Tizhoosh, H. R., & Pantanowitz, L. (2018). Artificial intelligence and digital pathology: Challenges and opportunities. *Journal of pathology Informatics*, 1-6.

Tunowski, R. (2015). Business Intelligence in Organization. Benefits, Risks and Developments. *Przedsiębiorczość i Zarządzanie*, 133-144.

Ulrich, P., Frank, V., & Timmermann, A. (2020). The dark side of data science - An empirical study of cyber risks in German SMEs. *Procedia Computer Science, 176*, 2615–2624. doi:10.1016/j.procs.2020.09.307

Veeramachaneni, K. (2016, December 7). *Why You're Not Getting Value from Your Data Science.* Retrieved from Harvard Business Review: https://hbr.org/2016/12/why-youre-not-getting-value-from-your-data-science

Vieira, R., Ferreira, F., Barateiro, J., & Borbinha, J. .. (2014). Data Management with Risk Management in Engineering and Science Projects. *New Review of Information Networking*, 49-66.

Vieira, R., Ferreira, F., Barateiro, J., & Borbinha, J. .. (2014). Data Management with Risk Management in Engineering and Science Projects. *New Review of Information Networking*, 49-66.

Walker, J. (2017, November 23). *Big data strategies disappoint with 85 percent failure rate.* Retrieved from Digital Journal: http://www.digitaljournal.com/tech-and-science/technology/big-data-strategies-disappoint-with-85-percent-failure-rate/article/508325#ixzz5JvGk3AV0

Wallace, L., Keil, M., & Rai, A. (2004). How Software Project Risk Affects Project Performance: An Investigation of the Dimensions of Risk and an Exploratory Model. *Decision Sciences*.

Waterman, K. K., & Bruening, P. J. (2014). Big Data analytics: Risks and responsibilities. *International Data Privacy Law, 4(2),* 89–95. doi:10.1093/idpl/ipu002

Worrell, J. L., Gangi, P. M., & Bush, A. A. (2013). Exploring the use of the Delphi method in accounting information systems research. *International Journal of Accounting Information Systems*, 193-208.

Yang, Q. (2010). Three challenges in data mining. *Frontiers of Computer Science in China, 4(3),* 324–333. doi:10.1007/s11704-010-0102-7

Yarnold, P. R. (2014). UniODA vs. Kendall's Coefficient of Concordance (W): Multiple Rankings of Multiple Movies. *Optimal Data Analysis*, 121-123 .

Attachments and Appendixes

Attachment A

RBS LEVEL 0	RBS LEVEL 1	RBS LEVEL 2
0. ALL SOURCES OF PROJECT RISK	1. TECHNICAL RISK	1.1 Scope definition
		1.2 Requirements definition
		1.3 Estimates, assumptions, and constraints
		1.4 Technical processes
		1.5 Technology
		1.6 Technical interfaces
		Etc.
	2. MANAGEMENT RISK	2.1 Project management
		2.2 Program/portfolio management
		2.3 Operations management
		2.4 Organization
		2.5 Resourcing
		2.6 Communication
		Etc.
	3. COMMERCIAL RISK	3.1 Contractual terms and conditions
		3.2 Internal procurement
		3.3 Suppliers and vendors
		3.4 Subcontracts
		3.5 Client/customer stability
		3.6 Partnerships and joint ventures
		Etc.
	4. EXTERNAL RISK	4.1 Legislation
		4.2 Exchange rates
		4.3 Site/facilities
		4.4 Environmental/weather
		4.5 Competition
		4.6 Regulatory
		Etc.

Figure 1- Risk Breakdown Structure (RBS) - Source: (PMI, 2017)

Attachment B

Table 1, represents the ranked risks listed in order of decreasing impact obtained in the third round of the Delphi study. This listing was prepared with the help of the mode calculation among the ranks. The classification of impacts was not always possible, as the risk was not always identified as existing, therefore, the impact cannot be estimated. For this reason, the total rating percentage in some cases will not reach the value of 100%.

