CARMEN PELAYO FERNÁNDEZ, 2022

LOCATION
RECOMMENDATION SYSTEM
FOR TECH BUSINESSES IN EUROPE

Executive Summary

Given the high importance of making the right business location choices, this project aims to provide a decision-making support solution for entities working in the ICT (Information and Communication Technology) industry to locate their operations in Europe.

With this purpose, the socio-economic factors influencing the location decisions of technological firms were carefully studied. To do this, the existing literature in this regard was reviewed, so as to identify the set of concrete dimensions that need to be considered in the process of locating a technological firm. Subsequently, the condition of these detected dimensions was examined in the individual European regions, in order to rank them and, ultimately, deliver a location recommendation.

For the recommendations to fit different business profiles, a tool based on a matchmaking algorithm was built, enabling customization. This way, users with different business characteristics and demands can access personalized and precise location recommendations. This tool was developed using data-mining techniques and was deployed in a web application, so as to enable the input of user data and deliver results in a visual and user-friendly manner.

Apart from supporting the location of technological businesses, the created solution can be configured to fit other purposes, like the search for specialized employment, the visualization of socio-economic data, or the discovery of available capital funding.

To examine the efficiency of the built system, the results obtained by the algorithm were compared against macroeconomic data of the European regions. It was observed that the generated location recommendations were significantly correlated to the regions with high technological activity, confirming the well-functioning of the recommendation system created.

Contents

Tables

Figures

Acknowledgments

To my father Enrique, my mother Gloria, and my sister Maria, who have been my main source of love and support my entire life, but especially in these months of hard work.

To my thesis director Jose Maria, for trusting me to carry out this interesting project and for guiding, supporting, and motivating me during its development.

To my friends and classmates, with whom I have shared unforgettable moments and who have also been a fundamental support during my time in college.

To the University Carlos III of Madrid and its teaching and administrative staff, for their work and dedication.

1. Introduction

"Knowledge is power" (*Scientia Potestas Est*) —as Francis Bacon rightly said in the 16th century [1], the more information we have on a subject, the better decisions we make. This is why information is something human beings have fought for since the beginning of history –whether it was to win a war, emerge victorious from a trial, or expose a corrupt individual. Today, the information war has moved into the economic arena, and it is precisely companies and other institutions that are fighting for their power. Good use of the right data is synonymous with good business management, as it allows better business decisions to be made.

One of the business decisions modern firms struggle most with taking is the location of their operations. This is a fundamental election that can dictate the future success or failure of an organization and will imply one of the greatest monetary investments for any type of company. Moreover, in a fully-globalized world, the location options are endless. That is why the creation of a system to drive location decisions, based on the appropriate analysis and utilization of multi-source socio-economic data, can provide a great advantage in the business arena.

On the other hand, a relatively young industry in the market is the so-called ICT (Information and Communications Technologies) field. Firms in this industry have special business features and location requirements, since they do not directly rely on natural resources or the physical presence of employees to operate. Being these technological companies recently established in the market, the research on their location decisions is yet limited.

With the aim of assessing the two concerns stated, this work intends to explore the decision-making process of firms in the ICT industry to locate their operations, and advance on this line of knowledge by proposing a system to optimize this practice.

1.1. Motivation

The location decisions of technological businesses not only have a direct impact on the productivity and profitability of individual firms, but also influence macroeconomic variables like the gross domestic product, employment, or technological knowledge in a country. For this reason, this section of the document is dedicated to exploring the effect that good location choices of technological firms can have in both the microeconomic and macroeconomic environments, as well as to present possible applications of the project and introduce the personal motivation to develop it.

1.1.1. Microeconomic Impact

Choosing the right location is a key decision that all businesses must undertake, and that will determine their future success. Failing to do so can lead companies to business failure or even bankruptcy, which will encourage them to relocate or close operations. Therefore, the choice of a physical location is an extremely costly action to unmake or modify because:

- Relocating or partially migrating a business involves the transfer of employees and/or recruitment of new ones, the movement and/or purchase of machinery and hardware equipment, and the filing and compliance with the new location's legal requirements, among other costs.
- Closing operations involves selling all assets (including fixed ones, like real estate and land properties) or paying redundancy costs.

Due to the highly onerous consequences of a bad choice, we can consider a business' location decision to be essentially a sunk cost (a cost that has already been incurred and cannot be recovered). For this reason, it is critical for any business to carefully study its location options before taking a decision, according to firm and industry-specific requirements and needs.

1.1.2. Macroeconomic Impact

Promoting business creation and economic activity in a country has proven to directly impact national productivity and wealth. This idea can be extrapolated as well, and with special consideration, to the ICT industry, whose economic activity has gained significant

importance in the market over the last years. As an illustration, it will be shown later in this document how the gross domestic product of the European countries maintains a strong positive correlation (over 93%) with the number of projects developed in the ICT field. This is due to the influence technological firms have on the emergence of economies of agglomeration (the so-called *tech clusters*), along with the stimulation of determinants of productivity like technological knowledge or human resources.

1.1.2.1. Formation of Tech Clusters

The term *economies of agglomeration* describes the urban concentration of firms in locations where cost savings can naturally arise [2]. When referring to the technological industry, where collaboration and the exchange of knowledge play an essential role, the term *tech cluster* is preferably used. Similar to the effect of economies of scale, the costs and benefits of agglomerating increase the larger the urban cluster becomes [2]. Being the advantages of technological clustering (e.g. the mobility of qualified staff across employers, the sharing of local inputs, etc.) significantly superior to its disadvantages (the environmental disruption caused by this industry is very limited), tech clusters become a scenario any urban planner or government organism would like to achieve [3]. Some examples of tech clusters include *Silicon Valley* and *Greater Seattle*, both in the United States, which have gained worldwide technological leadership thanks to the business agglomeration and technological specialization developed in these areas.

1.1.2.2. Stimulation of the Determinants of Productivity

Other effects of the implementation of ICTs in the economy are the consequent generation of employment and the advancement of technological knowledge. The use of ICTs in business is paired with the rise of task automation. Many functions that were traditionally conducted manually are now performed by machines, which have two positive implications on society:

- **Increase in productivity**. Workers who can work with machines are more productive, reducing economic and time costs, which is reflected in the decrease in prices of goods and services. [4]

- **Job generation**. As a result of the reduced prices, consumers gain purchasing power —they feel richer because their money can buy more goods—, so they spend more, which leads to the generation of jobs. [4]

There are also negative socio-economic effects associated with the automation of tasks brought by the incorporation of ICTs in firms. The main one is the employment destruction carried out by the cessation in need of manual processes. Even if new job positions arise, the technical skills associated with the use of technologies are yet limited in society, making it hard for a considerable portion of the population to fit those employment demands. However, this problem is progressively being addressed over time as new generations arise and the population gets adapted and educated in the use of new technologies. To motivate the digitalization of the workforce, universities are increasingly offering academic programs based on and directed to the use of technologies, and public organizations are continuously leading formative initiatives on the matter, apart from supporting every legal proceeding to be done through digital channels.

1.1.3. Thesis Applications

The main planned application of this thesis is the support of any type of firm in the ICT industry in their location decision-making process. Whether it is a small start-up exploring the artificial intelligence field, or a consolidated, big firm that seeks to innovate so as to keep in line with business trends and technological updates, the purpose of this project is to recommend where to geographically locate operations in order to leverage the resources that a specific region can offer.

Another interesting application of the project will be the job search. A person who is willing to specialize in a certain technological area should be able to leverage the resources provided by this thesis to get hints of the regions where the demand for workers in that area is tentatively the greatest, thus increasing his or her chances of being hired. There are also many other uses that can be given to the tool developed in this thesis.

1.1.4. Personal Motivation

The idea for this project came from a personal interest in strategic consulting and data science. These two knowledge fields are fully complementary, since the efficient use of data

is key to the execution of a successful business strategy. Pursuing a bachelor's degree in Management and Technology incentivized the author's interest in both business and technology management. Moreover, taking advanced classes in programming unfolded the huge power and extent of application of computing. Similarly, reading and working in various digital companies enabled the author to observe the complexity of internal business operations. Therefore, this project was initialized with the objective of combining knowledge in both fields to provide real value to the ICT business environment.

1.2. Objectives

Based on the needs outlined in section *1.1. Motivation*, the overall objective of the present project is to provide a decision-making support solution for entities operating in the European ICT industry. To achieve this goal, a set of sub-objectives will be sought:

1) Analysis of the relevant business dimensions considered in the location of technological firms.

- Researching and understanding the business context of ICTs, including the technological structure of companies and the effect of using ICTs to drive strategy and attain competitive advantage.
- Selecting and analyzing the socioeconomic dimensions that determine the success of organizations within the ICT industry.

2) Socio-economic study of the different location options in the European framework based on statistical data in relation to the business dimensions of an ICT company.

- Studying the relationship between the implementation of ICTs and productivity in a country.
- Obtaining valuable data on which to base the recommendation solution to be made.

3) Definition of a method to optimize the location of any business in the ICT industry according to the socio-economic features of the European regions.

4) Implementation of a useful and reliable tool that incorporates the method defined in the previous objective.

- Exploring and applying the advantages the use of data mining techniques, languages and modules can provide to the development of a location recommendation.
- Deploying a ready-to-use web application that presents the system created in a user-oriented manner, allows data input and displays visual representations of the results.

5) Examining the method built.

- Assess the accuracy of the location recommendation system created by comparing the results obtained with data on the digitalization of the European regions.
- Assess the usability of the tool created by testing it on various typical use-case scenarios.

1.3. Glossary

1.3.1. Acronyms

- **AI**: Artificial Intelligence
- **EU**: European Union
- **GDP**: Gross Domestic Product
- **HTTP**: HyperText Transfer Protocol
- **ICT**: Information and Communication Technology
- **M&A**: Mergers and Acquisitions
- **R&D**: Research and Development
- **SQL**: Structured Query Language

2. Methodology

Based on the motivation outlined before of leveraging the benefits of technology to support the business strategy into a single project, work will be structured to cover two phases.

2.1. Business Analysis Phase

This phase is intended to study the socioeconomic situation of the ICT industry in Europe, with the aim of understanding the context and determining the business dimensions that influence the location of technological firms. This will not only define the factors to consider in the design of the recommendation system to be built, but also enable the discovery and exploration of data on the matter, which will be later used in the technological integration phase of the project. This will be achieved through the uncovering of the motivation behind the location decisions of businesses (*1.1. Motivation*), the exploration of the concepts surrounding the matter (*3. Theoretical Framework)*, the study of the existing literature (*4. State of the Art*) and the identification and description of the factors found (*5. Business Strategy*).

Additionally, the results obtained after the tool is built will be leveraged to serve as economic indicators of the business activity in the different regions of the European territory (*7. Further Economic Insights*).

2.2. Technology Integration Phase

After determining the specific business dimensions to consider in location decisions, a system to deliver personalized recommendations will be created (*6. Technology integration*). It will be done following the CRISP-DM (*CRoss-Industry Standard Process for Data Mining*) methodology, which is a standardized methodology used universally to develop data-mining projects. This methodology divides work into different steps, as described in the following table [6]:

CRISP-DM Phase	Summary	Objectives
Business Understanding	This initial phase focuses on understanding the business objectives	● Determine business and data-mining

	and requirements in which to base the data-mining project.	goals ● Assess business situation ● Produce project plan
Data Understanding	This second phase attempts to explore the available datasets so as to get familiarized with the matter and formulate ways of solving the problem.	● Collect data ● Describe data ● Verify data quality
Data Preparation	This phase includes all the steps needed to build the final dataset from the initial raw data.	● Select data ● Clean data ● Construct data ● Integrate data ● Format data
Modeling	This phase consists of the selection and application of data modeling techniques so as to come up with the project solution.	● Select modeling technique ● Build model ● Assess model
Evaluation	This phase attempts to evaluate the model created in the previous step with the aim of uncovering and fixing bugs, while ensuring the proper functioning and reproducibility of the algorithm.	● Test results ● Review process ● Determine next steps
Deployment	This final phase adapts and presents the model created to the final user.	● Plan deployment ● Plan monitoring and maintenance ● Produce final report

***Table 1**. Phases of the CRISP-DM methodology.*

With CRISP-DM, the project life cycle consists of six phases, although they do not follow a strict sequence. The following figure shows the project roadmap, with arrows indicating the most important and frequent dependencies between phases. [7]

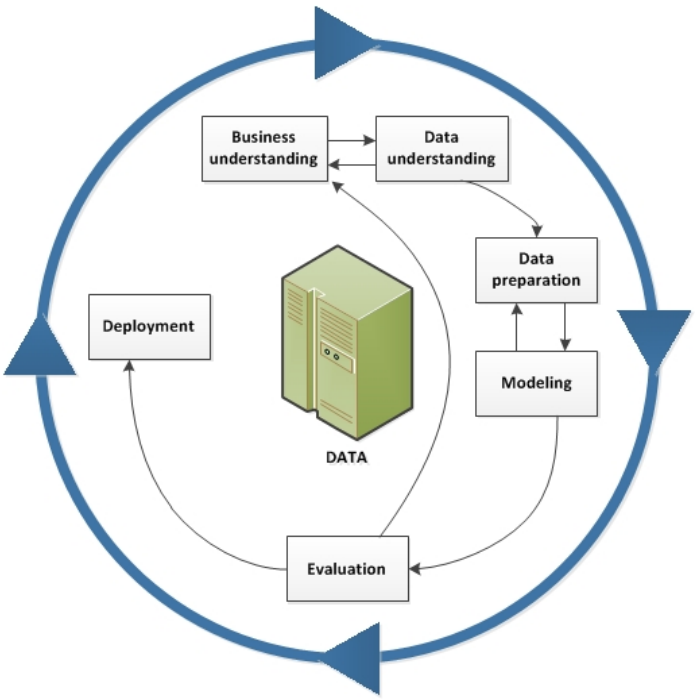


***Figure 1.** The life cycle of a business analytics project that follows CRISP-DM.* [7]

3. Theoretical Framework

In this section of the document all the relevant theoretical concepts involved in the subject of the project will be explained: what ICTs are, how they integrate with business processes to drive corporate strategy, and how location decisions take place in the business arena.

3.1. Information and Communication Technologies (ICTs)

ICT (Information and Communication Technologies) comprises the infrastructure of devices, networking components, and applications that enable modern computing. This field is the protagonist of the fourth industrial revolution (also referred to as *Industry 4.0)*, a concept that originated back in 2011 [8], which relies on interconnectivity, automation, real-time data, and machine learning to provide endless benefits to society. The main components of ICTs are the following:

ICT Component	Definition
Cloud Computing	*Clouds* are data centers available to different users over the Internet. They can be exclusive to a single organization (*enterprise clouds*), available to multiple organizations (*public clouds*), or a combination of both (*hybrid clouds*). [9]
Software	A *software* is a set of intangible instructions, programs, or data used to operate computers and execute tasks. [9]
Hardware	*Hardware* consists of all the physical and tangible elements that make up a computer or electronic system. This includes computer monitors, hard drives, memories, or the CPU (*Central Processing Unit*). [9]
Digital Data	*Digital data* refers to data that can be interpreted by various technologies. The most common data system is the binary system, which stores information (text, video, images, etc.) in the form of binary values (ones and zeros, or "on" and "offs"). [9]
Internet	The *Internet* is a global network of computers and other electronic

	devices, through which users communicate and share information. [9]

***Table 2**. Components of ICT.*

3.2. ICTs in the Business Environment

ICTs have impacted the way firms manage their operations in multiple ways, producing the benefits highlighted in the table below.

Business benefit	Explanation
Better decision-making	Thanks to ICTs, firms can store, clean, and process vast amounts of data, which is then transformed into valuable information. This information is then used to make decisions quickly and accurately, enabling rapid response to business opportunities and threats. [10]
Increased productivity	ICTs allow the automation of repetitive tasks, the accurate planning of production resources, the measurement of employee performance, and the optimization of manufacturing operations, among other improvements in the drivers of productivity. All these advantages lead to great cost and time savings. [10]
Improved customer service	The use of ICTs to collect and process market and customer data (purchase history, web tracking, etc.) enables firms to deploy user-centered applications, deliver personalized services, target promotions and provide fast-response customer support. All these applications boost client satisfaction and retention, an essential element in every business. [10]
Virtual collaboration	Communication networks enable individuals to effectively collaborate with each other from different locations. Videoconferencing sites permit holding virtual meetings, while collaboration software tools make it easy to jointly create files,

	distribute tasks, track progress, and share information. This contribution can happen between individuals within the same organization (project team members), or with externals (like suppliers or other business partners). In any way, the resulting improved collaboration creates stronger teams, which reduces overall project development time and effort. [10]
Improved financial performance	ICT solutions can help an organization minimize costs and maximize revenue, which optimizes business profitability. Analyzing stock data, predicting business performance in the market, or testing different investment alternatives are all ways in which companies can leverage ICTs to evaluate financial opportunities and mitigate risk. [10]

Table 3. *Benefits of implementing ICTs in a business.*

All the business benefits explained contribute to one or both the competitive strategies defined by Michael Porter in 1980-1985: *cost-leadership* and *differentiation*. Cost-leadership strategies aim to sell a standard product at the lowest price possible by attaining economies of scale and minimizing costs on non-manufacturing activities like R&D, sales, and advertising. On the other hand, differentiation techniques seek to offer a unique product or service in exchange for a price premium. Both strategies help create a competitive advantage for the company, which results in an increased market share [11].

