INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Data analysis applied to healthcare.

Ricardo Melo e Castro

Dissertation submitted as partial fulfillment of the requirements for the degree of Master in Telecomunications and Computer Engineering

Supervisor:
Doctor João Carlos Amaro Ferreira, Assistant Professor,
ISCTE-IUL

Co-supervisor:
Doctor Rúben Filipe de Sousa Pereira, Assistant Professor,
ISCTE-IUL

September, 2020

Acknowledgements

During this journey, which was the elaboration of the Master's thesis, that made me grow both professionally and personally, there were a number of people who were instrumental in achieving all this successfully and I would like to thank them.

First of all to my supervisor Dr. João Ferreira and co-supervisor Dr. Rúben Pereira who always showed themselves ready to help me overcome the difficulties I felt throughout the elaboration and helped me to carry out this thesis in the best possible way.

Next I would also like to thank the Garcia da Horta Hospital for providing the data that made this research possible.

I would also like to thank my colleagues Afonso Silva for the help they gave me when I needed it and my colleague Diogo Sousa for also having accompanied me throughout the development of this work.

Finally I would also like to thank my family, girlfriend and friends for all the love, motivation and comprehension that allowed me to successfully complete this important stage of my life.

Resumo

O serviço de urgências é um dos departamentos mais importantes de qualquer hospital, muitas vezes aberto 24 horas por dia é o serviço que mais fornece utentes a um hospital.

Posto isto, e perante uma população cada vez mais envelhecida, devido ao aumento da esperança média de vida, irá fazer com que a afluência a estes serviços venha a aumentar, assim é de extrema importância que o hospital seja capaz de realizar uma gestão eficaz dos recursos hospitalares disponíveis, de modo a que a qualidade de serviço que este pode oferecer aos utentes não seja comprometida e não se venha a verificar casos em que um utente não pode ser internado por falta de camas, ou que não tenha quarto disponível e tenha de ser colocado no corredor, entre outros.

Hoje em dia os hospitais com vista na possibilidade de análise de dados através de técnicas de Data Mining, que permitem a extração de conhecimento a partir das grandes quantidades de dados (Big Data) guardadas nos seus sistemas informáticos sob a forma de registos médicos eletrónicos (Eletronic Health Records).

A análise da Big Data produzida por um hospital vai permitir um olhar um para estes números e de certa forma ajudar a gestão do hospital a gerir melhor estes recursos de modo a melhorar os seus serviços, melhorando as condições dos utentes, deixando estes mais satisfeitos.

O autor desta tese irá analisar dados reais de um serviço de urgências fornecidos por um hospital da região de Lisboa, tratando destes dados e de seguida realizando uma análise descritiva que nos permita retirar conclusões em relação ao tempo que um doente passa no hospital (Length of stay), entre outras métricas, tendo em conta vários fatores que o possam influenciar, e de seguida elaborar uma análise preditiva do Length of Stay com vista a ajudar a apoiar o processo de tomada de decisões pelo hospital.

Palavras-Chave: Urgências Hospitalares, Big Data, Data Mining, Registos Médicos Eletrónicos, Dados Estruturados, Length of Stay, Gestão Hospitalar, Análise descritiva, Análise Preditiva.

Abstract

The Emergency Department (ED) is one of the essential departments of any hospital, often open 24 hours a day is the service that most provides patients to a hospital.

Advances in medicine have resulted in an increase in average life expectancy and, consequently, an ageing population, which will lead to an increase in affluence and need for healthcare, so it is of extreme importance that the hospitals are able to carry out effective management of available hospital resources, so that the quality of service (QoS) it can offer to patients is not compromised and there are no cases where a patient cannot be admitted due to lack of beds, or is not treated due to lack of medication, among others.

Hospitals today are aware of the growing importance of data analysis. Data Mining techniques allow extracting knowledge from the big data stored in their computer systems in the form of electronic health records (EHR). This analysis of big data produced by a hospital will, to some extent, help the hospital management to better manage these resources to improve its services, improving the quality of its service, making the patients more satisfied, which can contribute to a healthier society.

The author of this thesis will analyze data from an ED provided by an hospital in the Lisbon region, processing the data and then performing a descriptive analysis that allows us to draw conclusions regarding the time a patient spends in the hospital (Length of Stay), among other metrics, taking into account several factors that may influence it, and then perform a predictive analysis of the Length of Stay in order to help the hospital in the decision-making process.