Table 1-Risks impact results

Rank	Risk	Statistical values	Impact Rating Percentage				
		Mode	5- Very High	4- High	3- Medium	2- Low	1- Very Low
1	Project scope poorly defined	5	65%	25%	5%	0%	0%
2	Theft/leakage of information	5	65%	15%	0%	0%	0%
3	Negligence towards data security	5	45%	35%	5%	5%	5%
4	Poor management of customer expectations	5	45%	30%	15%	5%	0%
5	Cyber attacks	5	45%	20%	20%	5%	0%
6	Instability/degradation of the model or software (algorithm failure)	5	40%	25%	30%	5%	0%
7	Client unavailability	5	40%	25%	20%	10%	0%
8	Poor team management	5	35%	30%	30%	5%	0%
9	Complex data - faulty data	5	35%	25%	25%	15%	0%
10	Poor skills management in the team	5	30%	25%	25%	15%	0%
11	Inability to strengthen budget	5	30%	20%	20%	0%	10%

12	Lack of transparency and understanding of the project (lack of communication)	4	40%	45%	10%	5%	0%
13	Data system failures	4	25%	45%	20%	5%	5%
14	Little analysis/validation of model inputs and outputs model (data quality negligence)	4	25%	40%	25%	10%	0%
15	Lack of knowledge of data segmentation/organization practices	4	25%	35%	20%	5%	10%
16	Negligence in risk planning	4	15%	50%	15%	20%	0%
17	Very complex models	4	10%	50%	30%	5%	0%
18	Budget understatement	4	10%	30%	30%	15%	10%
19	Conflict of objectives among stakeholders	3	15%	35%	35%	15%	0%
20	Insufficient / lack of documentation	3	20%	15%	55%	10%	0%
21	Human operational errors in project development	3	15%	25%	30%	25%	5%
22	Poorly defined time estimates	3	15%	20%	50%	15%	0%
23	Complex projects	3	10%	25%	40%	15%	5%
24	Lack of awareness regarding future changes	3	0%	25%	45%	30%	0%
25	Introduction/application of a new technology	2	5%	25%	20%	35%	5%

Attachment C

This attachment session presents the analysis of the risks from the expert point of view collected in the third round.

I. Project scope poorly defined

“Scope is the biggest aspect of a project, yet, it is rarely well defined. The ill-defined scope starts when the management is unable to extract the client's real needs, which depends on the maturity of the company's project team. If the project goal is badly defined, everything that follows runs the risk of going wrong.

Sometimes, clients want everything. These situations usually occur when they do not know what they want, and not “everything” in a data project is possible. One of the first steps in a project is to know the problem that needs to be solved. It is always necessary that the project manager elaborates with the client the most concise and exact requirements possible. All other decisions stem from this. Developing a product aiming at the wrong target is the biggest failure of a project, even if it works technically and technologically well.

Depending on the methodology used, this risk can be controlled or trigger other negative events. Agile methodologies, allow the scope to be reviewed and changed on time if it does not make sense. It is essential to do a good data discovery. It is always necessary to evaluate if Data Science is the" right solution" for the customer's problem, because "poorly defined business questions” is one of the most serious project risks.”

II. Lack of documentation

“Either the documentation is outdated, does not reflect the current situation or it simply does not exist. It depends on the maturity of the team. Those who have more experience, document more because they know the importance of it, however, it is still very common to join a project without documentation. It is an even worse situation when the people who retain the knowledge of data and technology are no longer part of the business. Consequently, projects are longer lasting and more expensive. Due to lack of documentation, sometimes it takes additional months to complete a project and the understanding of the model needs to be done through reverse engineering. Many companies like to exclude Data Engineering from functional tasks, however, data engineers need to document the model because what was developed needs to be documented.

In terms of codes, a good alternative in the comments is the purpose of the instructions. Thus, allowing whoever is reviewing the code to know the meaning of what has been implemented. It is a manageable risk when you have a team with good knowledge”.

III. Communication failure and lack of transparency in projects

“It can be influenced by the lack of knowledge of the customer, and sometimes the customer does not pass the information correctly, influencing a different understanding to the team that will develop it. It also depends on the maturity of the client and team (who has more experience communicates more). Sometimes, the structure and formatting of the input data are not always well communicated or interpreted and end up taking a lot of time for the developer to structure it in an interpretable format by the development system.

