First Clustering Analysis of COVID in Portugal

Ana Teresa Ferreira[1], José Vieira[2], Manuel Filipe Santos[1], and Filipe Portela[1,2](✉)

[1] Algoritmi Research Centre, University of Minho, Guimarães, Portugal
cfp@dsi.uminho.pt
[2] IOTECH- Innovation on Technology, Trofa, Portugal

Abstract. There is an increasing need to understand the behavior of COVID-19, in this case, what type of medical preconditions can influence the recovery of the infected patient and what age groups are more affected. After the Directorate-General of Health of Portugal (DGS) made available the first records gathered from the infected, it became possible gather some conclusions. In this context, ioCOVID19 project arises, which wants to identify patterns and develop intelligent models able to support the clinical decision.

This article explores which typologies are associated with different outcomes to provide some insights regarding the consequences after the coronavirus infection. To understand which profiles, stand out, a clustering algorithm was used, 65 experiments were carried out, from which 192 clusters were obtained. From this study, the most relevant profiles are the following: the profile associated with death are patients with Diabetes – aged between 44 and 98 years old (19.74%); regrading hospitalized patients who died, the profile achieved was patients with Chronic Kidney Disease – aged between 52 and 102 years old (17.63%); for patients hospitalized in ICU who died the profile obtained was Cardiovascular Diseases – aged between 61 and 88 years old (26.23%); in regards to patients who died after being submitted to ventilatory support the correlated profile are patients with Cardiovascular Diseases – aged between 62 and 99 years old (32.17%). With the completion of this study it was possible to detect a set of profiles that are associated with different clinical conditions.

Keywords: COVID-19 · Clustering · Information Systems · Statistics · Public Health

1 Introduction

With the increase of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) cases, it becomes necessary to understand what outcome can be gathered from the data collected by medical health professionals in the field. Portugal is one of the European Countries where Coronavirus has a significant impact, and the number of cases and deaths increase every day. To provide some inputs to the decision process, a research project was released – ioCOVID19 aims to develop an intelligent decision support platform that predicts the evolution of the disease in a specific patient. Providing support to clinicals

J. M. Machado and H. Peixoto (Eds.): AISCOVID 2022, LNICST 485, pp. 51–61, 2023.
https://doi.org/10.1007/978-3-031-38204-8_5

in the fight against COVID-19. Given that, this article was developed as an integral part of the research project depicted and its essentiality falls within the phase of Data Understanding and Preparation using data mining techniques.

The aim of the study is to understand the trends of the virus considering the preconditions and age of the infected. In this context, this work helps health professionals in the moments of crucial decision making – based on previous cases of infected patients who can provide some output regarding the outcome that may result. For this, clusters were created based on the characteristics of those affected. In this way, it's possible to understand in which group the infected patient may be inserted in and if these preconditions influence in any way progression of the disease (coronavirus). The data found in this article goes back to June of 2020 and refers only to the Portuguese population. At this moment according to Directorate-General of Health (DGS) of Portugal, the average age of deaths due to COVID-19 is 81.4 years old [1].

The article presents the following structure: first, to situate the reader in the theme and problem addressed, a short introduction to the subject is presented. Next, it's described with more detail what type of data mining technique are used during the study, and an analysis of the data is also provided. The main themes of the article are detailed in the Background section, as well as information related to previous studies developed about the same matter. Then, it is portrayed in more detail which materials and methods were used for the development of the project, such as which methodologies were adopted and what kind of data was utilized. Regarding the Case Study point, the first 3 phases of CRISP-DM are exposed in more detail based on the project's objective. Then, in the Modeling phase, all the relevant information used during the data clustering process is presented. In the section referring to the Results, all relevant results and information obtained during the creation of the clusters are exposed. Lastly, in the Discussion section, all results disclosed in the previous points are discussed and evaluated in detail in a critical way.

2 Background

This section presents the relevant topics of the article, shows the Portugal situation at the time, and mentions some related works.

2.1 COVID-19

COVID-19 is the official name given by the World Health Organization (WHO) to the disease caused by the new coronavirus SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2), which can cause a serious respiratory infection such as pneumonia. This virus was first identified in humans in the Chinese city of Wuhan, Hubei province, at the end of 2019 [2].