Keywords: Emergency Department, Big Data, Data Mining, Electronic Medical Records, Structured Data, Length of Stay, Hospital Management, Descriptive Analyses, Predictive Analyses.

Contents

List of Figures

List of Tables

List of Equations

Abbreviations:

CRISP-DM - **C**ross-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining.

DM – **D**ata **M**ining.

ED – **E**mergency **D**epartment.

EHR – **E**letronic **H**ealth **R**ecords.

GDPR - **G**eneral **D**ata **P**rotection **R**egulation**.**

ICD - **I**nternational **C**lassification of **D**iseases.

ID – **Id**entification.

LoS – **L**ength **of** **S**tay.

MTP - **M**anchester **T**riage **P**rotocol.

NB – **N**aïve **B**ayes.

QoS – **Q**uality **of** **S**ervice.

RF – **R**andom **F**orest.

Chapter 1 – Introduction.

1.1 Motivation.

Technology, over the years, has been progressively present in human life, and it is in constant evolution and now is becoming an essential factor in many areas, healthcare is one of them.

Healthcare is by definition “the maintenance and/or improvement of health through prevention, diagnosis, and treatment of disease, illness, injury, and other types of problems related to the physical and mental states of human beings” [1].

The improvement of medicine also increased life expectancy, leading to an increasingly ageing population that can lead to ED crowding [2]. Crowding is the saturation of the ED services and can result in lousy service for the patients.

ED crowding is considered to be a significant international problem [3] and several authors have already studied its negative impacts on hospital metrics such as waiting teams and LoS , also influencing the level of patient satisfaction with the hospital [4].

The advances of medicine allied with the advances in technology open new possibilities in the way that diseases are faced. Hospitals are keeping tons of data in their databases under the form of Electronic Health Records, or Electronic Medical Records. The analyses of this data can provide information that can be helpful for the treatment and diagnosis of some diseases which will help improve patient care and the hospitals’ QoS [5].

In order to take advantage of the large volumes of data produced by hospitals, data scientists have used DM techniques, as these automatically extract relevant information from these large datasets [6]. This will allow to find patterns on the information and explore the potential of predictive analytics [7] to help make the ED process of admission and diagnose faster.

As will be further explained on section 1.3, this research aims to help improve hospital decision making through the development of a proof of concept of software using python language capable of finding patterns on the data as well as making a prediction of the patient's LoS with the goal of improving the hospital's decision-making process and ultimately the QoS and consequently the patients' satisfaction .

1.2 Research Questions.

With the elaboration of this thesis the author aims to answer the following research questions:

1- Can past data analytics help to predict patient length of stay (LoS)?

2- Can we identify patterns by analyzing past data that help us on the decision-making process?

3- Do big events near hospitals have any influence over the waiting times and LoS?

4- Do weather conditions and time of the year have any influence over the waiting times and LoS?

1.3 Objectives.

In order to be able to answer the research questions defined on section 1.2 the author will have to develop a proof of concept of software capable of:

- Extracting knowledge and medical information from structured data, with the aim of finding patterns and factors that influence hospital services.
- Given the input, the system gives a prediction of the patient's LoS, considering some factors.

To test the system, the author used a dataset from a Portuguese hospital containing structured data from the ED and preformed several DM techniques to process the information and find a model prototype for the system.

The system can be used both by the patients to obtain an estimate of how long they will stay at the hospital, and by the hospital management to identify factors that have influence over the quality of the services they provide and to get an idea of how long each patient will be occupying hospital resources (medical and nursing care, use of beds or appliances, etc.) in order to be able to manage these resources in the best possible way, with the goal of improving the hospital's QoS.

1.4 Research Method.

The research method used to develop this thesis will be the Data Science Research model [8]. The figure displayed below (Figure 1) represents the phases that constitute this method:

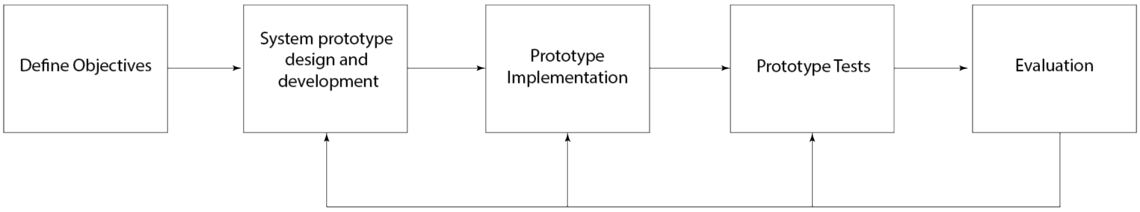


Figure 1 - Research Method Flow Chart.