There is also the scenario of lack of understanding of the project, influenced by poor communication within the development team, the lack of critical sense regarding the tasks and their implications, which undermines the success and quality of the product/service. Without the understanding of what is being done or needs to be done, the project will likely not go well. There has to be a certain rapprochement between the development team and the customer's team, the right customer's team. It is also necessary to have good management between the collaboration between the development team and the client's team. Avoid situations where the development team does not have business support and data translation by the client. This lack of communication and transparency is one of the great drivers of the ill-defined scope. In very frequent situations the customer rarely asks for something that will solve their problems and the team with little understanding of the business will not know how to ask the customer the right questions.”

IV. Lack of data quality - little input and output data analysis

“The great challenge of data projects is to ensure the quality of data entering and leaving the developed models. Bad data create bad Data Science, therefore, it is vitally important to take the time to ensure the data is of high quality. This is true of any analytics procedure and is certainly the case for Data Science. Junior teams tend to skip a few steps and have little awareness of possible failure scenarios. Thus, it is essential to spend enough time and necessary to make a thorough evaluation of the data surrounding the project, according to its nature (structured or not), source and format. The volume of data to be processed is often misestimated so, it is necessary to validate the scenarios before the models are published. Ideally having someone on the client-side do the data preparations, therefore, the development team does not have this concern. Many DS projects fail because the data is not handled correctly. There are many predictive models out there that if it inputs garbage, it outputs garbage as well. The so-called "garbage in, garbage out" concept. The team has to pay a lot of attention to the quality and formatting of the data that is fed into the models as well as the output quality. It is essential to purify and exclusively select the data that can confirm hypotheses, predict trends and detect weak points in the business.”

V. Wrong time estimate

“Data projects involve many uncertainties, from the technology to use, to the business areas involved. Sometimes tasks are estimated without knowing the data conditions and the real people

involved. It represents one of the biggest impacts on data projects. Nevertheless, the impact can be minimized with a scope with a review to evaluate what can be delivered first. Depending on the level of error in the estimates, it can even lead to project cancellation, because after a long waiting time, the client often loses interest in the project."

VI. Data system failures

"One of the assumptions in any project is that the system can fail, at some point of its lifecycle. However, it is necessary to be prepared for this, especially in a productive environment. Such failures can lead to service unavailability, which can cause great inconvenience to the company.

Often the customer does not have an integrated system, the Date Warehouse is not in good condition, and the processes can fail even in production and it happens occasionally. Lack of communication can and often leads to this risk."

VII. Complex or faulty data

"By nature, many projects already have unstructured data, although, it is not a problem when there are specific people, with appropriate skills to handle these situations. However, this risk always ends up consuming more time and creating more costs to the project. When the issue is defects and lack of data labeling, the problem can become serious or even an obstacle to the project's progress. Data processing is a delicate process, as in some cases the processing of some data itself involves legal issues, as is the case of names. There is always a limit to how far data can be cleaned without losing important information. One of the bad data prevention techniques is the implementation of data labeling rules at the front end, in order that the stored data can have some coherence."

VIII. Lack of awareness regarding future changes

"Even though it is not always possible to see a long-term scenario, there is still only a current concern, when a solution is built, without thinking regarding how the service can serve in the future. Some solutions are designed to solve only static and current problems. As for the functional level, depending on the type of project (methodology, deadline, etc.), there is the possibility to adapt and control the changes that occur, however, at the technological level, there is little space for adaptation over time. For those cases, the solution is to migrate the system to a more current technology capable of providing better answers, when it is possíble to do so. When working with large companies, there is a lot of adaptation, however, when the company does not have this financial capacity, this can represent a big problem, especially when support is discontinued. Companies always want better solutions, the technological market is constantly presenting something innovative and the solution is to bet on technologies that the organization's "pocket" allows. It is still very common for projects to be deployed and unable to follow the business changes and needs."

IX. Poor team management

"It is less critical when the team already has some maturity to guide itself and is critical regarding the tasks, time and organization of the project. The manager's performance greatly influences the productivity and motivation of his team, both positively and negatively."

X. Poor skill management

"Underutilizing talents, using unskilled people to provide input on issues that are not skilled does not always bring good results for the project. When the project requires it, it is always necessary to have the right specialists to carry out their tasks. It is very important to a project that team members bring different influences to it, and that there is always a good exchange of perspective and brainstorming on the direction of the project. The management of qualifications within a team is still a challenge, as often the skills that are hired are not the skills that are demonstrated throughout the project. Influencing volatility, lack of adequate specialists and the quality of the product delivered."