2.2 Portuguese Reality of COVID-19

According to official data, on June 30 of 2020, the scenario in which Portugal found itself was as follows: 42 171 confirmed cases, 27 505 recovered cases and 1 576 cases of

death [3]. To better understand the effect of the pandemic in Portugal, the mortality rate of the coronavirus on June 30, 2020, and the deadliest diseases in the country. According to available data, for the year of 2019, the three most deadly diseases were the following [4]: Diseases of the circulatory system (represents 29,9% of deaths), Malignant tumors (25,5%) and Respiratory system diseases (10,9%). In contrast, on 30 June, the mortality rate due to COVID-19 was 3.74% [5].

2.3 Project ioCOVID19

The article in question is linked to the project under development - ioCOVID19 – Intelligent Decision Support Platform NORTE-01-02B7-FEDER-048344. It aims to create an essential platform for clinicians to combat COVID-19, and its main objective is to analyze the available data referring to those infected by coronavirus in Portugal and to predict the evolution of the disease of a given patient from a set of predictive models. Through the use of open data accessible online and made available by the SNS (Portuguese National Health Service) and DGS, it is possible to categorize the type of patients, assess the impact that each variable has on the course of the disease and predict the type of patient discharge. A Web/Mobile platform - ioCOVID19 - is also being developed, which aims to allow doctors/nurses to access a set of essential data for decision making.

2.4 Data Mining

Data Mining refers to the ability to extract useful and relevant information from a large dataset. In this study field, the main objective is to find non-evident relationship between data or patterns, in other words, it's the process of discovering knowledge from data [6] that can add some value to the dataset owner.

2.5 Clustering

Data clustering is a technique used in Data Mining, to provide statistical data analysis. Clustering is the classification of similar objects in different groups, that is the separation of data into clusters. The data inserted in each cluster, ideally, has some common trait [6]. In general, clustering is about identifying a finite set of categories or clusters do describe the data.

2.6 Similar Works

Since the study focuses on a relatively recent disease, no study like the one presented in this article was found. However, studies related with the clinical field have already been developed using the Clustering technique. That said, the studied found was the following:

- Identification of clusters of symptoms that can interfere in the quality of life of patients with advanced cancer, thus allowing greater control of them by clinicians, to reduce side effects in those patients [7].

- Identification of risk factors associated COVID-19 deaths in Portugal, in order to allow a more efficient health services strategic interventions with a significant impact on deaths by COVID-19 [8].

3 Materials and Methods

The portrayed data was provided by DGS and SNS. It refers to patients infected with COVID-19, collected by clinicals between 2nd March and 30th June of 2020.

3.1 Design Science Research

Since this is a research project and to understand if it's possible to characterize the clinical typology of patients infected with coronavirus (as well as the outcome of the disease), two methodologies were followed - Design Science Research (DSR), as a research methodology and Cross-Industry Standard Process for Data Mining (CRISP-DM) as practical method. DSR consists of 6 phases [9]:

1. Identifying the problem and motivation;
2. Defining objectives of the solution;
3. Design and development;
4. Demonstration;
5. Evaluation;
6. Communication.

To set DSR in action, it's necessary to use a practical methodology to help drive the project, so, CRISP-DM was chosen.

3.2 CRISP-DM

CRISP-DM was the second methodology used. This method provides a global perspective on the life cycle of a data mining project. This cycle, shown in Fig. 1 - Project Workflow - is divided into 6 sequential phases. There are dependencies between them; however, it does not have a rigid structure. The phases of the current CRISP-DM model for data mining projects are Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. The information depicted in this document was achieved after the completion of the third phase - Data Preparation [10]. Regarding that, there is a more detailed description of the phases involved.

To drive this project is essential to do a relation between the research methodology and the practical method.

3.3 DSR and CRISP-DM

Since both methodologies are used concurrently, it's possible to point out the relationship between the two. This article portrays the three first phases of both CRISP-DM and DSR. For example, phase 1 and 2 of the DSR are directly linked to the first activity of CRISP-DM, as portrayed in the table. The remaining correlations are also shown in Table 1 – Crossover of CRISP-DM and DSR methodologies [9].