The first phase is Defining objectives, this is the phase where we need to study and define the main goals and the means (technologies, methods and review related work) that will be used to develop the system.

After the objectives are defined, comes the second phase. This phase is where the components of the system will be developed, process the data that will be analyzed and prepare it for analysis, choose the algorithm to be used, and prepare an interface to display the information obtained from the analyses.

The third phase is where we implement the components developed in the previous phase together and check for errors if there are none the project's prototype should be developed.

The fourth phase is to test the prototype and check if it meets the objectives chosen in the first phase, and if not make the necessary changes to make sure that it meets the requirements to reach the final prototype.

The last phase is where the final prototype is tested and evaluated.

If the prototype does not fulfil the requirements, the second, third and fourth phases need to be repeated until the final prototype tested on phase five fits all the objectives defined on phase one.

1.5 Dissertation Structure.

The remaining of the document is divided in 3 different chapters. The first one is Chapter 2 - Literature Review, where the author reviews the state of the art concerning similar studies he found, and is divided into 4 subchapters Big Data in Healthcare, Data Mining, Predictive Analytics in Healthcare and Emergency Department Crowding and patient LoS.

In Chapter 3 - Model Building, in which the author explains goes through and the different phases that a data analysis and predictive analysis project requires and it is divided into 4 subchapters, Data Collection, Data Preparation, Descriptive Analyses and Predictive Model.

Finally in Chapter 4 – Conclusions, the author draws the conclusions of this research and addresses the limitations as well as the possible future work.

Chapter 2 – Literature Review.

This chapter is an overview of related work and is divided into four main topics: Big Data in healthcare (1), Data Mining (2), Predictive analytics in healthcare (3) and Emergency Department crowding and patient length of stay (4).

For each one of these topics, there will be a review of related work developed over the last years, case studies and challenges that projects like the one developed by the author can face.

For the first topic, Big Data in Healthcare, there will be an overview of the type of data that will be used in this thesis.

The second step Data mining is the process of extracting information from electronic health records (EHR) and analyzing that information.

The third topic predictive analytics in healthcare use DM techniques in order to help in many areas of the healthcare industry and is an area that has been studied and explored in the recent past.

The last topic is Department crowding and patient length of stay where the concepts of crowding and length of stay are explained as well as factors that may influence those phenoms.

2.1 Big data in healthcare.

In the medical context, big data is the bio-medical information that hospitals store in their databases and it can be used to improve the healthcare system. The availability of big data represents a vast range of opportunities to improve our healthcare industry through the analysis of this information and is seen as the future trend for research and treatment according to R. D. Todor and C.V. Anastasiu [1]. In order to extract the critical part of that information data scientist apply multiple DM techniques.

In big data for healthcare specialists like Wullianallur Raghupathi and Viju Raghupathi [9] refer to the "Four VS": The first "V" stands for volume. There is an enormous amount of data for the data scientists to work on. The second one stands for variety, there is a variety of data that lets us connect different sources together (some sources can be the clinical record, smartwatches and fitness bands, medical images, sensor data, genomics, are some examples) , offer new clinical information that can be analyzed. The third one stands for velocity, sometimes data need to be processed and analyzed live. The last one is veracity, because there are lots of errors, missing data, irrelevant data that needs to be taken care of in order to obtain the best conclusions possible.

The authors also argue that in addition to clinical benefits such as detecting diseases in early stages that are easier and more effectively curable accordingly to Bernard Marr [10], data analysis can also bring monetary advantages of millions of dollars.

The process of analyzing this enormous amount of information is similar to the traditional health informatics analyses. Still, it requires a different processing method known as distributed processing as it would not be possible to do in a single laptop, and the tools used are also different from the traditional ones.

There are different methodologies but one that is used and simple, it requires 4 steps. The first step is where the team should develop a concept statement based on the 4 V's.

For the second step, the team will take the concept of step 1 and go into detail, as well as search for existing projects that can be similar and answer some questions that justify the need to develop the project, as the cost of big data analytics is high.

The third step is where the team identifies the variables, collect data, selects the platforms, the tools, create a conceptual model, applies analytic techniques and finally obtains the results.