XI. Data Security negligence

"There is still some negligence and lack of security sensitivity towards projects, although, the problems are not frequent. The data protection laws are very important and any breach puts the company in a very sensitive position. It is still possible to see cases of staff having access to information that they should not have. There have been cases where if the business was more attractive or generated more curiosity, the company could have been in a bad situation. Some companies already have people who are experts in data regulation matters."

XII. Human errors

"It happens, however, nowadays there are already many processes to track and control errors. They are workable and many mistakes are natural, therefore, there is always correction range. It becomes very critical when the error is too broad to the project."

XIII. Conflict of objectives among stakeholders

"Usually this risk is derived from the projects of large companies, with many bureaucratic processes and involved departments. Departments with divergent initiatives. It is a serious matter, as it can delay and sometimes terminate a project. It depends on the simplest to the most serious cases, from methodologies, technology to be used, how the product/service will be delivered. Often in a data project, the client asks for something, however, that thing will not always solve their problem, nor will that always bring them results. Moreover, this often causes conflicts because the project manager has a different view of the client's problem and its solution. In some cases, the conflict arises when the outcome of a project only benefits a part of the business."

XIV. Very complex models

"Complex models alone can represent opportunity as well as threat. It is closely linked to the assessment of customer needs. The complex model itself is not a problem (if it solves the customer's problem, it fulfills its purpose), the problem lies in its implementation and maintenance. It is not something that anyone could handle, due to its robustness and complex neural structure, it is necessary to have maturity and documentation. It is a delicate situation, when the complexity of the model interferes with the real needs of the customer, as has already been witnessed a few times. For this reason, it constitutes a time and cost risk factor to the project. Models do not always need to be complex to solve the customer's problem. Nevertheless, it is always essential that the developing team knows what the model is doing and the customer knows what the outcomes represent. A good way to prevent these risks from negatively influencing the company is in addition to development, the team also has a maintenance contract and well-documented models."

XV. Introduction/application of a new technology

"It depends on technology and business. It poses a risk when the technology is too constrained and expensive and may not be adaptable to changing approaches to data. Companies often bet on new technologies, however, it depends on technology and depends on the level of structure of the company. Generally, those who venture into new technologies are companies that have the conditions/capacities for it, and usually it brings great opportunities."

XVI. Poor management of customer expectations

"Very often, the impact is greater when combined with other noises such as miscommunication and wrong scope."

XVII. Complex projects

"Complex projects are a challenge, especially when it requires a lot of people and the right people. If the team is junior, there are more chances of getting into trouble.

Complex projects are opportune for the company as it enhances its visibility, nonetheless, it is necessary to have a structure prepared for it. If something goes wrong, the project effect ends up being the opposite. Therefore, a factor that can generate many risks and events to the project. A complex project is much more succinct to risks than simpler projects, as it involves more people, more structure, and/or a more complex business area."

XVIII. Information theft/leakage

"The impact of this risk depends on the company and its business. Some companies already have some data governance policies, to prevent information leakage. "

XIX. Model instability/degradation

"Failure of forecasting models is serious the larger the business is. If the risk is associated with the degradation of the technology, and/or team's lack of knowledge, and/or lack of documentation, which brings some inconveniences such as cost, the best thing to do is to redo the project with the best-known and most used technology.

The models have a lifetime, however, there are teams prepared for these scenarios and fixing them. Model failures are one of the major risks of Data Engineering and Artificial Intelligence, and the larger the business or service, the greater the maintenance and monitoring of its performance. It is essential to train models well."

XX. Cyber attacks

"It is a serious risk and depending on what is lost the impact can be catastrophic."

XXI. Lack of knowledge of data organigazation practices

"It is important to have a storage-oriented design and have an organized data infrastructure.

Good data query and storage practices are very important, especially for data-intensive projects. Data organization affects model performance at firsthand. Nowadays, there are already several high-volume data processing technologies, however, the way they are handled after this process is essential. Companies must invest in teams with good knowledge of data warehouses."