Due to the essential and impactful role technology plays nowadays in businesses, an increasing number of companies are establishing entire departments dedicated to managing their ICT operations. The CIO (*Chief Information Officer*) or CTO (*Chief Technology Officer*) is the president of the department, in charge of all the operations that compound the ICT system of a firm. The following figure shows the teams that are generally part of the ICT unit in a company.

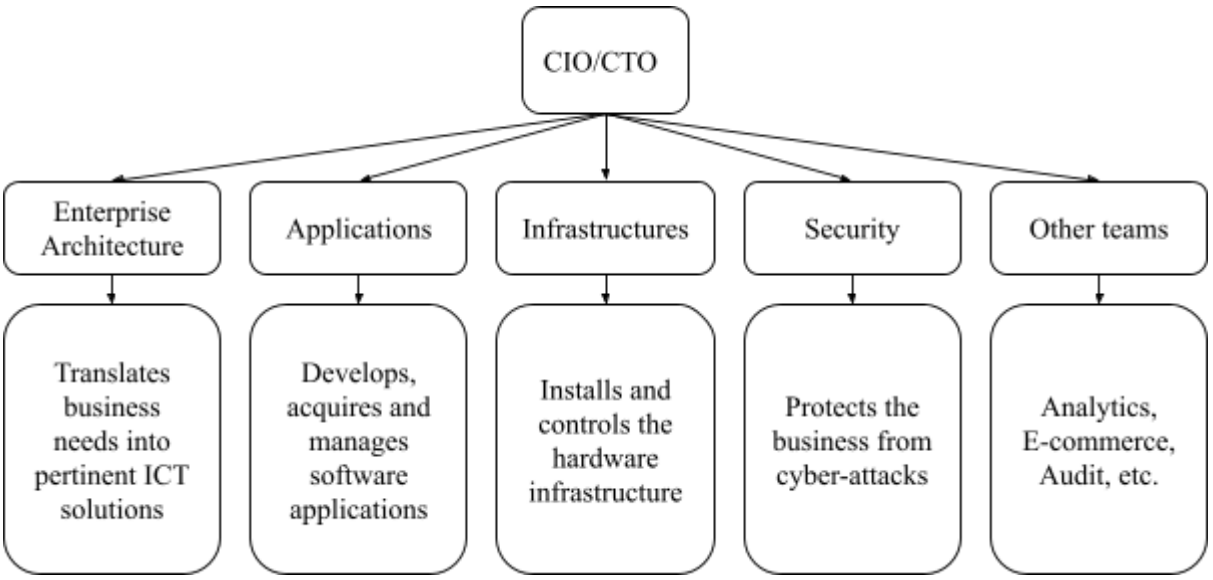


Figure 2. *Mainstream team distribution in the ICT department of a firm.*

3.3. Business Location Decisions

The location of a business refers to the geographical place where the company will provide services or sell products. For a small local restaurant, that will be the physical space where it welcomes its customers and serves tasty dishes. For a digital firm, the location will be where the main office is located, since that is where the software is developed.

Business location decisions can be classified into location or relocation settings, depending on the lifecycle phase when it occurs. *Location* decisions occur when a company searches for places where to locate its operations for the first time, while *relocation* decisions occur when a company that was already settled down somewhere decides to move its location (this can happen for multiple reasons, like an M&A, changing economic conditions or business expansion). Also, other levels of location decisions can happen in between, like *partial migration* (only a subset of a business changes its location). [12]

4. State of the Art

4.1. Factors Influencing the Location Decisions of Firms in the ICT Industry

As mentioned previously, the existing literature on the factors that drive location decisions of firms in the ICT industry is yet scarce, due to the recent incorporation of this special industry into the market. Despite the limitations, two main papers were identified when exploring the existing research on this topic.

The first one, *Selecting the location for a new business unit in the ICT industry* (Sanja & Nikolić, I. & Rakicevic, Jovana., 2018), found the following location determinants important:

- In a quantitative context → capital investment costs and operational costs. [13]
- In a qualitative context → political and economic environment, legal framework, and competition. [13]
- In an ICT-specific context → human resource availability, infrastructure availability, and cultural compatibility. [13]

In addition, the paper provides a ranking of the factors, assigning human resource availability the highest overall priority weighting, 28,4%. The political and economic environment was ranked second weighing 18,2%, followed by the competition (14.5 %). The study also suggests an interesting fact: political and economic strategies are starting to play an important role in the ICT market, as governments are getting more interested in this industry. [13]

The second resource, *A Study of the Factors Influencing the Location Selection Decisions of Information Technology Firms* (Rajkumar, P., 2013), studied the parameters that influence the location decisions of Indian technological firms The study identified the following dimensions [14]:

1. Manpower factor → Determined by the availability of manpower, the education systems, the educational level of manpower, the skill level of manpower, the wage rate and the labor union.
2. Technological factor → Defined by the infrastructure for the latest technology, the availability of the latest technology, and the telecommunication facilities.

3. Social factor → Explained by the quality of life, the standard of living of the community, the housing availability in the community, the attitude of community residents, the availability of quality schools, the availability of colleges and research institutions, the medical facilities, and the shopping centers.
4. Hedonistic factor → Described by the pleasant working environment, the recreational facilities, the relative humidity, and the monthly average temperature.
5. Industrial site factor → Determined by the cost of industrial land, accessibility, the closeness to the city, the infrastructure (roads, electric power, water and sewer, utilities, etc.), the cost of electric power and other utilities, the proximity to other high-tech firms, the airway, highway and railroad facilities and the transportation cost.
6. Economic factor → Defined by the corporate tax structure, the tax assessment basis, the government aids, incentives, and tax-free operations, and the local business regulations.
7. Governmental Factor → Explained by the governmental regulations, the prevalence of bureaucratic red tape, political stability, congenial environment for business, and supportive local authorities.

The econometric model conducted (using regression analysis) concluded that all seven factors play a significant role in the decision of a location. The manpower and technological dimensions are very predominant, while the hedonistic one is the least relevant in the decision-making process. [14]

Both papers agree some factors are the most influential in the decision-making process of firms in the ICT industry. According to both sources, the human resource dimension is the most important one. Other factors referenced as relevant in both studies are the technological infrastructure, the influence of government, the capital or economic structure, and the industry environment. Therefore, this project will be focused on analyzing and working on those five identified factors as the main drivers of location decisions for companies in the ICT industry.

4.2. Level of Digitalization in Europe

In an attempt to evaluate the suitability of the European territory to accommodate technological firms, the existing data on the matter is examined here.

The European Commission has been monitoring member states' digital progress through the Digital Economy and Society Index (DESI) reports since 2014. Each year, DESI includes country profiles that support member states in identifying areas requiring priority action as well as thematic chapters offering a European-level analysis across key digital areas, essential for underpinning policy decisions. [15]

Based on the DESI annual reports from 2014 to 2019, Deloitte prepared a report analyzing the impact of directing the EU's investment in the digitalization of Europe [16]. The results showed that an increase of 10% in the overall score assigned by the DESI to the countries in the EU27 group and the UK was equivalent to an increase of up to 0,65% of the GDP per capita (this is, the economic value of the possessions of the average resident of a country), keeping the rest of factors constant (employment, public expenditure, etc.).

In light of the observed benefits of digitalization, the EU started to support and incentivize the creation of R&D projects by financing programs and competitions, among which *Horizon 2020* stands out. This program seeks to raise funds from all participating countries of the EU27 —as well as from other associated countries outside the group—, and distributes these funds among the participants, in proportion to the relevance of the projects presented [17]. Each of these projects is evaluated by the program's committee of examiners in terms of the quality of the work, innovation, and application of the proposed technology. As a result, each project submitted receives a certain amount of European funds, which does not depend on the investment made by each country. Therefore, the amount contributed and received by each country are independent, motivating the participant countries to present the most quality projects to obtain the highest amount of funds possible.

Due to the relevance and scope of the program, the number of projects presented by a country can indicate the actual technological failure or success of the nation. For this reason, *Horizon2020* data will be used in this thesis to evaluate the European regions' digitalization level.

5. Business Location Strategy

While all companies need to consider multiple, sometimes hard-to-analyze factors in their location decisions, not all businesses give the same importance to the same factors. For example, a sunlight energy provider will need to study where the most sunlight can be obtained (place with the highest sun intensity or with the longest daytime), but also where energy-transfer infrastructures are available, or the geographical distance to energy consumers. On the other hand, a restaurant will choose to operate in a place with an adequate flow of individuals (like the city center) and where food ingredient suppliers can be accessed. Every company considers different parameters in their location decisions, according to both their business needs and their unique characteristics.

This section of the paper aims to clearly outline and describe all the factors that determine the location decisions of companies in the ICT field. This will include both the factors considered to be *industry-specific needs* (shared by all the organizations in the ICT industry) and the factors considered to be *firm-specific characteristics* (unique to each company).

5.1. Industry-Specific Needs

In the case of firms that heavily or entirely rely on information technologies for operating, some special factors are considered, which differentiate them from businesses in other industries that are more dependent on natural resources. This is due to the fact that companies in the ICT industry mostly sell intangible products or services (based on software), and make use of virtual channels to deal with business stakeholders (suppliers, distributors, customers, employees, etc.). However, likewise, certain factors play an important role in determining the location of these special businesses. The five main ones, identified through the review of the existing literature (*4.1. Drivers of the location decisions of technological firms*), are explained hereunder:

1. **Human resources.** In all industries, but especially in those like ICT —in which advanced technical knowledge is required—, attracting, recruiting, and retaining the right people is key. Innovation and technological advancement requires specialized education, and the supply of highly qualified profiles in the areas of engineering, computer science, and physics is limited. Like Tom Stringer —corporate real estate

advisor for BDO— said, "in the tech space, the hunt for talent is pretty extreme. That is driving a lot of the [location] decisions that are being made. They're looking at the university systems in the area and whether the ecosystem is there to support the workforce flow that they need" [18]. This implies that the availability and qualification of human resources in a region play an essential role in the selection of an ICT business location.

2. **Technological infrastructures.** Another key component of every ICT business is its technological infrastructure. The better hardware, device interconnectivity, and electrical power they have, the faster they can perform their computing tasks. Since all operations these firms perform are based on their technological infrastructure, they need to carefully study the availability of these resources in their location decisions. The availability of high-tech networks (like 5G), the price of electricity or the suitability of the space to support computer servers are some examples of factors that need to be taken into consideration.
3. **Capital structure**. The technological industry, being young in the market, relies heavily on capital investment. This capital is mostly concentrated in urban areas, where banks and other financial institutions exist in large numbers. Entrepreneurs are interested to locate their businesses where the offer of capital is abundant, so as to get credit at cheaper rates of interest. For this reason, the availability of capital is also a factor that greatly influences the location of technological firms. [19]
4. **Governmental influence.** The government establishes legal measures and public initiatives that can either favor or restrain the economic activity in a location. This affects the ability of technology firms to access or use certain resources (e.g. customer data), leverage tax concessions, or benefit from grants and other public opportunities [19]. During the last years, after observing the socioeconomic benefits the ICT industry can bring to a territory, many governing institutions are implementing public policies to boost the placement of technological companies in their locations (see *8.1. Legislation*).
5. **Industrial environment**. The collaboration between institutions (companies, universities, and technology centers) is especially significant in the ICT industry. When organizations specializing in different technologies associate, synergies are created that allow the research and development of more complex products. Moreover, this agglomeration of technological specialization attracts capital

investment and talent, which promotes further collaboration in an endless cycle, eventually driving technological prosperity and productivity in a region. This is what causes the previously explained tech clusters (see *1.1.2.1. Formation of tech clusters*). On the other hand, the physical closeness to competing firms is not a major matter of concern in the ICT industry. Since the product is delivered using digital channels, the customer does not need to physically move to the place of sale, thus the agglomeration of firms cannot have a significant negative effect on the firm's sales.

While the critical factors affecting the location decisions of technology firms are covered by the indicators explained, the list should not be considered exhaustive. These five were selected from an endless list of factors due to two reasons. Firstly, the selected indicators reflect the current availability of data, given the yet-highly-unexplored technological industry [20]. Secondly, some factors –not considered in this study– are tied to subjective judgment and personal location preferences of the decision-makers, meaning they cannot be measured accurately and in a standardized way.

5.2. Firm-Specific Characteristics

Every firm, independently of the industry of operation, has unique characteristics. To narrow down the definition of a firm into three dimensions, this project will consider the aspects described below.

- **Market area**. The Information and Communication Technologies industry encompasses the use of many different types of technologies: big data, cybersecurity, media and communication, robotics, etc. An organization focused on a specific technology (e.g. artificial intelligence) may be interested in targeting those regions where the specialization in that area is the highest. This way, it could take advantage of the existing tech clusters, have closer contact with potential partners, learn from observing the market functioning, and, in general, have a greater chance of success.
- **Company size**. Another aspect in which firms differ is their size, typically measured in the number of employees. Business managers consider this characteristic in their location decisions because that can greatly influence their relationship with other companies. Some people may prefer locating their businesses in locations with a greater proportion of start-ups and other types of small institutions, since this way collaboration and resource sharing can be perceived as more accessible or achievable.

Similarly, managers of larger companies could prefer locating their ICT operations close to other firms of the same size, since this can open up the possibility of M&A, deals, and negotiations. Competitiveness in an industry is also a factor that is greatly influenced by the size of its components. Larger entities tend to hold greater market shares (they can be conceived as the industry "leaders" because they managed to attain a competitive advantage that led them to grow into great size), while smaller companies are usually "followers" or are focused in a niche (a small subset of the market that demands special products).

- **Nature of work**. Firms, apart from specializing in a market area of operation, also specialize in the type of work they conduct in the market. This work can either be exploratory, constructive, or integrative (more details on this later). This is an interesting factor to analyze, because some business managers may strategically attempt to locate their operations close to other purpose-specific institutions. For example, a firm that conducts integrative tasks –by applying technology to products– can be interested in setting their operations near institutions that conduct exploratory work (like *deep-techs*), as they are the ones inventing the technology they will later incorporate into products.

The definition of a business according to the three characteristics explained allows the determination of business-specific location demands. This will be useful to configure the personalized location recommendations in the system to be built.

6. Technology Integration

This section studies the development and deployment of a location recommendation system. As previously indicated, the CRISP-DM approach will be followed.

6.1. Business Understanding

This section explores the business requirements to be assessed by the recommendation system to be built and come up with an action plan. This includes the design of a solution to attain the desired outcome (business goals), the study of the existing resources to develop the project (resource assessment), and the technical plan to reach the desired outcome (implementation plan).

6.1.1. Business Goals

After reviewing the state of the art on the matter, it can be observed that there are currently no definitive indicators or guidelines that help select the best location for a business within the scope of the European ICT industry. There is still a need to create a tool that will enable the recommendation and visualization of different location options depending on the individual firm's features. With this intent, the requirements for the tool from a business perspective, and their corresponding technical solutions are the following:

- **Tool that allows customization**. Not all businesses in the ICT industry are the same —each company focuses on a different technology, some of them require highly-qualified personnel while others do not, they also differ in size, etc—. Therefore, there is a need to create a tool in which the location search can be narrowed down to match specific business characteristics.
- **Tool that is accurate**. Due to the complex and extremely competitive nature of the current business environment, it is critical that the tool can give the best recommendations and can deliver a high success rate.
- **Tool that is simple to use**. The recommendation, apart from being complete and accurate, must be delivered in a clear way, so that any business decision-maker can easily understand what the results obtained are, and decide on the next steps to take. To do this, the tool needs to deliver results both quantitatively and visually.