The last step is for testing the models and validate the results obtained in the previous step, and the implementation is phased so it minimizes the risk of failure.

Caroll McDonld [11] states that the use of advanced data analysis technologies, will revolutionize the healthcare industry and Todor and Anastasiu [1] believe that the use of big data brings advantages to Patients, medical practitioners, hospital operators, Pharma and clinical researchers and healthcare insurers.

The implementation of Big Data in healthcare will face some challenges like the initial cost of establishing the BDA infrastructure, the cost of storing the data and the costs of data analysis, and the algorithms used are often not documented or verified

[12], which means that these technologies still have limitations and the result of the prediction may not always be reliable [13] and has its consequences, but in the end, it will help avoid preventable deaths.

2.1.1 Structured Data.

There are two main types of Data depending on its characteristics, one is Structured Data and the other one unstructured Data.

Structured data refers to data or information well-organized data where everything has proper naming and can be stored into columns and rows [14]. This makes management and analyses of this type of data easier, as it is usually inserted into RDBMS which makes it easy to access through DM techniques and algorithms or SQL queries.

Mohammad B. Ateya, Brendan C. Delaney and Stuart M. Speedie conducted a study to evaluate the importance of structured data in EHR [15] and compared it to previous studies made by other authors in that and the conclusions drawn from this study were that a large amount of patient's information can be found as structured data elements in EHR, and the use of this data can be useful to speed up the screening process for enrolling patients and thus improve patient care and clinical research.

2.2 Data Mining.

Data mining is the process of automatically finding useful information in large data repositories, revealing patterns that may be relevant for analysis, and the use of these techniques is increasing in the healthcare industry [6].

DM has a method and procedures that are chosen accordingly to the data. One methodology used is the one explained by D. Olson and D. Delen in their book [16], Cross-Industry Standard Process for Data Mining or CRISP-DM which has 6 steps as shown in Figure 2 retrieved from the book mentioned above:

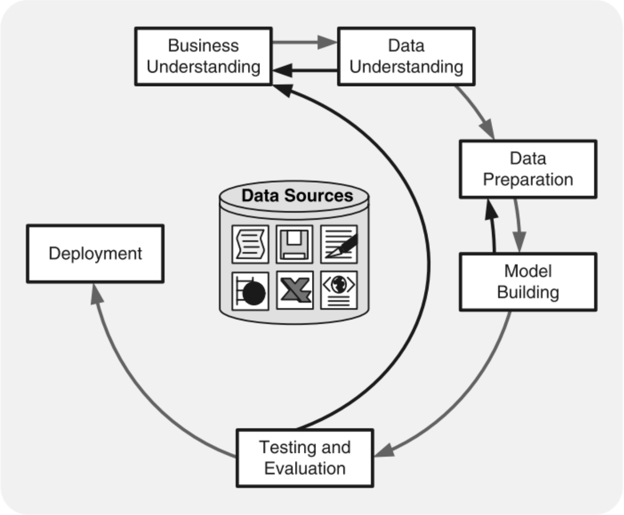


Figure 2 - CRISP-DM phases [12].

The first phase, Business Understanding, is to define objectives and goals for DM and develop a project plan. The second phase Data Understanding is to collect data and identify Data sources and verifying data quality. The third phase, Data Preparation is where data cleaning and transformation processes take place. The fourth phase Model Building is to carry an initial analysis of data for further understanding and then apply proper models on given data. The division of data into training and test sets is also done on this phase. The Evaluation phase is where the models prepared in the previous phase are tested to check if they fit the requirements identified in the first phase. This is an iterative process that may require moving towards previous phases until the model finally satisfies the objectives. The last phase is a deployment where an efficient model is obtained through the previous phases of the process.

Studies conducted by G. Stiglic et al. [17] revealed that the presence of corrupted or missing information in the analyzed datasets affect the results obtained through these analyses, so the data should be prepared in such a way as to ensure its quality in order to obtain results with the maximum possible veracity.

2.3 Predictive analytics in Healthcare.

The use of predictive analytics and data mining applied to management in the healthcare industry has increased in the last years.

According to Delen and Demirkan cited in [18], "Data mining and predictive analytics aim to reveal patterns and rules by applying advanced data analysis techniques on a large set of data for descriptive and predictive purposes".