XXII. Customer unavailability

"Customer unavailability often hinders the progress of work and service delivery. The impact depends on how long the customer takes to get back to the team. Customer support is essential on a project."

Expert two

"… 'Successful project' depends on the type of methodology used, and it depends on what was delivered. Huge projects, the "official" definition of success is somewhat unrealistic. Success turns out to be finishing/delivering the project with the least possible setback. Because these are projects that their structure makes difficult the development..."

Expert 10

"…The risks almost always have the same consequence, which is the cost…. Many times, they end up being a waste of investment. Yes, there is a disconnection between the management and the technical team. This is because the technicians are often only included at the start of the project, and not during project planning. Some risks are very technical, and often on paper, it is easy to say that certain things

are possible, although, it is not always. Many times, technically it takes a long time, there is no information, and there is no data required for that…."

"….The risk and the project are the whole teams, as well as the success and failure. The team is responsible for handling all problems, risks and barriers…"

Attachment D

Table 2- DSR process elements for IS, presented over the years (Peffers, et al., 2006)

Objectives for a design science research process model	Archer (1984)	(Takeda et al. 1990)	Eekels and Roozenburg (1991)	Nunamaker et al (1991)	Walls et al (1992)	(Rossi et al. 2003)	(Hevner et al. 2004)
1. Problem identification and motivation	Programming Data collection	Problem enumeration	Analysis	Construct a conceptual framework	Meta-requirements Kernel theories	Identify a need	Important and relevant problems
2. Objectives of a solution			Requirements				Implicit in "relevance"
3. Design and development	Analysis Synthesis Development	Suggestion Development	Synthesis, Tentative design proposals	Develop a system architecture Analyze and design the system. Build the system	Design method Meta design	Build	Iterative search process Artifact
4. Demonstration			Simulation, Conditional prediction	Experiment, observe, and evaluate the system			
5. Evaluation		Confirmatory evaluation	Evaluation, Decision, Definite design		Testable design process/product hypotheses	Evaluate	Evaluate
6. Communication	Communication						Communication

Appendix A

This appendix contains the Delphi study invitation and the instructions presented to the experts in the first round.

Invitation

"Dear Mr/Mrs Dr/ Prof X,

I would like to invite you as a specialist in the field of Data Science in Portugal, to participate in a Delphi study process. The study aims to recognize specialists' opinions on risks and challenges in Data Science projects in Portugal, and the tools and techniques used to avoid and respond to those risks and challenges.

This study is a part of my master's degree research, which is being built to contribute to the creation of solid and scientific information regarding Data Science in Portugal based on Portuguese Data Science professional inputs.

I realize that you are extremely busy in your respective area, however, because of the important input you can bring to the project, I hope that you will agree to participate in it. In practical terms, I would require no more than 10 minutes of your time, on three separate occasions (if possible).

I send in attachments all the details regarding the study and your participation if you agree with it.

I would appreciate it if you gave me your answer, by accepting or declining your participation.

Thank you in advance for your time.

Yours sincerely,

Cristina Varela"

Study Explanation (attached to the message)

"The Delphi technique is a structured prediction method based on the assumption that combining the anonymous opinion of a group of experts will result in a more accurate prediction of a truth.

This Delphi Process will consist of three rounds. Experts will submit their responses via Google forms by filling out specially prepared forms.

This Delphi process will involve answers to research questions round one, ranking the answers generated by experts in round two, reaching consensus in round 3.

Answers may be changed between rounds, based on aggregated information from the previous round. The Delphi process ends when (the predefined level of) consensus is reached, or when the predefined number of rounds has been completed.

The first round is scheduled to start on March 1st, 2021, and consists of risk identification based on a pre-identified list. There is also the possibility of adding recurring risks to Data Science projects not identified in the form.

The second round is scheduled to start on March 22th , 2021, where all risks identified in round 1 will be ranked by their frequency, impact and Data Science area. In this round, the response tools and strategies will also be ranked in relation to the risks presented.

The third round will start on April 19, 2021, and will summarize the rankings in the order obtained in round two, for another concordance ranking.

The result of the study will be tallied in May/June 2021 and all participants will have access to the result.

Observation:

The list of participants and their opinions will be treated anonymously, avoiding any compromise of the participants' data in this study."

First round instructions

"Dear X,

I hope you are well.