6.2.3. Resource Assessment

6.2.3.1. Tools

To create a tool that addresses the identified business goals, data mining and business analytics skills will be required. The programming language *Python,* commonly used in the data science field, offers multiple advantages to do so with modules like *pandas* (for data analysis and manipulation [21]), *geopandas* (for the use and representation of geospatial data [22]), or *numpy* (for scientific computing [23]). Moreover, open-source notebooks like *Project Jupyter* or *Google Colab* offer convenient coding environments where to build the tool.

6.2.3.2. Scope of Analysis

The scope of study of this project will be constrained to the European territory. The possible location options will be regions classified under level 2 of the Nomenclature of Territorial Units for Statistics (NUTS). The reasons for the selection of this location division are the following:

- To narrow down the data used to a specific geographical area, so as to attain a balance between bias and accuracy. By focusing on regions instead of countries (NUTS 1 level), we acquire a level of detail that brings significance to the application of the study, since countries are ample and their technological parameters can greatly vary from one region to another. At the same time, by studying regions instead of cities (NUTS 3 level), we favor the availability of data, reducing the bias and providing reliability to the study.
- The NUTS 2 level is commonly used by the EU members for the application of their public policies, and thus is the most appropriate nomenclature to analyze the different business factors that influence location decisions, especially in the case of the governmental influence factor. [24]

6.1.3. Implementation Plan

6.1.3.1. Technique

Considering the business goals of the project, the most reasonable solution is to create a matchmaking tool that will compare the user's business requirements with the characteristics of each European region, and return the locations with the highest degree of similarity.

6.1.3.2. Resources

To create the matchmaking system, the resources needed are:

- **A matrix of regional vectors**. Each European region will have a unique vector formed by *m* dimensions of values between 0 and 1. Each dimension will correspond to a business parameter to be evaluated, and their values will be the score those parameters have in each region. An array will be built containing all these evaluation vectors, so its size will be *k* rows (*k* is the number of European regions to be evaluated) and *m* columns (*m* is the number of business parameters to evaluate).
- **An input vector reflecting the user preferences**. Its values will be provided by the tool user. This input vector will be formed by *n* dimensions (each corresponding to a business characteristic to be selected) of values equating to 0 or 1. The vector will then be adapted to match the number of dimensions of the evaluation vectors (*m*). To do so, 1s will be added to the parameters that do not depend on business characteristics, but rather affect the ICT industry as a whole (capital, human resources, innovation, government, and technical infrastructure).
- **An input vector reflecting the weight of each evaluation block**. Its values will be optionally provided by the tool user. This input vector will be formed by *l* dimensions (each corresponding to the weight or importance associated with each business factor) of values between 0 and 1 (being 0 a null importance and 1 the maximum importance). The sum of all the weights will equate to exactly 1. If not provided, the weight vector will be automatically generated, giving the eight evaluation blocks the same importance (this is, all the dimensions in the vector will have the value of ⅛). Finally, the vector will be adapted to match the number of dimensions of the ending input vector of search preferences (*m*).

- **A target vector**. The target vector will represent the definitive user's customized search preferences. It will have *m* dimensions (one for each parameter to be evaluated in the match). The target vector will be created by multiplying the values of the two final input vectors (the one reflecting the search preferences and the one reflecting the importance of each evaluation block).
- **An output matrix of score vectors**. This will be the array of vectors returned by the matchmaking system. It will correspond to the ranking of regions recommended by the tool. The array will contain the measure of similarity of each regional vector with the target vector, so its size will be *k* rows (*k* is the number of European regions evaluated) and one column (this will correspond to the similarity score).

6.1.3.3. Design

Following the analysis conducted on the factors that influence the location decisions of businesses in the ICT industry, the variables selected to represent the *m* dimensions of the vectors will be classified into eight evaluation blocks. Out of these blocks, some will be selected by the user, whereas others will not require a selection of values. The elective blocks will be:

1. *Technological area.* This block intends to observe the technological specialization of each region.
2. *Company size.* This block intends to observe the size of the entities in a region.
3. *Technological maturity.* This block intends to observe the work purposes of the entities in a region (either exploratory, constructive or integrative).

On the other hand, there will be non-elective blocks. These correspond to the factors found to be drivers of business location decisions for firms in the ICT industry, as indicated previously in the study. It will be assumed that the greater the values for the variables in these blocks are, the greater similarity to the input vector will the regional vectors have. The non-elective blocks are the following:

1. *Capital.*
2. *Human resources.*
3. *Innovation ecosystem.*

4. *Legal framework.*
5. *Technological infrastructure.*

Each evaluation block can be subdivided into multiple parameters, which will all together correspond to the *m* dimensions of the target vector and all the regional vectors. With all this data, the match will be made, returning a ranking of *k* similarity measures (one for each regional vector). This ranking, sorted in descending order, will represent the final location recommendation, delivered as a table providing the score each region receives when compared to the target vector.

The following table summarizes the matchmaking system design, with the eight evaluation blocks in gray, the twenty-one evaluation parameters (electives in green, non-electives in yellow), the input vectors in blue (the search-preferences vector will contain fourteen values of zeros or ones, whereas the weights vector will contain eight values between zero and one), the created target vector of twenty-one dimensions in orange, and the evaluation vectors in red (one for each region).

Dimension	Technological Area									Company Size		Technological Maturity			Capital		
P	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
Parameter	Artificial intelligence (AI)	Big Data	Computation	Cybersecurity	Internet	Internet of Things (IoT)	Media & Communication	Robotics	Software	Small and Mid-Size Enterprise (SME)	Large Enterprise (LE)	Deep Tech	Development	Integration	EU Grants	Capital Investment	
User Preferences	✓				✓					✓		✓	✓				
Search Vector	1	0	0	0	1	0	0	0	0	1	0	0	1	0	1	1	
Weights Vector	0.2									0.1		0.05			0.15		
Target Vector	0.1	0	0	0	0.1	0	0	0	0	0.1	0	0	0.05	0	0.05	0.05	
Region 1	0,3	0	0,1	0	0	0	0	0	0	0,01	0,2	0	0,4	0,3	0,01	0,2	
Region 2	0	0	0,3	0	0,1	0	0	0	0	0,2	0,1	0	0,1	0,2	0,01	0,1	
...																	
Region N	0	0	0	0	0	0	0,2	0	0	0,3	0	0,4	0	0,1	0,02	0	

Figure 3. *Matchmaking system design.*

6.2. Data Understanding

6.2.1. Data Collection and Description

To fulfill the requirements of the project, four databases will be used, as detailed below.

1. ***ICT_H2020.xlsx***. This dataset, in Microsoft Excel Open XML Spreadsheet (XLSX) format, was extracted from *Data.europa.eu* in April of 2022 [25]. It includes valuable data about the ICT projects presented and the entities that participated in the Horizon 2020 program. For the purposes of this project, only those belonging to the information and communication technologies were selected. The entries correspond to the entries received in the Horizon 2020 program, being a single entity able to participate in several projects, and a project able to be developed by multiple entities. The program in question received entries from 2014 to 2020, therefore the entries in this dataset were added at different times. The values are in absolute terms, so they will eventually be normalized to fit in the proposed model.

	Framework Programme	Call ID	Topic Code	Topic Descr	Project Nbr	CORDIS link	Project Acronym	Project Title	Partner Role
0	H2020	ECSEL-2014-1	ECSEL-01-2014	ECSEL Key Applications and Essential Technolog...	661933	http://cordis.europa.eu/project/id/661933	RobustSENSE	Robust and Reliable Environment Sensing and Si...	COORDINATOR
1	H2020	ECSEL-2014-1	ECSEL-01-2014	ECSEL Key Applications and Essential Technolog...	661933	http://cordis.europa.eu/project/id/661933	RobustSENSE	Robust and Reliable Environment Sensing and Si...	PARTICIPANT
2	H2020	ECSEL-2014-1	ECSEL-01-2014	ECSEL Key Applications and Essential Technolog...	661933	http://cordis.europa.eu/project/id/661933	RobustSENSE	Robust and Reliable Environment Sensing and Si...	PARTICIPANT

Figure 4. *Preview of the database ICT_H2020.xlsx*

2. ***Subfactor score table.csv***. The second dataset contains information on the readiness of European countries to deploy and adopt 5G technologies. It is in Comma-Separated Values (CSV) format and was collected in April 2022. This dataset was obtained from a report published by *Incites Consulting S.A.* in 2020, and gives the different European countries a score (in relative values) for each of the factors that determine the level of digitalization of a nation [26].

Country	4G coverage	Fiber coverage	Internet BW per user	5G commercial networks	# of IXPs	# & maturity of 5G pilots	Time to get electricity	4G launch year	5G spectrum auction plans	...
Finland	89.5	37.5	62.05	100.0	10.11	30.15	86.55	90.91	66.67	...
Switzerland	89.6	30.3	61.50	100.0	30.32	23.60	87.54	72.73	66.67	...
Germany	76.9	8.5	55.92	100.0	80.85	64.24	91.15	90.91	66.67	...
Denmark	88.6	64.4	62.59	25.0	12.63	10.49	87.87	90.91	33.33	...
Sweden	91.1	72.2	58.91	25.0	45.48	7.87	83.27	100.00	50.00	...

Figure 5. *Preview of the database Subfactor score table.csv*

Since data at a regional level —rather than at a national level— is needed, it will be assumed that all the regions in a country have the same scores.

3. ***digital-innovation-hubs.xlsx***. This dataset, in XLSX format, was extracted from Eurostat in May of 2022 [27]. It quantifies the number of tech hubs each region in Europe has (in absolute terms, so they will be normalized). The source collected the data needed to construct this dataset in 2020.

DIH Name	Location	City	Country	Website	NUTS2 Code
Aachen DIH Center for Robotics in Healthcare	Pauwelsstraße 30	Aachen	Germany	http://www.robotics.ukaachen.de/	DEA2
Aalen University / Transfer Platform Industry 4.0	Beethovenstraße, 1	Aalen	Germany	http://www.hs-aalen.de/	DE11
Aarhus University Centre for Digitalisation, B...	Finlandsgade 22	Aarhus	Denmark	http://www.digit.au.dk	DK04
Accelerating Photonics innovation for SME's (A...	Brussels Photonics B-PHOT, Vrije Universiteit ...	Brussel	Belgium	http://www.actphast.eu	BE10
ADAPT Centre	O'Reilly Building, Trinity College Dublin	Dublin	Ireland	https://www.adaptcentre.ie/	IE02

Figure 6. *Preview of the database digital-innovation-hubs.xlsx*

4. ***unicorns.xlsx***. This dataset, in XLSX, was released by CBInsights in 2022 [28]. It quantifies the number of unicorn start-ups in each city in the world (in absolute terms, so they will be normalized). Since data at a regional level —rather than at a city level— is needed, the region to which each city belongs will be obtained, and it will be assigned that number of unicorns.

Company	Valuation	Date Joined	Country	City	Industry
1047 Games	$1.5	9/14/2021	United States	Zephyr Cove	Internet software & services
1KMXC	$1	8/30/2021	China	Hangzhou	Hardware
1Password	$6.8	7/8/2021	Canada	Toronto	Cybersecurity
4Paradigm	$2	12/19/2018	China	Beijing	Artificial intelligence
56PINGTAI	$1.08	1/25/2021	China	Shanghai	Supply chain, logistics, & delivery

Figure 7. *Preview of the database unicorns.xlsx*

6.2.2. Data Quality Verification

Before starting to process the data, a quality check will be conducted on each dataset, with the aim of assessing whether they fulfill the data needs of the project (data relevance) and whether the provided information is reliable (data reliance).

ICT_H2020.xlsx

This dataset was initially selected because, after observing how ICTs increase the productivity of individual businesses, it could be thought that these technologies should also have a positive impact on the aggregate productivity of a country. Therefore, to quantitatively prove that the data on *Horizon2020* entries is a relevant source of information from an economic perspective, the Gross Domestic Product (GDP) of the European countries will be analyzed.

By computing the correlation coefficient between the total European investment in H2020 projects presented by each country and the GDP of these countries (see *Computation of the correlation between the investment in H2020 projects and the European countries' GDP* in the Annex*)*, it can be demonstrated that the use of ICTs in businesses (explained by the investment in H2020 projects) also affects overall national wealth. Since a correlation coefficient defines the strength of the relationship between the two variables studied, getting a high value (close to 1) would mean that a country's investment in ICT is strongly related to its GDP.

The resulting correlation coefficient was 0.93, which concludes that the use of the *ICT_H2020.xlsx* dataset is reasonable for the purpose of this project.

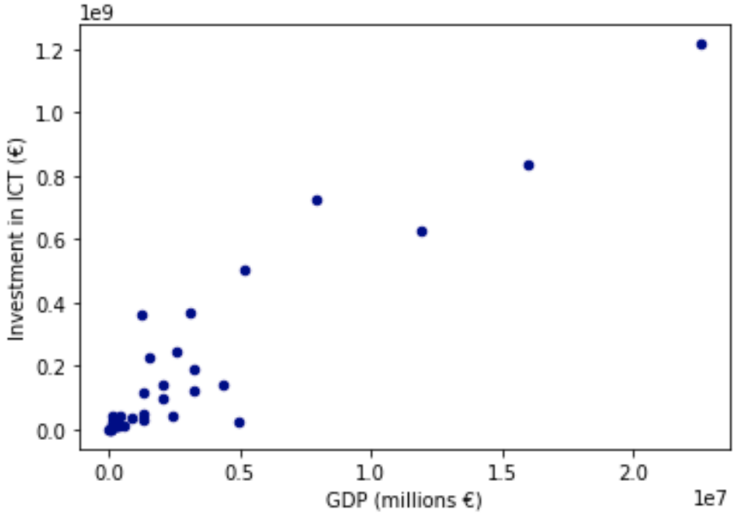


Figure 8. *Correlation between the investment in ICTs (in euros) and the GDP (in millions of euros) of the European countries.* Self elaboration.

The dataset is also reliable because it was elaborated by a trusted source —the European Commission, which is also the organizer of the Horizon2020 program—.

Subfactor score table.csv

The information provided by this dataset is directly related to the purpose of the project —obtaining a vector of scores for each region—, so assessing whether it fulfills the data needs will not be necessary. The data reliance is also confirmed after observing that the sources the dataset creator used were recognized organizations like the World Bank, the World Economic Forum, UNESCO, or the European Commission. [29]

digital-innovation-hubs.xlsx

This dataset is relevant to the study because the presence of tech hubs, by definition, is a direct indicator of the ease of operating a technological business in a location. This is because a tech hub is a physical space —whether it is a city or a suite of offices— designed to foster and support tech start-ups [30]. It is also reliable because it was published by Eurostat.

unicorns.xlsx

This dataset is relevant to the study because the presence of unicorns is an indicator of favorable investment conditions in a location, since unicorns are start-ups that started small and used vast amounts of capital to grow exponentially, until attaining a value of a billion dollars [31]. It is also reliable because it was published by a trusted institution.

6.3. Data Preparation

6.3.1. Data Selection

The four datasets analyzed include data on a different number of European regions. To standardize this and have information from all the datasets for the same number of European regions, only those regions for which data is available in all four datasets will be considered in the study. This corresponds to 280 European regions, identified by their unique NUTS 2 code. 218 of those selected regions belong to countries in the EU 27 group (Germany, Austria, Belgium, Bulgaria, Cyprus, Czech Republic, Croatia, Denmark, Slovakia, Slovenia, Spain, Estonia, Finland, France, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, and Sweden), whereas the remaining 62 belong to countries outside the EU 27 (Liechtenstein, North Macedonia, Norway, Turkey, Switzerland, Iceland, Montenegro, Serbia, United Kingdom, Albania). The EU 27 group is formed by the twenty-seven countries that currently form the European Union, as of June 2022. [32]

6.3.2. Data Cleaning

All the datasets contain extra information that is not necessary for the study. Therefore, the data frames are filtered to keep only the relevant data (this is, the columns that communicate information).

6.3.3. Data Construction

The next step is to attain the regional values for each of the business parameters of interest. From the datasets available, to suit the eight evaluation blocks proposed, twenty-three parameters can be obtained. For each of these parameters, a dictionary will be created, with keys corresponding to regional NUTS 2 codes and values being the values of that parameter for the corresponding region. The next tables explain, for each evaluation block, what parameters will be obtained and why, the computation for each of them to get the dictionaries, and the data source.