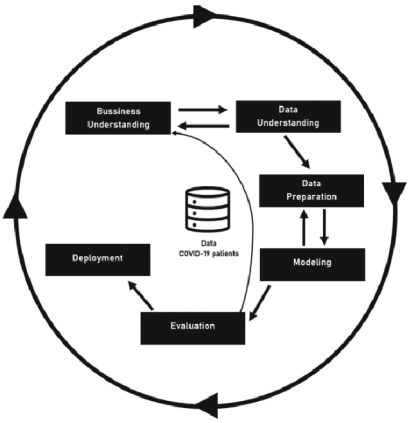

Fig. 1. Project Workflow

Table 1. Crossover of CRIPS-DM and DSR methodologies

Methodology	Activities	DSR Phases					
		1	2	3	4	5	6
CRISP-DM	Business Understanding	X	X				
	Data Understanding		X	X			
	Data Preparation			X			
	Modeling			X			
	Evaluation				X	X	
	Deployment				X	X	

3.4 Tools

To develop the project depicted, Python programming language was used. For such, to apply this language a set of libraries were applied to enable the preparation and the analysis of the data in question. The libraries employed were the following:

- Panda's library (allows the manipulation of data organized by columns);
- Matplotlib (provides different information visualization options);
- Scikit-learn or sklearn (is a machine learning software that provides various algorithms such as Classification, Regression and Clustering).

The Clustering algorithm used to support the study was K-Means Clustering. This algorithm seeks to divide all the observations under analysis into "k" clusters, with each observation belonging to the closest centroid. In this way, it's possible to survey observations that have some determining characteristic in each of the groups.

4 Case Study

The description of the case study goes through the methodology, presented in the CRISP-DM section, as it's possible to understand in the following points.

4.1 Business Understanding

The first phase, Business Understanding, focuses essentially on understanding the objectives and requirements of the project regarding a business perspective. From there, it's then possible to design a preliminary data mining project that can achieve the outlined objectives. Therefore, this project intends to develop a platform for clinicians to combat COVID-19, with the primary objective of predicting the evolution of the disease of a specific patient - evaluating the impact that each variable has on disease and predicting the type of outcome. For this study, the aim is to categorize the types of clinical patients in Portugal.

4.2 Data Understanding

At this stage, data analysis is carried out to search for possible quality problems and, consequently, obtain a better understanding of them. Since the data concerning COVID-19 was recorded by a wide range of health professionals, it was shown that it was inconsistent. The data provided has 38,545 records, which were collected between the 2nd of March and the 30th of June. To gather a better perception of the data, an annex - COVID-19 Data Analysis - was created where all the relevant information is exposed. Nevertheless, a global perception of the data is exposed in Table 2 – General Analysis:

Table 2. General Analysis

Study Groups	Number of records	% of records
Patients with Comorbidities	8326	21.60%
Admission	4327	11.23%
Deaths	1155	3.00%
Ventilatory Support	4327	11.23%
ICU Admission	253	0.66%
Recovered Patients	17 046	44.23%

4.3 Data Preparation

In this phase, all the changes applied are portrayed to build the final dataset, that is, the data that later is used to feed the modelling tools. The main tasks performed were cleaning data, selecting attributes, and building new attributes from existing data. Some of the main corrections/changes made were:

- The correction of fields that indicated pregnant men. This error was present in 3 records and represents 0.0078% of the data;
- Ignore records that precede the first confirmed official case in Portugal. This error was present in 23 records, which represented 0.060% of the data and were subsequently eliminated from the data;
- The creation of columns about other comorbidities through the column "If another, which one?" - such as the addition of the columns Cardiovascular Diseases, Obesity, Smoking, Tuberculosis, Rhinitis, Bronchitis, among others. This change affected 4970 columns, that is, 12.90% of the records.

5 Modeling

At this point, the entire process concerning the modelling and elaboration process of the clusters is exposed. To understand the best clustering algorithm to use, several have been studied and considered and several experiments were carried out to obtain the presented results and conclusions. That said, 33 experiments (3 Scenarios * 11 Targets * 1 Methods) were carried out and documented to better understand the preconditions or other factors such as gender, age, hospitalization, ICU service and the need for ventilatory support. Other experiments were also carried out to find the correlation between the various diseases under study, however, it wasn't possible to obtain relevant information on this point. In total 112 experiences were elaborated, which gives a total of 325 clusters. However, about half of these experiments refer to tests performed, to understand what kind of information and correlation it would be possible to extract from the data. Given that 33 experiments were then documented, as they were the ones that demonstrated to have relevant and usable information.