As a follow-up to our previous conversation, I am providing the link to the survey for the study I am developing. The link is: https:"…"

I reinforce once again that the answers are anonymous and handled directly by me.

Thank you for your collaboration in the study and if you have any questions, please feel free to contact me.

Best regards,

Cristina Varela"

Instructions on the form**:**

"This survey is part of a Delphi study that aims to identify recurring risks in Data Science projects in Portugal.

This first iteration was structured based on the risks of Data Science projects identified in studies published to date.

In the course of this survey, a set of risks grouped by category will be presented. The goal is to validate the risks that you have already identified in a Data Science project or that you consider relevant and to present new suggestions for risks and their categorization."

Appendix B

This appendix contains the Delphi study invitation and the instructions presented to the experts in the second round.

Second round invitation:

"Dear X,

I hope everything is going well with you.

Following the first phase of the Delphi study regarding the risks of Data Science projects in Portugal, I am making available the second survey of the study.

In this second phase, the survey is available in an excel file. Therefore, I suggest that you open it on your computer for a better user and viewing experience.

Thank you for your participation and if you have any questions, please feel free to contact me."

Instructions on the file:

"This second phase has the objective of classifying the risks identified in the first phase by the probability and impact they represent to the organization and analyzing the opportunities and threats arising from them.

To start filling out the survey, click on the "Start" button, choose your area(s) of specialization in Data Science and proceed to the risk rating page by clicking on the "Next" button.

Thank you for your willingness to participate in this study.

Best regards,

Cristina Varela"

Appendix C

This appendix contains the Delphi study invitation and the instructions presented to the experts in the third round.

Thrid round invitations:

"Dear X,

I hope that everything is going well with you.

Following the Delphi study that you have cordially agreed to participate in, I would like to invite you to a 20-30 minute meeting to follow up on the third and final round of this study.

The purpose of this meeting is to discuss the results of the second round and to obtain a conclusion from your opinion, plus some other information regarding the risks and their identified factors.

In this sense I would like to ask, when can you make this time available for us to talk?

I look forward to hearing from you.

Best regards,

Cristina Varela"

****

"Dear X,

I hope everything is going well with you.

I would like to invite you as an expert in the Data Science area in Portugal, to participate in a Delphi study process.

The purpose of the study is to identify the experts' opinions regarding risks and challenges in Data Science projects in Portugal, and the tools and techniques used to avoid and respond to them.

This study is part of the research for my master's thesis in Integrated Decision Support Systems, which is being developed in order to contribute to the creation of a list of risks and strategies for action regarding Data Science projects in Portugal.

I know you are extremely busy in your respective area, however, because of the important contribution you can bring to the project, I would like to invite you for a brief conversation, in order to

obtain your feedback, regarding the risks under study. This would be regarding 20 to 30 minutes of your time. If you accept, when can you make this time available for us to talk?

Thank you in advance for your time.
I look forward to hearing from you.

Best regards,
Cristina Varela"

Interview's guideline:

"Dear expert,

First of all, thank you for accepting the invitation and to have reserved this little time to talk to me. It is a pleasure to talk to you personally and thank you for your previous participation.

…As mentioned before, this study is part of my master's research, and it has been very important to highlight the concept of the risks in Data Science projects in Portugal and your testimony has been very important in this process…..

…I have summarized the 25 most identified risks as frequent in DS projects, and I would like you to classify them once again by frequency, the level of impact and which business areas you consider to be impacted by the risk.…..

…Besides the rankings, I would like to have your opinion regarding some additional information.

- According to the evaluation of the results of the second round of the study, the most frequent risks are those related to project management, project scope and communication. Do you agree? (Yes or No), why?
- In the projects that you have participated in, is there risk planning?
- What is normally done when a risk is identified (the response)?
- Is a risk mitigation plan drawn up after a risk is planned? Or only when it occurs, if it does occur?
- Does the project follow a certain development methodology?
- In the projects, you've been part of, an estimate, what percentage of those have had complete and delivered successfully (in time, cost and scope)?
- And how many were not delivered (canceled, postponed...)?

… Once again I would like to thank you for your valuable participation and availability. This study is yet to be completed, and when the result is out I would like to share it will you with the perspective of bringing some input into the perception of risks and hazards in their data projects."