Evaluation Block 1 → *Technological Area* (elective)

Using the information provided by the column “Area” in the *Horizon2020.xlsx*

database, we can determine the market area of operation of the participant entities. Each participation in H2020 can be categorized into multiple different market areas, but ten general classes were selected to simplify the classification. The classes are detailed hereunder:

- **Artificial Intelligence** (*AI*) refers to the set of tasks that machines perform simulating human behavior. Therefore, all projects working on imitating or augmenting human capabilities (such as visual perception, speech recognition, or decision-making) will fall under this category. This includes any project in the following AI subtopics: reactive machines, limited memory, theory of mind, self-aware, artificial narrow intelligence, artificial general intelligence, or artificial superintelligence. [33]
- ***Big Data*** projects are those which handle extremely large volumes of data to reveal patterns, trends, and associations.
- ***Computing*** projects refer to those working on creating or expanding the operating capabilities of machines. This class of projects is focused on the study and experimentation of algorithmic processes and the construction of computer systems.
- ***Cybersecurity*** projects are those focused on building security systems to protect devices (computers, mobile devices, etc.) and systems (networks, applications, etc.) from the criminal or unauthorized use of electronic data. [34]
- ***Internet*** projects are those relative to the enhancement of the communication architecture and protocols, making them more secure, private, and decentralized. This type of project includes those involved in Blockchain technologies (like those presented by Fintech companies), along with those related to privacy and identity management. [35]
- ***The Internet of Things*** (IoT) is understood to be the interconnection via the internet of computing devices embedded in everyday objects, enabling them to send and receive data. [36] Therefore, projects categorized in the market area "IoT" are intended to design and develop those physical objects with sensors, processing ability, software, and other technologies.
- ***Media & Communication*** refer to those projects specialized in dealing with electronic communication channels. This includes all the devices used to store data (hard drives, CD-ROMs, diskettes, etc.) as well as the ones used to transmit it (cables, wires), or propagate it in its many forms (videos, sounds, podcasts, etc.). Projects using the Internet for communication purposes (social media) also fall into this category. [37]

- ***Robotics*** is a field of ICTs dedicated to the design, construction, operation, and application of robots. Therefore, projects in robotics work with robots in any field (aerospatial, industrial, medical, educational, drones, entertainment, etc.). [38]
- ***Software*** projects are those focused on developing computer programs. This category includes both applications software (word processors, database management software, sight and sound programs, and Internet browsers) and system software (operating system, gadget drivers, firmware, and utilities). [39]

Parameter	Definition	Computation	Data Source
AI (Artificial Intelligence)	Number of participations in AI-related H2020 projects in each region.	Count of entries whose value in the column "Area" is "AI" among the entries with the same value in "NUTS 2 Code"	*Horizon2020.xlsx*
Big Data	Number of participations in big-data-related H2020 projects in each region.	Count of entries whose value in the column "Area" is "Big Data".	*Horizon2020.xlsx*
Computing	Number of participations in computing-related H2020 projects in each region.	Count of entries whose value in the column "Area" is "Computing".	*Horizon2020.xlsx*
Cybersecurity	Number of participations in cybersecurity-related H2020 projects in each region.	Count of entries whose value in the column "Area" is "Cybersecurity".	*Horizon2020.xlsx*
Internet	Number of participations in internet-related H2020 projects in each	Count of entries whose value in the column "Area" is "Internet".	*Horizon2020.xlsx*

	region.		
IoT (Internet of Things)	Number of participations in IoT-related H2020 projects in each region.	Count of entries whose value in the column "Area" is "IoT".	*Horizon2020.xlsx*
Media & Communication	Number of participations in media or communication-related H2020 projects in each region.	Count of entries whose value in the column "Area" is "Media & Communication".	*Horizon2020.xlsx*
Robotics	Number of participations in robotics-related H2020 projects in each region.	Count of entries whose value in the column "Area" is "Robotics".	*Horizon2020.xlsx*
Software	Number of participations in software-related H2020 projects in each region.	Count of entries whose value in the column "Area" is "Software".	*Horizon2020.xlsx*

***Tabe 4**. Computation of the different parameters in the "Technological Area" block.*

Evaluation Block 2 → *Company Size* (elective)

Using the information provided by the column "SME Flag" in the *Horizon2020.xlsx* database, we can determine whether a participant is an SME or not. To be categorized in size, the participant entities in H2020 will be divided into two big groups: SMEs and LEs.

- **Small and Mid-sized Enterprises** (*SMEs*) are those entities in European territory with no more than 250 employees. [40]
- **Large Enterprises** (*LEs*) are those entities in European territory with more than 250 employees. [40]

Parameter	Definition	Computation	Database
Small and Mid-Size Enterprises *(SMEs)*	Number of times small and mid-sized enterprises presented H2020 projects in a region.	Count of entries whose value in the column "SME Flag" is "Yes" among the entries with the same value in "NUTS 2 Code".	*Horizon2020.xlsx*
Large Enterprises *(LEs)*	Number of times H2020 projects were presented by non-SMEs in a region (this mainly corresponds to large enterprises).	Count of entries whose value in the column "SME Flag" is "No" among the entries with the same value in "NUTS 2 Code".	*Horizon2020.xlsx*

***Table 5**. Computation of the different parameters in the "Company Size" block.*

Evaluation Block 3 → *Technological Maturity* (elective)

From the column "Legal Entity Type" in the database *Horizon2020.xlsx* we can obtain information on the entities' work purposes. This is useful to see which level of technological maturity each region is mostly in (focused on researching, developing, or integrating technology). The different maturity levels that an entity can be in are the following:

- ***Research***. This refers to organizations focused on developing new technological offerings based on engineering innovation or scientific discoveries and advances [41]. It will be assumed that the entities that develop these exploration and innovation-based projects are universities or other higher-level research institutions.
- ***Development***. This technological maturity level corresponds to organizations focused on R&D, to build and enhance existing technologies. The organizations that do these development jobs are usually technological centers or research organizations.
- ***Integration***. This corresponds to entities dedicated to integrating existing technologies with products or services in order to expand business capabilities. Therefore, this type of job is performed by private profit entities (companies).

Given that there are five different possible values in the column describing the entity

type of the H2020 participants, these will be categorized to fit into one of the three maturity levels explained above. Entries categorized as "Other" or "Public body (excl. research and education)" under the column "Legal Entity Type Descr" will be excluded, since they do not directly fit into any of those maturity levels.

Parameter	Definition	Computation	Database
Research	Number of times deep-tech entities presented projects to H2020 in each region.	Count of entries whose value in the column "Legal Entity Type Descr" is "Higher or secondary education" among the entries with the same value in "NUTS 2 Code"	*Horizon2020.xlsx*
Development	Number of times development entities presented projects to H2020 in each region.	Count of entries whose value in the column "Legal Entity Type Descr" is "Research organizations" among the entries with the same value in "NUTS 2 Code"	*Horizon2020.xlsx*
Integration	Number of times integration entities presented projects to H2020 in each region.	Count of entries whose value in the column "Legal Entity Type Descr" is "Private for profit (excl. education)" among the entries with the same value in "NUTS 2 Code"	*Horizon2020.xlsx*

***Table 6**. Computation of the different parameters in the "Technological Maturity" block.*

Evaluation Block 4 → *Capital* (non-elective)

To determine the capital availability in each European region, three databases providing different information each will be taken into consideration. From each database a parameter will be obtained.

- The ***EU grants*** are the euro amount that the European Union has given to technological entities through the Horizon2020 program. This parameter indicates the easiness of regions to obtain funds from public organisms as the EU. A region that receives a higher amount of grants has more money to invest in technology.
- ***Capital*** corresponds to a national score summarizing several variables in capital investment, calculated as the average of the following sub-parameters:
    - *FDI & technology transfer*. Foreign Direct Investments (FDI) are the purchases of an interest in a company by an investor (generally another company) located outside its national borders [42]. Therefore, the higher the score for a country in this aspect, the more foreign investment it has obtained so far.
    - *R&D expenditure (% of GDP)*. This value corresponds to the proportion of a nation's gross domestic product (GDP) that is allocated in performing research and development (R&D). This last term refers to the work directed toward the innovation, introduction, and improvement of products and processes (in any field, not only ICT) [43].
    - *VC availability.* Venture capital (VC) is a form of private equity that funds start-ups and early-stage emerging companies with little to no operating history but significant potential for growth [44]. This means that SMEs in high-scoring countries have greater opportunities to find capital investment.
- ***Unicorns*** are start-up companies valued at more than a billion dollars, typically in the technology sector [31]. A region that has unicorns is an indicator of favorable investment conditions in that location, since unicorns are entities that started small and used vast amounts of capital to grow exponentially.

Parameter	**Definition**	**Computation**	**Database**
EU Grants	Total amount of funds that the European Commission has granted each region through the H2020 program (in millions of	Sum of values in the column "H2020 Net EU Contribution" while grouping the dataset by "NUTS 2 Code"	*Horizon2020.xlsx*

	euros).		
Capital	Score (out of 100) indicating the easiness to access capital in each country.	Average from the values in the columns "FDI & technology transfer", "R&D expenditure (% of GDP)" and "VC availability" for each entry.	*Subfactor score table.csv*
Unicorns	Number of unicorn start-ups in each region.	Count of entries for each value after merging the databases unicorns.xlsx and Horizon2020.xlsx by "NUTS 2 Code" and grouping them by "NUTS 2 Code"	*unicorns.xlsx*

***Table** 7. Computation of the different parameters in the "Capital" block*

Evaluation Block 5 → *Human Resources* (non-elective)

The human resources block will be defined by a single parameter, which will be computed as the mean of five sub-parameters, all obtained from the *Subfactor score table.csv* database, as indicated below:

- *Researchers in R&D* corresponds to the number of people engaged in Research & Development (R&D), expressed as per million, in a country. Researchers are professionals who conduct studies and build or enhance concepts, theories, instrumentation, or any type of technology. R&D covers basic research, applied research, and experimental development. [29]
- *University-industry collaboration* refers to the extent to which universities collaborate with companies in R&D projects in each country. [29]
- *Skillset of university graduates* corresponds to the extent to which recent graduates possess the skills needed by businesses in each country. [29]
- *Extent of staff training* refers to the extent to which companies invest in training and

employee development. [29]

- *Availability of scientists & engineers* corresponds to the extent to which scientists and engineers are available in a country. [29]

Parameter	Definition	Computation	Database
Human Resources (*HHRR*)	Score (out of 100) indicating the availability and quality of human resources to engage in technological projects in each country.	Average from the values in the columns "Researchers in R&D", "University-industry collaboration", "Skillset of university graduates", "Extent of staff training" and "Availability of scientists & engineers" for each entry	*Subfactor score table.csv*

***Table 8**. Computation of the different parameters in the "Human Resources" block*

Evaluation Block 6 → *Innovative Ecosystem* (non-elective)

This block is intended to measure the capacity to innovate and create a productive ecosystem of each region. It will be measured by the following factors:

- *Tech Hubs* measures the number of tech hubs in each European region. A high number of tech hubs in a region is synonymous with it being a productive and proliferating location, given the abundance of opportunities and advantages for tech entities.

Parameter	Definition	Computation	Database
Tech Hubs	Number of tech hubs in a region.	Count of values in any column while grouping the dataset by "NUTS2 Code"	*digital-innovation-hubs.numbers*

***Table 9**. Computation of the different parameters in the "Innovative Ecosystem" block*

Evaluation Block 7 → *Legal Framework* (non-elective)

This block measures the legal easiness to open and operate an ICT business in each region. It will be defined by a single parameter, "Government", based on the following sub-parameters:

- *Govt ensuring policy stability* refers to the extent to which the government in each country ensures a stable policy environment for doing business. [29]
- *Legal fwk's adaptability to digital BMs* measures how fast the legal framework of each country is adapting to digital business models (e.g. e-commerce, sharing economy, fintech, etc.). [29]
- *Efficiency in settling disputes* measures how efficient the legal and judicial systems for companies in settling disputes are in each country. [29]
- *Efficiency in challenging regulations* measures how easy it is for private businesses to challenge government actions and/or regulations through the legal system. [29]
- *Burden of govt regulation* is a score measuring how difficult it is for companies to comply with public administration's requirements (e.g. permits, regulations, reporting). [29]
- *# of days to start a business* refers to the time (in calendar days) required to complete the procedures to legally operate a business. [29]
- *e-Gov services* is a score reflecting the e-participation index (EPI) of each country. [29]. This index measures the use of online services to facilitate the provision of information by governments to citizens ("e-information sharing"), interaction with stakeholders ("e-consultation"), and engagement in decision-making processes ("e-decision making"). [46]

Parameter	**Definition**	**Computation**	**Database**
Government	Score (out of 100) indicating the legal easiness to open and operate an ICT business in each region.	Average from the values in the columns "Govt ensuring policy stability", "Legal fwk's adaptability to digital BMs", "Efficiency in settling disputes", "Efficiency in	*Subfactor score table.csv*

		challenging regulations", "Burden of govt regulation", "# of days to start a business" and "e-Gov services" for each data entry.	

***Table 10**. Computation of the different parameters in the "Legal Framework" block*

Evaluation Block 8 → *Technological Infrastructure* (non-elective)

This block measures the availability and quality of the technological infrastructure in each region. It will be defined by a single parameter, "Infrastructure", based on the following sub-parameters:

- *4G coverage* measures how consistently accessible 4G networks are in each country. This score is based on the proportion of time users have access to a particular network. [29]
- *Fiber coverage* reflects the number of homes passed by FTTP (fiber-to-the-premise) at a national level. [29] FTTP refers to the installation of optical fiber directly to individual buildings and businesses to provide high-speed broadband access. [29]
- *Internet BW per user* is a metric based on the sum of the capacity of all Internet exchanges offering international bandwidth measured in kilobits per second (kb/s). [29] It is meant by "international bandwidth", the maximum quantity of data transmission from a country to the rest of the world. [47]
- *5G commercial networks* measures the level of a country's commercial 5G network development. [29]
- *# of IXPs* is a score measuring the number of internet exchange points (IXPs) in each country. [29] An IXP is a physical location through which Internet infrastructure companies, such as Internet Service Providers (ISPs) and Content Delivery Networks (CDNs), connect with each other. [48]
- *# & maturity of 5G pilots* is a score reflecting the sum of the 'maturities' of all the 5G pilots in each country. [29] 5G pilots are projects designed to test the efficiency of the implementation of the new 5th generation of mobile communication technology. [49] *Time to get electricity* is a measure that captures the number of days to obtain a

permanent electricity connection. It is computed as the median duration that the electricity utility and experts indicate is necessary in practice, rather than required by law, to complete a procedure. [29]

- *4G launch year* is a metric reflecting the date when the first commercial 4G network went live in the country. [29] The earlier this network was in operation in a country, the higher this score is.
- *5G spectrum auction plans* measures the country's readiness in allocating the three main frequency bands that will be used for 5G (as these have been identified by the International Telecommunication Union): 700MHz, 3.6GHz and 26GHz. [29]

Parameter	Definition	Computation	Database
Infrastructure	Score (out of 100) indicating the availability and quality of the technological infrastructure in each region.	Average from the values in the columns "4G coverage", "Fiber coverage", "Internet BW per user", "5G commercial networks", "# of IXPs", "# & maturity of 5G pilots", "Time to get electricity", "4G launch year" and "5G spectrum auction plans" for each entry.	*Subfactor score table.csv*

Table 11. *Computation of the different parameters in the "Technological Infrastructure" block*

6.3.4. Data Integration

Each of the created dictionaries will be appended to a data frame. This data frame will consequently contain the information for all 280 regions of all the 23 parameters computed, thus having a measure of 280 rows and 23 columns.

6.3.5. Data Normalization

The resulting values in the data frame are in absolute terms (e.g. the number of

unicorns per region) —rather than in relative terms (e.g. the presence of unicorns in a region, in proportion to the rest of the regions) —. Therefore, the data needs to be normalized so that all the parameters are measured on the same scale (values from zero to one).

There are two popular ways of performing data normalization: through the *Z-score* (or *standard score*) method, and through the *Min-Max* method.

- The *Z-score* method normalizes the distance between each score and the mean score to its standard deviation.

$$\frac{X-\mu}{\sigma} = \frac{X - mean}{standard\ deviation}$$

- The *Min-Max* method transforms a set of scores such that all of them fall in the domain [0, 1]. This is done by assigning a value of 1 to the highest score and 0 to the lowest one, and falling the rest of scores between these two limits proportionally.

$$X\ new = \frac{X - min(x)}{max(x) - min(x)}$$

In the research paper "*Comparative Analysis of KNN Algorithm using Various Normalization Techniques"* [50] both techniques were compared, concluding that the average accuracy of the Min-Max normalization algorithm in a K-means problem is greater than the one returned by the Z-score normalization algorithm. For this reason, the former method was selected to normalize the built dataframe.

The resulting normalized data frame will therefore form the array of evaluation vectors that will be compared in similarity with the target vector. This data frame will be extracted in XLSX format for the purpose of future algorithm deployment.