Therefore, the targets under consideration were as follows:

```
{
"Scenarios":{
        "S1":"All Comorbidities ",
        "S2":"Risk Comorbidities ",
        "S3":"Risk Comorbidities except Cardiovascular Diseases"
},
"Targets":{
        "TG1":"All infected people who died",
        "TG2":"All Hospitalized",
        "TG3":"All Hospitalized who died",
        "TG4":"Everyone who needed an Intensive Care Unit (ICU)",
        "TG5":"All who needed ICU and died",
        "TG6":"All infected who needed ventilatory support",
        "TG7":"All those infected who needed ventilatory support who died",
        "TG8":"All Recovered",
        "TG9":"All Hospitalized who recovered",
        "TG10":"All who needed ICU and recovered",
        "TG11":"All who needed ventilatory support and recovered"
},
"Methods":{
        "ALG1":"K - Means with Elbow Method"
}
}
```


Regarding all the mentioned targets, studies were carried out considering all the preconditions portrayed, all the preconditions considered at risk, all the risk preconditions (except for cardiovascular diseases), and risk preconditions with age or gender, however no relevant results were gathered by the last two, therefore, they were discarded. Comorbidities considered at risk are: Neoplasm, Diabetes, Human Immunodeficiency Viruses (HIV) and others Immunodeficiencies, Chronic neurological and/or Neuromuscular diseases (CNND), Asthma, Chronic Lung Disease, Hepatic pathology, Chronic Hematological Diseases, Chronic Kidney Disease, Chronic Neurological Disability, Obesity, Smoking and Cardiovascular Disease (since is considered the deadliest disease in Portugal.). However, as the designation of cardiovascular disease is very general (since it can represent more serious or less severe illnesses within the same branch), the study of risk comorbidities without this aspect was elaborated, to discover what other pre-conditions would also stand out.

To discover the most appropriate cluster number to be used for each target, the elbow method was used. This method analyses the percentage of variance explained as a function of the number in clusters. The first clusters add information but, after a certain point, the difference between the clusters decreases to the point of not adding relevant information. So, for the targets under study, with the help of the Elbow Method, it was found which number of clusters is the most appropriate and for the study in question, the number of clusters obtained varies from experience to experience. That said, the number of insured clusters varies between 2 and 4 clusters.

6 Results

In this section only the best results will be exposed. However, an annex - Clustering Results - with all collected clustering results will also be made available. Therefore, in Table 3 - "Best Clustering Results", the results of the attributes that demonstrate having some relevant information were exposed. For each target, the number and percentage of records for that target will be exposed, which scenario has obtained the best cluster, as well as, the affected ages, the highlighted disease and finally, from this sample, how many patients have the highlighted profile in the cluster.

Table 3. Best Clustering Results

Targets	Nº and % of records	Scenarios	Affected Ages	Comorbidities	% of records per cluster
TG1	1155 (3.00%)	S1	[44–98]	Diabetes	19.74%
		S3	[52–100]	Chronic Kidney	13.25%
			[42–99]	Neoplasm	6.23%
TG3	72 (0.19%)	S1	[57–98]	CNND	10.33%
		S2	[52–100]	Chronic Kidney	17.63%
		S3	[44–95]	Neoplasm	13.64%
TG4	253 (0.66%)	S1	[18–92]	Cardiovascular Disease (C.D.)	30.04%
			[0–96]	Chronic Lung	11.46%
		S3	[0–96]	Chronic Kidney	13.83%
TG5	61 (1.28%)	S1	[61–88]	C.D	26.23%
		S3	[52–84]	Chronic Kidney	11.48%
TG6	493 (1.28%)	S1	[0–97]	C.D	59.63%
		S3	[51–94]	Chronic Lung	17.04%
TG7	115 (0.30%)	S1	[62–99]	C.D	32.17%
		S2	[56–91]	Chronic Lung	22.61%
		S3	[55–97]	Chronic Kidney	20.00%
TG8	17046 (44.22%)	S1	[25–99]	C.D	6.20%
		S3	[12–96]	Neoplasm	2.37%
TG9	1645 (4.27%)	S1	[12–100]	Diabetes	19.70%
		S3	[34–95]	Neoplasm	17.26%
TG10	81 (0.21%)	S3	[44–80]	Diabetes	35.80%
TG11	167 (0.43%)	S1	[44–97]	Diabetes	35.93%
		S3	[56–94]	Chronic Lung	16.17%

7 Discussion

As already portrayed, the aim of the study is to find different profiles of infected patient. From the results shown in the annex – Clustering Results – it's possible to identify Cardiovascular Diseases and Diabetes in all the targets studied. This is because, a significant percentage of the Portuguese population is affected by cardiovascular diseases, which naturally results in the highlight of this comorbidity in almost all targets studied. Because of this, it's not possible to associate patients who have this type of comorbidities (cardiovascular disease and diabetes), to a positive or negative outcome, since it presents itself in targets referring to patients who died, recovered, admitted, admitted to the ICU, etc. The highlighted comorbidities were as follows: Cardiovascular Disease, Diabetes, Chronic Kidney Disease, Neoplasm, Chronic Neurological and/or Neuromuscular, Chronic Lung Disease.