Finally, the resulting evaluation data frame was saved to XLSX format with the name "*RegionalVectors.xlsx*", in order to be used in deployment. Hereunder a preview of the resulting database is shown:

NUTS 2 Code	AI	Big Data	Computing	Cybersecurity	Internet	IoT	Media & Communication	Robotics	Software	SMEs
AL02	-0,43703	-0,41737	-0,383788786	-0,406285372	-0,38578	-0,5024	-0,401886546	-0,51014	-0,46624	-0,61048
AT11	-0,43703	-0,20141	-0,383788786	-0,406285372	-0,38578	-0,5024	-0,394314413	-0,51014	-0,46624	-0,5143
AT12	-0,43703	-0,41737	-0,383788786	-0,322138468	-0,38578	-0,2703	-0,390342147	-0,16691	-0,46624	-0,2578
AT13	0,583968	1,570928	0,701357027	3,09661975	-0,009	1,531967	0,615225943	0,687647	1,508856	1,184985
AT21	-0,43703	-0,13512	-0,090830314	0,336389749	-0,09999	0,339682	-0,3284741	-0,2772	-0,22534	0,223127
AT22	0,362639	0,76263	-0,216748395	0,791994163	-0,20646	0,654498	0,158359409	-0,36176	0,549518	1,249108
AT31	0,811729	0,067519	-0,199450853	-0,037787503	-0,29723	-0,38318	0,107664615	0,519813	0,112308	-0,03337
AT32	-0,43703	-0,04579	-0,383788786	-0,201382197	-0,38578	-0,5024	-0,350005024	-0,20694	-0,46624	-0,45017
AT33	-0,43703	-0,2793	-0,009973882	-0,406285372	-0,38578	-0,06846	-0,401886546	-0,51014	-0,46624	-0,57842
AT34	-0,43703	-0,41737	-0,383788786	-0,406285372	-0,38578	-0,5024	-0,401886546	-0,51014	-0,46624	-0,61048
BE10	1,705654	1,108713	-0,216166554	0,377409712	1,63846	1,512494	2,121106729	0,850397	0,840333	2,307151
BE21	-0,22852	-0,21835	-0,383788786	-0,406285372	-0,26817	-0,11831	0,525701662	-0,11734	-0,17624	0,159003
BE22	-0,43703	-0,41737	-0,383788786	-0,406285372	-0,38578	-0,5024	-0,401886546	-0,51014	-0,46624	-0,5143
BE23	-0,43703	-0,28806	0,440606519	-0,406285372	-0,27881	0,573293	-0,123602367	-0,37474	-0,46624	0,062818
BE24	0,172805	1,117146	-0,139986508	1,796162025	4,590964	0,846322	1,257724709	0,684457	-0,33816	0,543746
BE25	-0,43703	-0,41737	-0,383788786	-0,122881879	-0,30075	-0,36657	-0,341681885	-0,474	-0,39527	-0,45017
BE31	-0,43703	-0,32145	-0,151103332	0,875315513	1,823412	-0,5024	-0,387446365	-0,40524	-0,46624	-0,38605

Figure 9. *Final Evaluation Database.*

6.4. Modeling

The location recommendations provided by the system will be based on the distances between the evaluation vectors (the built data frame) and the target vector (to be created based on the input vectors). This means that the regions whose evaluation vector maintains a shorter distance with the target vector will be first in the recommendation results.

6.4.1. Modeling Technique

To determine the degree of similarity between the mentioned vectors there are seventeen different similarity-measurement techniques that can be employed [51]. Measuring the distance between them was chosen to be the matchmaking method due to its properties, which suit the characteristics of the available vectors [52]:

1. $d(x, y) \geq 0$ (non-negativity or separation)
2. $d(x, y) = 0$ if and only if $x = y$ (coincidence axiom)
3. $d(x, y) = d(y, x)$ (symmetry)
4. $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

Similarly, there are different methods to compute the distance between two vectors of the same length (same number of dimensions). The article *"A Comparative Analysis of Similarity Measures to find Coherent Documents*" [53] reviews the most important ones and concludes that both *Jaccard similarity* and *cosine similarity* give good results in both numerical and binary data. A similar conclusion is reached in the paper "*Comparison Jaccard*

similarity, Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method" [54] where it is shown that results of cosine similarity have the highest value in comparison with Jaccard similarity and the joint between Cosine and Jaccard similarity in KNN applications. In fact, cosine similarity has been used in a large number of applications, especially with high-dimensional data. Cosine similarity has also been used in the country comparison, which is somewhat similar to the prospective recommendation application that is being built on this project, for example in the paper "*A country comparison of place-based activity response to COVID-19 policies*". [55]

Therefore, the *cosine distance* was finally selected. The formula used to compute the cosine distance between the regional vectors ($RV = [rv_1, rv_2, ..., rv_{280}]$) and the target vector (tv) will therefore be:

$$cosine\ distance\ (rv_i, tv) = 1 - cos(rv_i, tv) = 1 - \frac{rv_i \cdot tv}{||rv_i|| ||tv||}$$

6.4.2. Matchmaking Algorithm

To create the matchmaking algorithm, the next steps will be followed:

Step 1. Creation of the final target vector.

This will be done by computing the scalar product between the adjusted input vector reflecting the search selection and the adjusted input vector reflecting the importance of each evaluation block (both will have the same number of dimensions).

Step 1.1. Creation of the adjusted input vector reflecting the search selection. The user will be asked to enter its search criteria over the three elective evaluation blocks (technological area, company size, and technological maturity). Therefore, the user will give values of either 0 (representing “no”) or 1 (representing “yes”) to the next 15 parameters:

1) From the *Technological Area* block, responding to the question “Are you interested in locating your business in regions prevalent in the following technologies?”

- $d1$: AI → 0 or 1
- $d2$: Big Data → 0 or 1
- $d3$: Computing → 0 or 1

- $d4$: Cybersecurity → 0 or 1
- $d5$: Internet → 0 or 1
- $d6$: IoT → 0 or 1
- $d7$: Media & Communication → 0 or 1
- $d8$: Other → 0 or 1
- $d9$: Robotics → 0 or 1
- $d10$: Software → 0 or 1

2) From the *Company Size* block, responding to the question "Are you interested in locating your business in regions populated with small and medium-sized enterprises (SMEs) or with large enterprises (LEs)?"

- $d11$: SMEs → 0 or 1
- $d12$: LEs → 0 or 1

3) From the *Technological Maturity* block, responding to the question "Are you interested in locating your business in regions prevalent in institutions working with the following purposes?"

- $d13$: Deep Tech → 0 or 1
- $d14$: Development → 0 or 1
- $d15$: Integration → 0 or 1

The resulting input vector will therefore have 15 dimensions:

$$Input\ Vector\ (search) = [d1, d2, d3, d4, d5, d6, d7, d8, d9, d10, d11, d12, d13, d14, d15]$$

However, it needs to be adjusted to match the same number of dimensions as the regional evaluation vectors (23). This will be done by extending the existing input vector with eight dimensions of values equating to 1.

$$Adjusted\ Input\ Vector\ (search) =$$

$$[d1, d2, d3, d4, d5, d6, d7, d8, d9, d10, d11, d12, d13, d14, d15, 1, 1, 1, 1, 1, 1, 1, 1]$$

Step 1.2. Creation of the adjusted input vector reflecting the importance of each evaluation block. The user will be asked to enter the weight that he or she wants to assign to each of the evaluation blocks (technological area, company size, technological maturity, capital, human

resources, innovative ecosystem, legal framework, and technological infrastructure). Therefore, the user will give eight values between 0 and 1, which should all together sum an exact total of 1:

$$Input\ Vector\ (weights) = [w1, w2, w3, w4, w5, w6, w7, w8]$$

$$subject\ to\ 1 = w1 + w2 + w3 + w4 + w5 + w6 + w7 + w8$$

However, it needs to be adjusted to match the same number of dimensions as the adjusted input vector reflecting the search selection (23). This will be done by assigning each weight to the positions in the vector corresponding to the parameters of each block, distributed to the parameters the user chose to have a value of "1" in *Step 1.1*. Consequently, the number of "Yes" choices will be taken into account in the following way:

- Number of technological areas (AI, big data, computing, cybersecurity, Internet, IoT, media & communication, robotics, software, other) selected as "Yes" → $n1\ s.t.\ n1 \in [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$
- Number of company sizes (SMEs, LEs) selected as "Yes" → $n2\ s.t.\ n2 \in [1, 2]$
- Number of technological maturities (deep-tech, development, integration) selected as "Yes" → $n3\ s.t.\ n2 \in [1, 2, 3]$

$$Adjusted\ Input\ Vector\ (weights) =$$

$$\left[\frac{w1}{n1}, \frac{w1}{n1}, \frac{w1}{n1}, \frac{w1}{n1}, \frac{w1}{n1}, \frac{w1}{n1}, \frac{w1}{n1}, \frac{w1}{n1}, \frac{w1}{n1}, \frac{w1}{n1}, \frac{w2}{n2}, \frac{w2}{n2}, \frac{w3}{n3}, \frac{w3}{n3}, \frac{w3}{n3}, w4, w4, w4, w5, w6, w6, w7, w8\right]$$

Step 1.3. Composition of the final target vector. The final target vector will result after performing the scalar product of the adjusted input vectors:

$$Target\ Vector\ (tv) = Adjusted\ Input\ Vector\ (search) \times Adjusted\ Input\ Vector\ (weights) =$$

$$[d1, d2, d3, d4, d5, d6, d7, d8, d9, d10, d11, d12, d13, d14, d15, 1, 1, 1, 1, 1, 1, 1, 1]$$

$$\times$$

$$\left[\frac{w1}{n1}, \frac{w1}{n1}, \frac{w1}{n1}, \frac{w1}{n1}, \frac{w1}{n1}, \frac{w1}{n1}, \frac{w1}{n1}, \frac{w1}{n1}, \frac{w1}{n1}, \frac{w1}{n1}, \frac{w2}{n2}, \frac{w2}{n2}, \frac{w3}{n3}, \frac{w3}{n3}, \frac{w3}{n3}, w4, w4, w4, w5, w6, w6, w7, w8\right]$$

$$=$$

$$[d1\frac{w1}{n1}, d2\frac{w1}{n1}, d3\frac{w1}{n1}, d4\frac{w1}{n1}, d5\frac{w1}{n1}, d6\frac{w1}{n1}, d7\frac{w1}{n1}, d8\frac{w1}{n1}, d9\frac{w1}{n1}, d10\frac{w1}{n1}, d11\frac{w2}{n2}, d12\frac{w2}{n2},$$
$$\mathrm{d13}\frac{w3}{n3}, d14\frac{w3}{n3}, d15\frac{w3}{n3}, w4, w4, w4, w5, w6, w6, w7, w8]$$

Step 2. Computing the distances between the target vector and the regional vectors.

Using the cosine distance method, the similarity between the target vector (tv) and each regional vector (from the matrix of evaluation vectors *RV*) will be examined as:

$$Similarity\ Score\ (tv, rv_i) = 1 - cos(rv_i, tv)$$

$$\forall\ rv_i \in RV$$

This will lead to 280 scores (one for each region) between 0 and 1, being 1 the highest score possible.

Step 3. Delivering the obtained scores.

Scores will be sorted in descending order so that the regions with the highest scores are displayed first in the recommendation results. Finally, a data frame showing the resulting scores, along with the region NUTS2 code, region name, and country name will be returned as output for the user.

6.5. Evaluation

This section of the paper is dedicated to evaluating whether the location recommendations provided by the built algorithm are accurate and whether the system can be effectively used. Therefore, the algorithm will be doubly tested (through both a recommendation accuracy and a system usability assessment).

6.5.1. Recommendation Accuracy Assessment

To evaluate the accuracy of the results obtained, the recommendation scores need to be compared against other dataset indicating the presence of businesses in the ICT industry in the European regions.

Although indirectly, this dataset can be acquired. To do this, the total number of companies in each European region will be obtained from Eurostat [56], keeping the average of the last

three years. On the other hand, the activity of each technological area in the regions (this is, the percentage of companies in a region that use a specific technology) can also be obtained from different sources:

- AI, IoT and Computing → Activity of each of these technological areas in each region. [57]
- Cybersecurity → Proportion of enterprises in each region using any ICT security measure. Obtained from the *Security policy: measures, risks and staff awareness* database (Eurostat, 2019). [58]
- Big Data → Proportion of enterprises in each region using big data analytics internally from any data source. Obtained from the *Big data analysis* database (Eurostat, 2020). [58]
- Robotics → Proportion of enterprises in each region using industrial or service robotics. Obtained from the *3D printing and robotics* database (Eurostat, 2020). [58]

Having the proportion of activity of each technology in a region, along with the total number of companies in each region, the number of companies in a region that use a certain technology can be calculated. The formula is:

$$\#\ companies\ in\ a\ tech\ area\ in\ a\ region\ =$$

$$\#\ companies\ in\ a\ region\ \times\ \%\ tech\ activity\ in\ a\ region$$

Therefore, after applying normalization to the resulting data on the number of companies using each technology in each region, an objective dataset to evaluate the results of the recommendation system can be obtained. The assessment will be conducted by computing the correlation between the number of companies using each technology in a region and the score obtained by that region using the recommendation system (configured in a way that only the technological area and the non-elective parameters are weighted in the target vector, having the tech area the largest weight (50%) and only being selected that specific technology as a value of "Yes" (1) in the input vector reflecting the search preferences). As an illustration, the search configuration to execute the recommendation of regions prevalent in AI will thus be the following:

$$Input\ Vector\ (search)\ =\ [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$$

$$Input\ Vector\ (weights)\ =\ [0.5,\ 0,\ 0,\ 0.1,\ 0.1,\ 0.1,\ 0.1,\ 0.1]$$

By executing the algorithm using that search configuration ($weights$ remain the same whereas in $search$ only the position of the "1" changes) for the six technological areas that we have data on the number of companies ("AI", "IoT", "Computing", "Cybersecurity", "Big Data" and "Robotics"), the correlation coefficients got are:

- AI → 0.52
- Big Data → 0.51
- Computing → 0.59
- Cybersecurity → 0.66
- IoT → 0.67
- Robotics → 0.69

All the pairs got significant positive correspondence (over 50% in all of them, some being close to 70%), which implies that the recommendation system delivers scores which are to a great extent realistic and accurate.

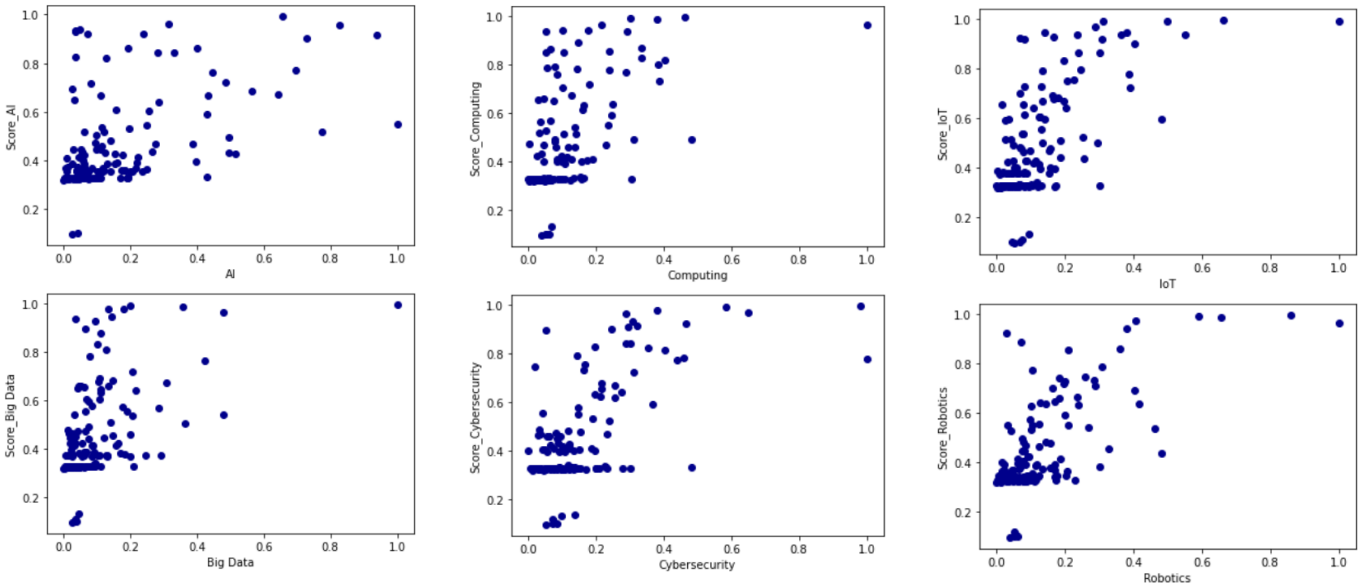


Figure 10. *Correlations between the scores obtained for the regions by the built system and the activity in each tech area for the regions.*

6.5.2. System Usability Assessment

Hereunder, a set of case scenarios will be proposed, so as to demonstrate the system's modus operandi and observe the results obtained under typical use conditions.

Case scenario 1 → Any technological area, any company size, any technological maturity, only giving importance to capital, human resources, innovative ecosystem, legal framework, and technological infrastructure (five equal weights of 0.2).

This combination of inputs returns the European regions that are the most complete in all technological and business factors. The recommendations obtained represent the locations where a random organization (with any business requirements) would most likely be successful in the ICT field.

```
#Test Scenario 1
selection1 = [1,1,0,0,1,0,0,0,0,1,0,1,0,0]
weights1 = [0,0,0,0.2,0.2,0.2,0.2,0.2]
test1 = recommendation(selection1, weights1)
test1
```

	Score	Region	Region Name	Country Name
0	98.262026	FR10	Ile-de-France	France
1	97.226010	EL30	Αττική	Greece
2	97.094505	NL41	Noord-Brabant	Netherlands
3	96.656906	DED2	Dresden	Germany
4	96.003474	DE21	Oberbayern	Germany

Case scenario 2 → AI as technological area, any company size, Deep Tech as technological maturity, only giving importance to the technological area and technological maturity (two equal weights of 0.5).

This combination of inputs returns the European regions that are most specialized in the investigation of artificial intelligence (deep-tech). This scenario is interesting for individuals or organizations that seek to identify tech hubs in AI.

```
#Test Scenario 2
selection2 = [1,0,0,0,0,0,0,0,0,0,1,0,0]
weights2 = [0.5,0,0.5,0,0,0,0,0]
score2 = recommendation(selection2, weights2)
score2
```

	Score	Region	Region Name	Country Name
0	99.998723	ITH3	Veneto	Italy
1	99.998309	ES52	Comunitat Valenciana	Spain
2	99.994405	UKJ1	Berkshire, Buckinghamshire and Oxfordshire	United Kingdom
3	99.898701	ITI2	Umbria	Italy
4	99.898701	SI03	Vzhodna Slovenija	Slovenia

Case scenario 3 → Any technological area, any company size, any technological maturity, only giving importance to the capital and legal framework (equal weights of 0.5)

This combination of inputs returns the European regions where it is the easiest to operate an ICT organization, exclusively from a business perspective. In this recommendation, the parameters examined by region are economic (net amount of grants received by the EU, availability of venture capitals, cross-border investment (FDI), R&D expenditure, and the number of unicorn start-ups) and legal (ability of the national government to ensure policy stability, government's efficiency in settling disputes and challenging regulations, period of days needed to start a business, e-Gov services and level of adaptability of the legal framework to adapt to digital business models).

```
#Test Scenario 3
selection3 = [1,0,0,0,0,0,0,0,0,1,1,0,0]
weights3 = [0,0,0,0.5,0,0,0.5,0]
score3 = recommendation(selection3, weights3)
score3
```

	Score	Region	Region Name	Country Name
0	98.068909	FRK2	Rhône-Alpes	France
1	97.757899	FR10	Ile-de-France	France
2	97.719971	DED2	Dresden	Germany
3	96.581443	UKI3	Inner London — West	United Kingdom
4	96.103613	UKJ1	Berkshire, Buckinghamshire and Oxfordshire	United Kingdom

Case scenario 4 → Any technological area, SME as company size, any technological maturity, only giving importance to company size and innovative ecosystem (equal weights of 0.5).

This combination of inputs returns the European regions where there is the maximum concentration of ICT SMEs, forming an innovative ecosystem. The factors considered in this recommendation are the amount of SMEs (start-ups), the degree of project collaboration between companies (networking), and the presence of tech hubs in the regions.

```
#Test Scenario 4
selection4 = [1,0,0,0,0,0,0,0,0,0,1,1,0,0]
weights4 = [0,0.5,0,0,0,0.5,0,0]
score4 = recommendation(selection4, weights4)
score4
```

	Score	Region	Region Name	Country Name
0	99.950963	NL22	Gelderland	Netherlands
1	99.903302	BE10	Région de Bruxelles-Capitale/ Brussels Hoofdst...	Belgium
2	99.801297	DEG0	Thüringen	Germany
3	99.793566	FI19	Länsi-Suomi	Finland
4	99.706917	DK05	Nordjylland	Denmark

6.6. Deployment

The created tool is to be primarily used by business people, who frequently lack the technical and programming abilities to access, execute and configure the algorithm using programming tools. For this reason, a simple and user-friendly version of the system will be presented, using common-use digital channels.

6.6.1. Deployment Plan

The idea is to deploy the algorithm into a public web application that any individual can instantly access through the Internet using a link. A web application is a software that runs on a web browser (unlike software programs that run on the local operating system of computers) and is delivered on the World Wide Web (WWW) to users with an active network connection [59]. Web applications are made up of two primary components:

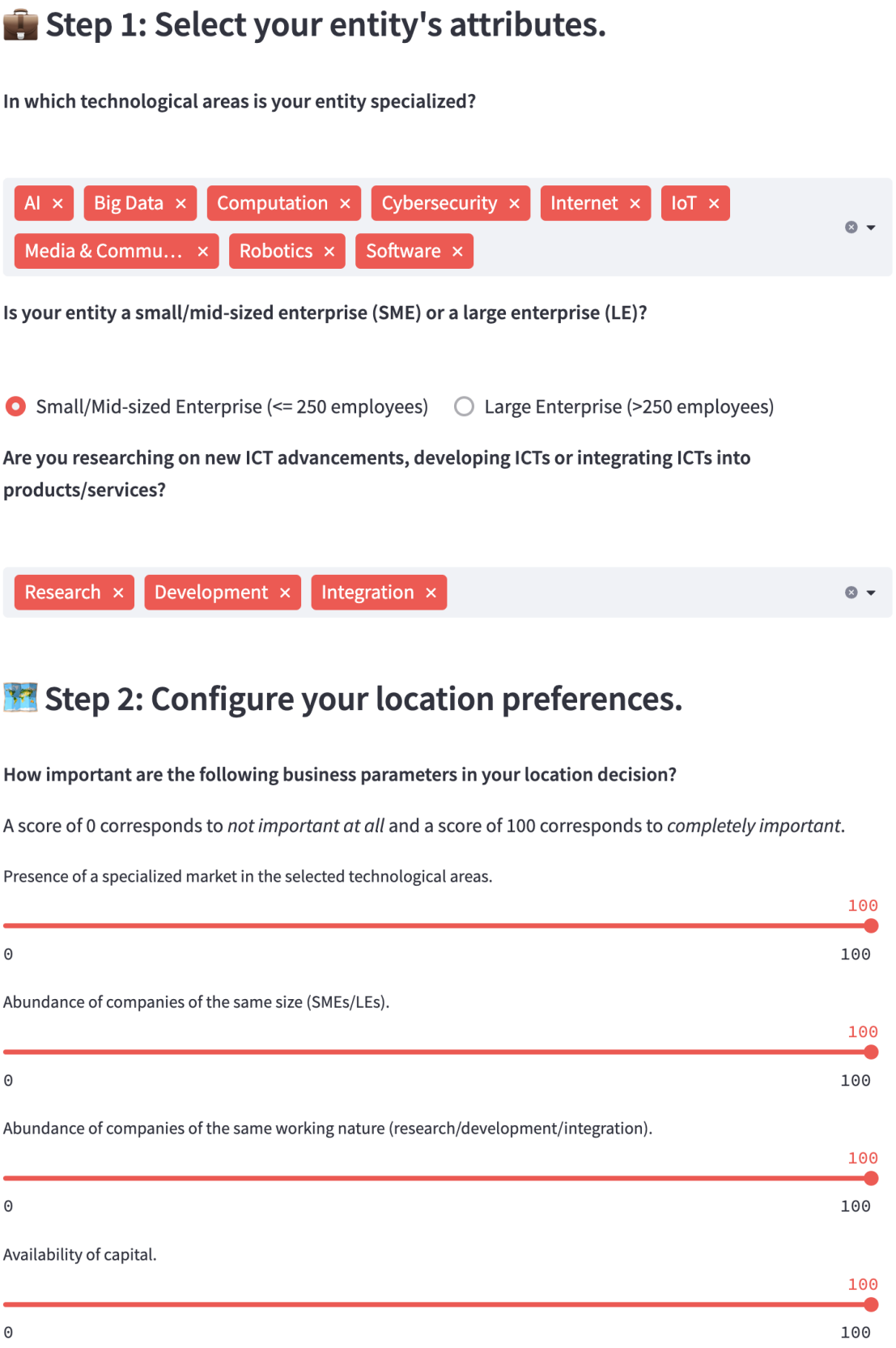
- **Server-side (back-end component)**. It is a computing device (computer) that controls the business logic, handles databases, and responds to HTTP requests. The server-side code should be written in a Python notebook (like the one used to create the algorithm on this project). This notebook interacts with a SQL (Structured-Query Language) database. The notebook and database together then form a project that runs on the server computer to deliver certain outputs, that will later be transmitted to the client side.
- **Client-side (front-end component).** It is where the interaction with the user takes place, through web pages written in HTML (Hyper-Text Markup Language) or other languages.

The traditional approach to deploying a web application is to wrap the API (Application Programming Interface), via the use of web frameworks such as *Django* and *Flask*. However, a much simpler approach is to use a low-code solution such as *Deepnote* or *Streamlit* to create a web application.

In this case, the objective of the deployment is just to collect input data from the user and display the recommendation results obtained. Therefore a simpler version, which does not require complex programming, is enough. With this intention, the latter program will be used: *Streamlit Cloud*. Streamlit is an open-source app that enables the quick deployment, management, and sharing of web applications. [60]

6.6.2. Deployment Visualization

To deploy the recommendation system in a web application through Streamlit, the code and files in the following GitHub repository were used: https://github.com/carmenpelayo/Carmen-Pelayo/tree/main/Location-Recommender [61]. Streamlit integrates with the mentioned repository hosting service, GitHub, so it is not necessary to perform additional programming of the algorithm. Only a few lines of code are needed to customize the web application with images, text, and buttons. The final layout of the web application (with the address https://locationrecommender.streamlit.app [62]) can be observed in the following screenshots:


💼 Step 1: Select your entity's attributes.
In which technological areas is your entity specialized?
AI × Big Data × Computation × Cybersecurity × Internet × IoT × Media & Commu… × Robotics × Software ×
Is your entity a small/mid-sized enterprise (SME) or a large enterprise (LE)?
Small/Mid-sized Enterprise (<= 250 employees)
Large Enterprise (>250 employees)
Are you researching on new ICT advancements, developing ICTs or integrating ICTs into products/services?
Research × Development × Integration ×
🗺️ Step 2: Configure your location preferences.
How important are the following business parameters in your location decision?
A score of 0 corresponds to not important at all and a score of 100 corresponds to completely important.
Presence of a specialized market in the selected technological areas.
100
0
100
Abundance of companies of the same size (SMEs/LEs).
100
0
100
Abundance of companies of the same working nature (research/development/integration).
100
0
100
Availability of capital.
100
0
100

🏆 Your results!

Based on your entity's attributes and location preferences, these are the recommended regions in Europe:

	Score	Region	Region Name	Country Name
0	95.3960	DE30	Berlin	Germany
1	95.0089	DEA2	Köln	Germany
2	92.1339	FI1B	Helsinki-Uusimaa	Finland
3	87.8218	DE12	Karlsruhe	Germany
4	85.9568	SE11	Stockholm	Sweden
5	84.0263	DE11	Stuttgart	Germany
6	78.8398	UKI3	Inner London — West	United Kingdom
7	77.8637	DE71	Darmstadt	Germany
8	76.1001	CH04	Zürich	Switzerland
9	75.9505	NL41	Noord-Brabant	Netherlands

***Figure 11**. Preview of the Location Recommender application on Streamlit.*

7. Further Economic Applications

The database built containing the regional vectors (*RegionalVectors.xlsx*) can be further leveraged to obtain valuable economic insights.

7.1. Geographical Representation of Data

The following maps were obtained by plotting the values for all the European regions in the following parameters: "EU Grants", "HHRR", "Innovation", "Government" and "Infrastructure".

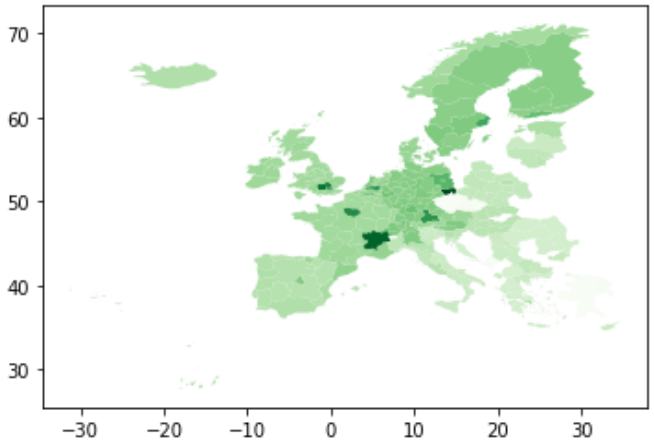
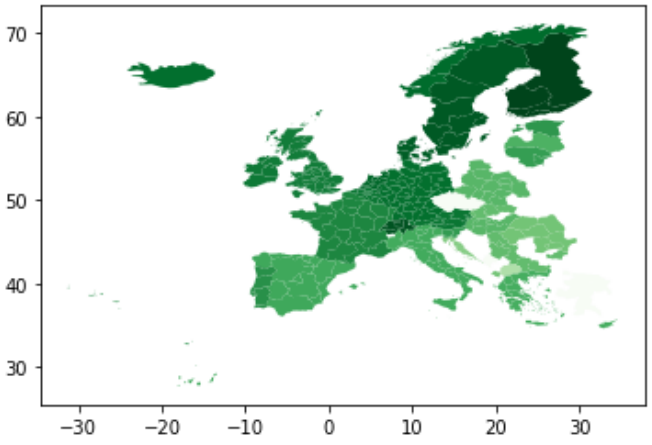


***Figure 12**. Representation in the European map of the EU Grants parameter.*

***Figure 13**. Representation in the European map of the HHRR (Human Resources) parameter.*



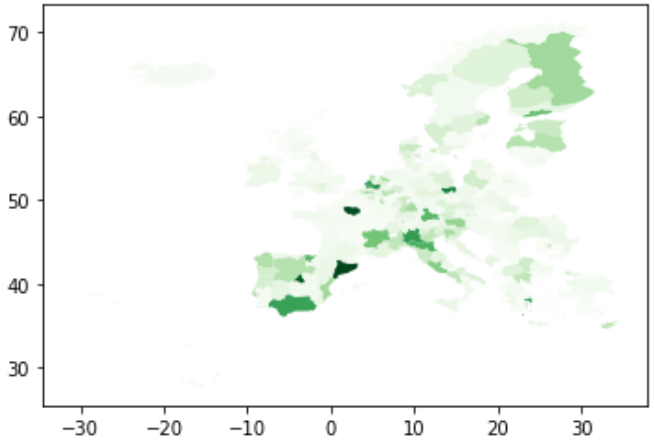
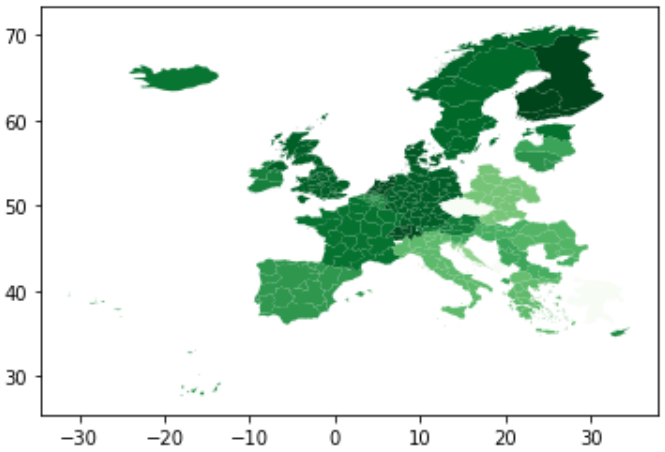