Regarding the Chronic Kidney Disease comorbidity, it's typically associated with patients over 52 years of age, and is present in the targets TG1 (13.25%), TG3 (17.63%), TG4 (13.83%), TG5 (11.48%), TG7 (20.00%), and all these targets are associated with patients who died, who needed hospitalization and/or died and needed ICU. As for the comorbidity Neoplasm, it is typically associated with patients older than 34 years, and is present in the following targets: TG1 (6.23%), TG3 (13.64%), TG8 (2.37%) and TG9 (17.26%), that is, it's related with patients who required hospitalization, but does not have a specific outcome, since it is present in patients who have died and who have recovered. About Chronic Neurological and/or Neuromuscular Disease, it's present in patients between 57 and 98 years old, who required hospitalization and died, that is, associated with TG3 (10.33%). Finally, for Chronic Lung Disease, it's typically in patients over 50, who needed ICU, ventilatory support, being present both in patients who required ventilatory support who recovered or died, in other words correlated with the targets TG4 (11.46%), TG6 (17.04%), TG7 (22.61%), TG11 (16.17%).

8 Conclusion

The study depicted represents only a part of the ioCOVID19 project in development, and any updates to the results set out in this article can be found on the project's website. Important to remember that the results obtained represent only a small sample of patients infected by COVID-19 in Portugal, since it is an analysis of the first data provided.

Hereupon, it was possible to outline different types of patients with the data provided, that is, the aim of the project was achieved. So, from the clustering investigation carried out and evaluation of the results obtained it was assessed which comorbidities were associated with each target. That said, the main conclusions, in summary form, after the analysis of the data by the clustering technique were that the Chronic Kidney Disease, is mostly associated with targets that resulted in the patient's death, Neoplasia is associated with patients who needed hospitalization and Chronic Lung Disease is mostly associated with patients who have ICU or ventilatory support. However, it was not possible to outline a profile for patients with Diabetes and Cardiovascular diseases, since they are present in all targets studied. In the exposed article, only the most relevant results were portrayed, the interested reader should consult the official page of the project for more information, https://iocovid19.research.iotech.pt/.

Acknowledgements. This work has been developed under the scope of the project NORTE-01-02B7-FEDER-048344, supported by the Northern Portugal Regional Operational Programme (NORTE 2020), under the Portugal 2020 Partnership Agreement, through the European Regional Development Fund (FEDER). This work has also been supported by FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

References

1. DGS COVID-19 homepage. https://covid19.min-saude.pt/media-de-idades-dos-obitos-por-covid-19-e-81-4-anos/. Accessed 14 May 2021
2. COVID-19 | SNS24 [Internet]. SNS24. 2020 [cited 27 October 2020]. https://www.sns24.gov.pt/tema/doencas-infecciosas/covid-19/
3. Relatório de Situação. In: COVID-19. https://covid19.min-saude.pt/relatorio-de-situacao/. Accessed 14 May 2021
4. Óbitos por algumas causas de morte (%). In: Pordata.pt. https://www.pordata.pt/Portugal/%C3%93bitos+por+algumas+causas+de+morte+(percentagem)-758. Accessed 14 May 2021
5. Sete gráficos com a evolução da covid-19. Doentes internados em máximos de dois meses. In: Jornaldenegocios.pt. https://www.jornaldenegocios.pt/economia/coronavirus/detalhe/sete-graficos-com-a-evolucao-da-covid-19-em-portugal-taxas-de-crescimento-com-tendencia-de-queda. Accessed 14 May 2021
6. Bharati, M., Ramageri, M.: Data mining techniques and applications (2010)
7. Walsh, D., Rybicki, L.: Symptom clustering in advanced cancer. Support Care Cancer **14**, 831–836 (2006)
8. Nogueira, P.J., et al.: The role of health preconditions on COVID-19 deaths in portugal: evidence from surveillance data of the first 20293 infection cases. J. Clin. Med. **9**, 2368 (2020)
9. Fernandes, G.: Pervasive Data Science Applied to the Services Society. Master's Thesis, University of Minho, Guimarães, Portugal (2019)
10. Wirth, R., Hipp, J.: CRISP-DM: towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK, 1–13 April 2000