***Figure 14**. Representation in the European map of the Innovation parameter.*

***Figure 15**. Representation in the European map of the Government parameter.*



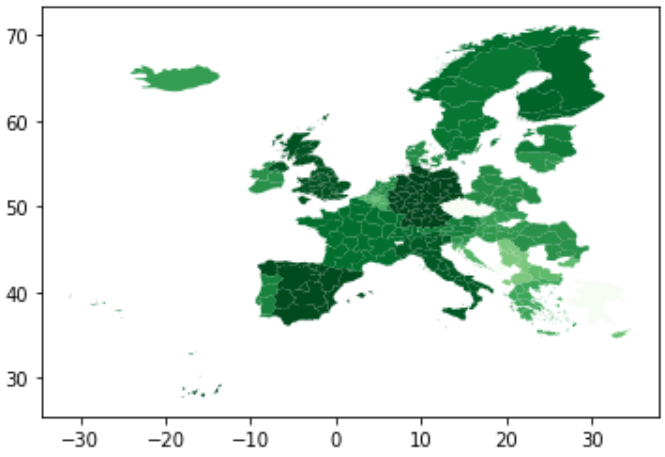


***Figure 16**. Representation in the European map of the Infrastructure parameter.*

7.2. Region Comparator

The application found at the following URL: https://europeanregioncomparator.streamlit.app/ was built using the mentioned database and can be employed to compare European regions in any of the twenty-one socio-economic parameters described in this paper. Hereunder is a preview of the results obtained by executing a comparison of the ES30 (Madrid, Spain), AT22 (Steiermark, Austria), BE23 (Prov. Oost-Vlaanderen, Belgium), and DE11 (Stuttgart, Germany) regions in the application:

	ES30	AT22	BE23	DE11
AI	3.4867	0.3626	-0.4370	-0.2497
Big Data	7.1692	0.7626	-0.2881	-0.1817
Computing	2.0934	-0.2167	0.4406	-0.0414
Cybersecurity	4.4072	0.7920	-0.4063	0.0893
Internet	3.9179	-0.2065	-0.2788	-0.2930
IoT	3.4664	0.6545	0.5733	0.9571
Media & Communication	6.6426	0.1584	-0.1236	0.4228
Robotics	3.2798	-0.3618	-0.3747	0.5958
Software	5.0638	0.5495	-0.4662	-0.4662

Figure 16. *Preview of the European Region Comparator application on Streamlit.*

8. Legal Framework

8.1. Legislation

The execution of this project was possible thanks to the availability of open-access data on multiple business and technological indicators. All the datasets employed were legally and freely obtained from reliable online sources, including public organizations and other entities, which voluntarily disclose information with the aim of providing transparency and boosting research. The Spanish Presidency Office recognizes the importance and value of the information generated by the public authorities, due to the interest it attracts from the private sector, which consequently uses it for economic growth and job creation, as indicated in the Boletín Oficial Del Estado of November 8th, 2011. Law 37/2007, of November 16th, establishes the right to openly access public information for both commercial or non-commercial purposes, following the guidelines of the 2003/98/CE Regulation, of November 17th, 2003, of the European Parliament [63]. In 2009, the "Aporta" project was initialized in Spain with the aim of promoting the reuse of public information (RISP), by communicating the power of data usage for economic value creation and reinforcement of democratic practices, increasing the participation of citizens in public politics. This initiative provides access to online public data repositories (like *datos.gob.es*), information catalogs, educational materials, and Open Data events [64].

Effective on December 10th, 2021, the "Startup Law" was approved in Spain as a means of easing and speeding up the creation of *startup* businesses, with the objective of attracting talent and capital investment to the country, and consequently directing it towards becoming an economic leader in Europe [65]. In line with the purposes of this law, the execution of the present project can be viewed as a support tool for entrepreneurs and startups.

Additionally, the European Commission launched in 2013 the so-called "Entrepreneurship 2020 Action Plan", which planned to "reignite the entrepreneurship spirit in Europe", due to the need of creating jobs, opening new markets, and nurturing new skills and capabilities [66]. Again, the present project serves the aims of this initiative, by providing an accessible and valuable tool for individuals considering starting up a business within the European frontiers.

8.2. Licenses

Apart from the multiple open data sources used in the study of the economic and business framework of the project (cited in the bibliography), there were five main datasets used in the study and development of the location recommendation tool The license pertaining to each dataset is reviewed hereunder:

- Datasets on *Horizon 2020 Projects*, *Digital Innovation Hubs* and *Smart Regions RIS3* → They were published by the European Commission's Joint Research Centre (JRC), whose website states that everybody has the right to "fully, freely, openly and timely share and use JRC data". [67]
- *Unicorns* dataset → It was published by CB Information Services, Inc., whose Terms of Service agree to the legitimate download and use of their data for legal purposes, given that it does not compromise the activity of the firm. [68]
- *GDP of NUTS3 regions* dataset → It was published by Eurostat, whose Copyright Notice and Free Reuse of Data note encourage the free re-use of its data, both for non-commercial and commercial purposes [69].

There were five software tools involved in the development of the present project. These include the text-editing software *Google Docs*, the cloud-based file storage and synchronization *Google Drive*, the computing notebook *Project Jupyter*, the version control, and source code management software *GitHub,* and the web application development software *Streamlit*.

- *Google Doc*s and *Google Drive*→ As part of the *Google Workspace* program group, Docs and Drive need a user license in order to be used [70]. The academic license provided by the Carlos III University of Madrid was used for this purpose.
- *Project Jupyter* → This project is licensed under the terms of the 3-Clause BSD License, which permits the redistribution and use in source and binary forms of the produced code, with or without modification, given that it retains the copyright notice, and that the name of the copyright holder or their contributors is not used to endorse or promote products derived from the software without specific prior written permission [71].
- *GitHub* → It is licensed under the Creative Commons Zero license, as indicated in its terms of service. [72]

- *Streamlit* → According to its terms of service, Streamlit grants a limited, non-exclusive, non-transferable, non-sublicensable license to any 18-year-old or older individual to use their services. [73]

9. Planning and Budget

9.1. Planning

Since this is a project of advanced complexity, time organization has played a key role in its success. According to the guidelines established by the Universidad Carlos III de Madrid, an ECTS (European Credit Transfer and Accumulation System) academic credit corresponds to approximately 30 hours of work by the student [74]. Therefore, consisting the development of a thesis of 6 ECTS in the degree in *Management and Technology* [75], the time investment was expected to be 180 hours for the author of the thesis (who is considered to have the role of "project developer" here). Additionally, the thesis director invested 30 hours of time in the role of "project manager". Therefore, the total time used was 210 hours: 180 in the project development term (which took place from February 7th, 2022, to June 21st, 2022) and 30 hours in the project planning term (which took place from January 31st, 2022, to May 17th, 2022). A total of 142 days were employed in the execution of this project, being classified into different phases: planning, analysis, execution, evaluation, and deployment. For a clear representation of the distribution of tasks and progress monitoring within each project phase, a Gantt Chart is shown below.

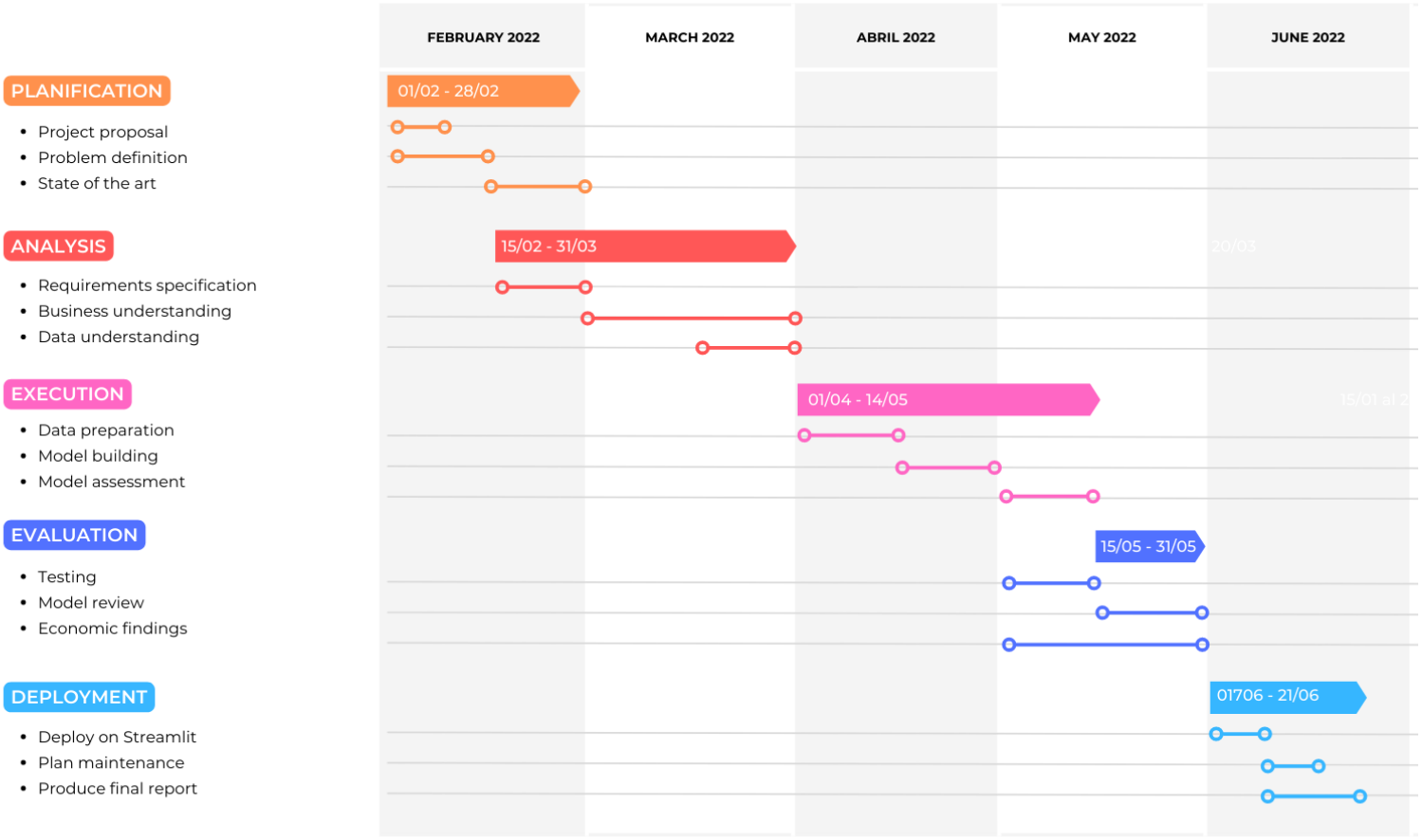


***Figure 17**. Gantt chart.*

9.2. Budget

The budget of the present project includes all direct (labor and capital expenses) and indirect costs necessary for the development of the project.

9.2.1. Labor Costs

This project was carried out by the joint work of two individuals: a project manager (in charge of setting the organizational guidelines and project requirements to be followed) and a project developer (who conducted the research and development of the project). In the calculation of labor costs, the number of hours attributable to the different tasks of the project will follow the work time proposed in *Chapter 6.1. Planning*. The salaries are established in accordance with the average salary provided by the job portal *Indeed*, which gathers data on salaries from users, published job offers, and companies. For the role of project manager, the average hourly salary obtained is 16,82 € [76], whereas the average hourly salary for a programmer analyst position is 15,96 € [77]. However, the provided values are in gross terms, thus the Social Security (SS) cost (30% of the base salary) should be added to obtain the net values. The SS cost includes the following: common contingencies, the general unemployment rate for an indefinite-term contract, Wage Guarantee Fund, and vocational training. In addition, for the hourly cost calculation. Therefore, the net hourly cost of the project manager rises to 21,86 €, while the net hourly cost of the project developer rises to 20,75 €.

Project Phase	**Project Manager**	**Project Developer**	**Total**
Planification	12	20	35
Analysis	8	50	55
Execution	5	70	75
Evaluation	3	30	33
Deployment	2	10	12
Total hours	**30**	**180**	**210**
Hourly salaries cost	21,86 €	20,75 €	

Total cost	**655,8 €**	**3735 €**	**4390,8 €**

Table 11. *Estimated labor costs.*

9.2.2. Capital Costs

This section of the budget includes all material costs, which are mainly hardware and software, since the cost of consumables is essentially negligible due to their scarcity and low value.

Hardware Costs

The project was developed using a single laptop: an *Apple MacBook Pro (15-inch, 2018)*, which is estimated to have a lifespan of 7 years [78], and was purchased for 1.699,99 €.

Product	**Laptop 1**
Model	Apple MacBook Pro (15-inch, 2018)
Net Price (with IVA)	1.699,99 €
Lifespan (in years)	7 years
Lifespan (in hours)	61.320 hours
Hourly Cost	0,0277 €
Hours Employed	180
Total Cost	4,99 €

Table 12. *Estimated hardware costs.*

Software Costs

The cost associated with the software material has been nil due to the academic environment in which this project was developed. The licenses of all the different software products used were free of charge.

9.2.3. Total Costs

The total project budget is detailed hereunder, including both direct and indirect costs, along with the safety margin, benefit margin, and corresponding sales tax applicable. Some notes to consider in the budget computation:

- The safety margin (estimated at 10%) serves as a buffer in the case of unexpected expenses.
- The benefit margin serves as an estimation of the expected profit to be obtained in the case the project was sold.
- Since 2012, sales tax in Spain (the so-called *Impuesto Sobre el Valor Añadido* or *IVA*, in short) supposes 21% of the final price value.

Labor Cost	4390,8 €
Capital Cost	4,99 €
Total Direct Costs	**4395,79 €**
Total Indirect Costs (15%)	659,37 €
Total Costs	**5055,16 €**
Safety Margin (10%)	505,52 €
Benefit Margin (20%)	1011,03 €
Tax fee (21%)	1061,58 €
Total Cost	**7633,29 €**

***Table 13**. Estimated project budget.*

Therefore, based on the estimated budget, the total cost of this project amounts to **SEVEN THOUSAND SIX HUNDRED AND THIRTY-THREE EUROS AND TWENTY-NINE CENTS** (7633,29 €), tax fee included.

10. Conclusions and Future Work

10.1. Conclusions

After completing all the phases of the present project, some conclusions were obtained based on the fulfillment of the objectives initially set (see *1.2. Objectives*):

1. The business dimensions observed to be the most relevant in the evaluation of possible locations for firms in the ICT industry are: the availability of capital, the qualification of human resources, the legal framework, the technological infrastructure and the existence of innovative ecosystems.
2. The implementation of ICTs is proven to have a direct relationship with the level of digitalization of the European regions and the productivity of the corresponding countries.
3. The most efficient way of delivering customized location recommendations was found to be the creation of a matchmaking algorithm, which compares the unique characteristics of a business with those from the different regions so as to communicate the compatibility between the different pairs and rank them accordingly.
4. The matchmaking algorithm was successfully designed, implemented, and deployed, allowing the fulfillment of the previously defined business and technical requirements.
5. The results delivered by the created system were assessed both in terms of accuracy and usability, leading to positive results.

All these conclusions confirm that the overall goal of the project of providing a decision-making support solution for entities operating in the European ICT industry was achieved.

10.2. Future Work

As the socioeconomic environment evolves over time, firms constantly need to adapt to changing conditions, resulting in new challenges that modify business characteristics and needs. A clear illustration of this was the coronavirus (COVID-19) pandemic, which demanded a drastic change in the working environment, with the proliferation of virtual collaboration and remote communication. Therefore, the dimensions considered in the

process of choosing a business location can also evolve (some dimensions can increase or decrease in importance, or new essential dimensions can emerge), thus they need to keep being monitored.

Similarly, the technological situation of a region is not static, but evolves constantly over time, which is why the recommendations provided by this system will never dictate a definitive location solution. The built tool is based on data collected from 2014 to 2020, since these are the most recent databases available on the matter. This implies the recommendations provided at the current time are not totally up-to-date. However, the created system is reproducible, meaning that the algorithm can be applied to future updated databases.

With respect to the matchmaking algorithm, other modeling techniques could be tested in the future to observe the performance. Although the *Min-Max scale* was used as the normalization technique, the *Z-score* is another option. Similarly, in the computation of the similarity between the vectors, the *cosine similarity* was used, although others like the *Jaccard similarity* also apply to the problem.

In conclusion, the developed project intends to serve as a support tool in the location decision-making process of technological firms, rather than a provider of conclusive location solutions.

11. Bibliography

[1] Rodrígez García, J. M. (n.d.). *Scientia potestas EST – knowledge is power: Francis Bacon to Michel Foucault - Neohelicon*. SpringerLink. Retrieved June 19, 2022, from https://link.springer.com/article/10.1023/A:1011901104984

[2] Ellison, G., & Glaeser, E. L. (1999). *The Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration?* The American Economic Review, 89 (2), 311–316. Retrieved June 19, 2022, from http://www.jstor.org/stable/117127

[3] Kerr, William R., and Frederic Robert-Nicoud. 2020. *Tech Clusters*. Journal of Economic Perspectives, 34 (3): 50-76. DOI: 10.1257/jep.34.3.50

[4] Holzer, H. J. (2022, March 9). *Understanding the impact of automation on workers, jobs, and wages.* Brookings. Retrieved June 20, 2022, from https://www.brookings.edu/blog/up-front/2022/01/19/understanding-the-impact-of-automation-on-workers-jobs-and-wages/

[5] *Data, information, knowledge, and Wisdom*. (n.d.). Retrieved June 19, 2022, from https://homepages.dcc.ufmg.br/~amendes/SistemasInformacaoTP/TextosBasicos/Data-Information-Knowledge.pdf

[6] Rüdiger Wirth. (n.d.). *CRISP-DM: Towards a standard process model for data mining.* Retrieved June 19, 2022, from http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

[7] *CRISP-DM Help Overview*. IBM Documentation. Retrieved June 20, 2022, from https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=dm-crisp-help-overview

[8] G. Aceto, V. Persico and A. Pescapé, *A Survey on Information and Communication Technologies for Industry 4.0: State-of-the-Art, Taxonomies, Perspectives, and Challenges, in IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3467-3501, Fourthquarter 2019, doi: 10.1109/COMST.2019.2938259.

[9] E. U. R. O. P. E. Y. O. U. (1970, May 2). *What is information and Communication Technology?* Retrieved June 19, 2022, from https://europeyou.eu/es/what-is-information-and-communication-technology/

[10] III, G. N. R. (2021, November 20). *Effectiveness, efficiency and organizational structure.* Bizfluent. Retrieved June 19, 2022, from https://bizfluent.com/info-7829850-effectiveness-efficiency-organizational-structure.html

[11] Datta, Y. (2010). *A critique of Porter's cost leadership and differentiation strategies*. Chinese Business Review, 9 (4), 37.

[12] Camila Balbontin, David A. Hensher, *Firm-specific and location-specific drivers of business location and relocation decisions*, Transport Reviews, Volume 39, Issue 5, 2019, Pages 569-588, ISSN 0144-1647, https://doi.org/10.1080/01441647.2018.1559254.

[13] Marinkovic, Sanja & Nikolić, I. & Rakicevic, Jovana. (2018). *Selecting location for a new business unit in ICT industry*. Zbornik Radova Ekonomskog Fakultet au Rijeci. 36. 801-825.

[14] Rajkumar, P. (2013). A STUDY OF THE FACTORS INFLUENCING THE LOCATION SELECTION DECISIONS OF INFORMATION TECHNOLOGY FIRMS. *Asian Academy of Management Journal*, *18*, 35–54. https://www.researchgate.net/profile/Rajkumar-Paulrajan/publication/259442687_A_Study_of_the_Factors_Influencing_the_Location_Selection_Decisions_of_Information_Technology_Firms/links/00b7d52b972c0b6616000000/A-Study-of-the-Factors-Influencing-the-Location-Selection-Decisions-of-Information-Technology-Firms.pdf

[15] *The Digital Economy and Society Index (DESI)*. (2021, November 12). Shaping Europe's Digital Future. Retrieved June 19, 2022, from https://digital-strategy.ec.europa.eu/en/policies/desi

[16] Deloitte. (2021, February). *Digitalisation, an opportunity for Europe.*

[17] *Horizon 2020*. (2020). European Commission - European Commission. Retrieved June 19, 2022, from https://ec.europa.eu/info/research-and-innovation/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-2020_en

[18] Tom Gresham. U.S. Regional Reports. (2018, September 20). *Technology upending the location decision process.* Area Development. Retrieved June 19, 2022, from

https://www.areadevelopment.com/advanced-manufacturing/Q3-2018/technology-upending-the-location-decision-process.shtml

[19] Team, K. (2017, July 18). *13 Factors Affecting Location Of Industries*. Business Finance and Accounting Blog. https://www.knowledgiate.com/location-of-industries/

[20] *Indicators of entrepreneurial determinants - OECD*. (2020). OECD. Retrieved June 19, 2022, from https://www.oecd.org/sdd/business-stats/indicatorsofentrepreneurialdeterminants.htm

[21] *pandas - Python Data Analysis Library*. (2022). Pandas. https://pandas.pydata.org/

[22] *GeoPandas 0.11.0 — GeoPandas 0.11.0+0.g1977b50.dirty documentation*. (2022). Geopandas. https://geopandas.org/en/stable/

[23] *NumPy*. (2022). Numpy. https://numpy.org/

[24] Cantwell, J., & Piscitello, L. (2002). The location of technological activities of MNCs in European regions. *Journal of International Management*, *8*(1), 69–96. https://doi.org/10.1016/s1075-4253(01)00056-4

[25] *Data.europa.eu*. (n.d.). Retrieved April 20, 2022, from https://data.europa.eu/data/datasets/cordish2020projects?locale=es

[26] *Europe 5G Readiness Index*. (2020). inCITES Consulting. Retrieved April 20 2022, from https://www.incites.eu/europe-5g-readiness-index

[27] *Digital Innovation Hubs*. (2020). Smart Specialisation Platform. Retrieved April 20 2022, from https://s3platform.jrc.ec.europa.eu/digital-innovation-hubs-tool

[28] *The Complete List Of Unicorn Companies*. (2022, June). CBInsights. Retrieved April 20 2022, from https://www.cbinsights.com/research-unicorn-companies

[29] *Europe 5G Readiness Index Report*. (2020). Incites Consulting S.A. https://www.incites.eu/incites-map/Europe_5G_Readiness_Index_Report.pdf

[30] John. (2021, May 21). *The top 10 african tech hubs. African Vibes*. Retrieved June 20, 2022, from https://africanvibes.com/the-top-10-african-tech-hubs/

[31] **Unicorn English definition and meaning. *Lexico Dictionaries | English*. Retrieved June 20, 2022, from https://www.lexico.com/en/definition/unicorn

[32] *Current list of all 27 European union countries. List of 27 European Union member countries*. (n.d.). Retrieved June 20, 2022, from https://www.countries-ofthe-world.com/european-union-countries.html

[33] Joshi, N. (2022, April 14). *7 types of artificial intelligence*. Forbes. Retrieved June 20, 2022, from https://www.forbes.com/sites/cognitiveworld/2019/06/19/7-types-of-artificial-intelligence/?sh=1d6f0c32233e

[34] *Cybersecurity English definition and meaning*. Lexico Dictionaries | English. Retrieved June 20, 2022, from https://www.lexico.com/en/definition/cybersecurity

[35] *The NGI Initiative: An Internet of Humans*. (2022, June 15). Next Generation Internet. https://www.ngi.eu/about

[36] *Internet of things English definition and meaning*. Lexico Dictionaries | English. Retrieved June 20, 2022, from https://www.lexico.com/en/definition/internet_of_things

[37] *What is media? - definition from Techopedia*. Techopedia.com. Retrieved June 20, 2022, from https://www.techopedia.com/definition/1098/media

[38] Juliet, Gabbie, Go, N., & Abigail. (2022, May 31). *Robotics in 2022: Types of robots that we use*. Robots.net. Retrieved June 20, 2022, from https://robots.net/robotics/types-of-robots/

[39] *Types of software*. GeeksforGeeks. (2020, August 10). Retrieved June 20, 2022, from https://www.geeksforgeeks.org/types-of-software/

[40] Ward, S. (2020, June 29). What are smes? The Balance Small Business. Retrieved June 20, 2022, from https://www.thebalancesmb.com/sme-small-to-medium-enterprise-definition-2947962

[41] Pahwa, A. & consultant, A. P. A. startup. (2021, April 25). *What is Deep Tech? - use cases, examples & future.* Feedough. Retrieved June 20, 2022, from https://www.feedough.com/what-is-deep-tech/

[42] Team, T. I. (2022, June 11). *Foreign Direct Investment (FDI).* Investopedia. Retrieved June 20, 2022, from https://www.investopedia.com/terms/f/fdi.asp

[43] *Research and development. English definition and meaning*. Lexico Dictionaries | English. Retrieved June 20, 2022, from https://www.lexico.com/en/definition/research_and_development

[44] Baldridge, R. (2022, April 18). *Understanding Venture Capital*. Forbes. Retrieved June 20, 2022, from https://www.forbes.com/advisor/investing/venture-capital/

[45] *Gross domestic product (GDP) at current market prices by NUTS 3 regions*. (2022). Eurostat. Retrieved April 2022, from https://ec.europa.eu/eurostat/databrowser

[46] *EGOVKB | United Nations - E-participation index*. United Nations. Retrieved June 20, 2022, from https://publicadministration.un.org/egovkb/en-us/About/Overview/E-Participation-Index

[47] BDTwebSupport. (n.d.). *Next events*. Retrieved June 20, 2022, from https://www.itu.int/ITU-D/finance/work-cost-tariffs/events/tariff-seminars/Cost_modeling_training_Geneva08/index-results.html

[48] **What is an Internet exchange point? | How do IXPs work? *|* Cloudflare. Retrieved June 20, 2022, from https://www.cloudflare.com/es-es/learning/cdn/glossary/internet-exchange-point-ixp/

[49] *5G pilots: 5G technology to Mobile Robotics*. Robotnik. (n.d.). Retrieved June 20, 2022, from https://robotnik.eu/projects/pilotos-5g-en/

[50] Pandey, Amit & Jain, Achin. (2017). *Comparative Analysis of KNN Algorithm using Various Normalization Techniques*. International Journal of Computer Network and Information Security. 9. 36-42. 10.5815/ijcnis.2017.11.04.

[51] Harmouch, M. (2022, January 7). *17 types of similarity and dissimilarity measures used in data science.* Medium. Retrieved May 2022, from https://towardsdatascience.com/17-types-of-similarity-and-dissimilarity-measures-used-in-data-science-3eb914d2681

[52] MDMCourse. (n.d.). *Lecture 3: Similarity and distance measures*. https://mycourses.aalto.fi/pluginfile.php/1261250/course/section/168493/distance%20%281%29.pdf

[53] Goswami, M., Babu, A., & Purkayastha, B. (2018). *A Comparative Analysis of Similarity Measures to find Coherent Documents*. International Journal of Management, Technology And Engineering Volume 8, Issue XI, NOVEMBER/2018, 8(XI). http://www.ijamtes.org/gallery/101.%20nov%20ijmte%20-%20as.pdf

[54] Zahrotun, L. (2016). *Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method*. Computer Engineering and Applications, *5*(1). https://core.ac.uk/download/pdf/86430721.pdf

[55] McKenzie, G., & Adams, B. (2020). *A country comparison of place-based activity response to COVID-19 policies.* National Library of Medicine. https://doi.org/10.1016/j.apgeog.2020.102363

[56] *Communication: Digitising European industry - reaping the full benefits of a Digital Single Market. Shaping Europe's digital future*. (n.d.). Retrieved June 20, 2022, from https://digital-strategy.ec.europa.eu/en/library/communication-digitising-european-industry-reaping-full-benefits-digital-single-market

[57] *Concept. Smart Specialisation Platform.* (n.d.). Retrieved June 20, 2022, from https://s3platform.jrc.ec.europa.eu/s3concept

[58] *Home. Smart Specialisation Platform.* (n.d.). Retrieved June 20, 2022, from https://s3platform.jrc.ec.europa.eu/homem (europa.eu)

[59] Wikimedia Foundation. (2022, May 19). *Web application*. Wikipedia. Retrieved June 20, 2022, from https://en.wikipedia.org/wiki/Web_applicationWikipedia

[60] *Cloud • Streamlit*. (2022). Streamlit Cloud. Retrieved May 2022, from https://streamlit.io/cloud

[61] Pelayo, C. (2022, June). *GitHub - carmenpelayo/Location-Recommendation-System-for-ICT-Businesses*. GitHub. Retrieved June 2022, from https://github.com/carmenpelayo/Carmen-Pelayo/tree/main/Location-Recommender

[62] Pelayo Fernandez, C. (2022, June). *Location Recommendation System for Businesses in the European ICT Industry*. Streamlit Cloud. Retrieved June 2022, from https://locationrecommender.streamlit.app/

[63] : *Real Decreto 1495/2011, de 24 de octubre, por el que se desarrolla la Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público, para el ámbito del sector público estatal*. BOLETÍN OFICIAL DEL ESTADO, 269, 8/NOV/2011

[64] *Reutilización de la Información del Sector Público*. PAe. (n.d.). Retrieved June 20, 2022, from https://administracionelectronica.gob.es/pae_Home/pae_Estrategias/pae_Gobierno_Abierto_Inicio/pae_Reutilizacion_de_la_informacion_en_el_sector_publico.html#.YrCmni8lNhB

[65] La Moncloa. 10/12/2021. *El Gobierno aprueba el Proyecto de Ley de Startups para favorecer el emprendimiento innovador [Consejo de Ministros/Resúmenes]*. (n.d.). Retrieved June 20, 2022, from https://www.lamoncloa.gob.es/consejodeministros/resumenes/Paginas/2021/101221-rp_cministros-extraordinario.aspx

[66] COM(2012) 795 – *Entrepreneurship 2020 Action Plan: Reigniting the entrepreneurial spirit in Europe.*

[67] *2011/833/EU: Commission Decision of 12 December 2011 on the reuse of Commission documents* OJ L 330, 14.12.2011, p. 39–42 (EN)

[68] *Terms of service*. CB Insights. (2022, April 11). Retrieved June 20, 2022, from https://www.cbinsights.com/terms-of-service/

[69] *Copyright notice and free re-use of data - Eurostat.* (n.d.). Retrieved June 20, 2022, from https://ec.europa.eu/eurostat/about/policies/copyright

[70] Google. (n.d.). How licensing works. Google Workspace Admin Help. Retrieved June 20, 2022, from https://support.google.com/a/answer/6309862?hl=en

[71] *Licensing terms for Project Jupyter code - Project Jupyter Governance.* (n.d.). Retrieved June 20, 2022, from https://jupyter.org/governance/projectlicense.html

[72] *GitHub terms of service. GitHub Docs.* (n.d.). Retrieved June 20, 2022, from https://docs.github.com/en/site-policy/github-terms/github-terms-of-service#g-intellectual-property-notice

[73] *Terms of use • streamlit*. Streamlit. (n.d.). Retrieved June 20, 2022, from https://streamlit.io/terms-of-use

[74] *Trabajo de Fin de Grado | UC3M*. (2022). Trabajo de Fin de Grado. https://www.uc3m.es/ss/Satellite/SecretariaVirtual/en/TextoDosColumnas/1371241563580/Trabajo_de_Fin_de_Grado

[75] *Grado en Empresa y tecnología*. UC3M. (n.d.). Retrieved June 20, 2022, from https://www.uc3m.es/grado/empresa-tecnologia#programa

[76] *Salario de Jefe de proyecto en España.* (n.d.). Retrieved June 20, 2022, from https://es.indeed.com/career/jefe-de-proyecto/salaries

[77] *Salario de Data scientist en España*. (n.d.). Retrieved June 20, 2022, from https://es.indeed.com/career/data-scientist/salaries?from=top_sb

[78] Mora, A. (2021, July 7). *¿Cuánto Tiempo te va a durar El Mac? Macworld España*. Retrieved June 20, 2022, from https://www.macworld.es/articulos/mac/anos-vida-mac-3783831/

12. Annex

12.1. Computation of the correlation between the investment in H2020 projects and the European countries' GDP.

In the computation of the correlation coefficient the next steps were followed:

1. Extract the necessary datasets: *Horizon_2020.xlsx* [25] for data in Horizon 2020 projects, and *GDP_countries.xls*x [45] for data in EU27 GDPs.
2. Select the variables to study:
    - From *Horizon_2020.xlsx* → columns `'Country Code'` and `'EU Contribution (€)'`. Then group `'EU Contribution (€)'` values by country while applying a sum to obtain the aggregate (at a national level) investment in ICT → `df_p.groupby('Country Code').sum().reset_index()`
    - From *GDP_countries.xls*x → columns `'country'` and `'gdp'`.
3. Once we have two columns in each dataset, we can merge them with respect to the country code, so that we get a single dataframe of three columns: one corresponding to the countries' code, other to the countries' investment in ICT and the last one to the countries' GDPs. → `'pd.merge(df_gpd, df_countries, how='inner', on='country')'`
4. The final step consists of computing the correlation coefficient between the columns on investment and GDP → `'df['gpd'].corr(df['grants'])'`. When printing the results, this gives us a correlation coefficient of 0.86, which is a very significant value.