The evolution of technologies like Cloud computing, Cognitive Reasoning and Pseudo-Intelligence, combined with the fact that more and more data is being generated and need to give the machines the upper hand as accordingly to [19] machines have become more reliable than humans in this matter.

The use of predictive analytics in healthcare can be useful to help predict diseases based on the analyses of patterns. Through the reports of data using machine learning and DM techniques, decision trees can be used for clinical decision making as shown in [20].

Predictive analysis was already used in areas such as politics to make predictions for elections, and is now also used in the healthcare area to help diagnose various types of cancer, the presence of myocardial infarction and also to make predictions for the mortality rate of diseases such as pneumonia [21].

In the same study, conducted by A. T. Janke, D. L. Overbeek, K. E. Kocher and P. D. Levy, we can read some of the opportunities predictive analytics give in emergency care. When a patient enters the ED, they undergo several tests and examinations to identify the problem that led the patient to the ED and make the diagnosis. The use of the predictive analysis' capabilities can decrease the number of necessary procedures, reducing the expense of resources and costs on the hospital side and making this screening process more agile.

However, to be able to use efficiently predictive analytics, it requires enough patient data, the more data we have, the more accurate the results will be, and these analyses are more useful during the diagnoses and therapeutic interventions.

Authors such as Ben Van Calster, Laure Wynants, Dirk Timmerman, Ewout W Steyerberg and Gary S Collins in [22] argue that the performance of algorithms must always be verified, since they behave differently from case to case. They also defend that the algorithms should be more transparent and should be described in full detail, to ease the institutions to choose wheather or not the algorithm fits its needs.

Haritha Chennamsetty, Derek Riley and Suresh Chalasani [23] developed a system to analyze Electronic Health Records based on Hive which is a scalable and dynamic data warehouse which can easily generate reports, graphs and charts that doctors and researchers can use to better understand the effects of medication on patients, based on their electronic medical records.

Another example of usage of this type of analytics in healthcare was made by Gonçalves F. [24], who developed a system that gave a prediction for the ED's waiting times given input and in order to choose the algorithm for the model compared the two different ones, the Naïve Bayes (NB) and Random Forest (RF), and the second one was the algorithm with best results.

Medicine is a very old profession with several centuries of history, and as such with traditions created regarding the doctor-patient relationship. Seuli Bose Brill Karen O. Moss and Laura Prater [25] have studied the impact that the use of Big Data analysis technologies and predictive analysis have on this relationship and have concluded that it must evolve and adapt to include the use of these technologies in medicine with a view to improving the health of the population.

2.4 Emergency Department Crowding and Patient Length of Stay.

There are several steps that a patient needs to follow when entering an ED, the check-in, assessment, treatment and final outcome.

Ida Mentzoni, Stig Tore Bogstrand and Kashif Waqar Faiz [26] studied the relation between crowding and patient's LoS using data from a norwegian hospital the Akershus University Hospital after they expanded the hospital's catchment area and consequently the number of patients.

They define crowding as a situation in which the identified need for emergency services exceeds available resources for patient care in the ED, hospital, or both, and reduces the quality of the services and can lead to hostile outcomes.

The LoS is the time spent since the patient checks into the hospital until he receives the discharge and goes home.

The results of this study showed that LOS increased proportionally to crowding. The increase in LOS means that patients will spend more time in hospital spending resources such as medical or nursing care, occupying beds, using machines, doing the service provided by the hospital slower.

Chaou, C. H. et al. on [27] studied what patient-related factors and the influence they had on the ED LOS for discharged patients. To conduct this study, they used EHR of all discharged patients from an of a tertiary teaching hospital in 2013 and a multivariate accelerated failure time model to determine the variables and its influence on LOS. In their study, they identified and quantified some influential demographic factors for the ED LOS such as: patient's age, patient's category, triage acuity level, gender, arrival time, and transfer status.

A group of researchers conducted a study of the impact of emergency crowding on the waiting room, treatment and boarding times [28]. To do that they used data of four different Eds collected for over a year, and the results they obtained were the time spent on the waiting room was increased, but it didn't influence much the treatment times.

Crowding is associated to increasing odds of hospital admission and mortality after discharge from ED on adults, and Quynh Doan et al. made a study to evaluate the association between crowding and the odds of some adverse outcomes among children seen at a pediatric ED [29], they used data from eight different Canadian pediatric EDs between 2010 and 2014 and analyzed the mean LOS for each hospital visit and hospital admission within seven days or death within 14 days of ED's discharge using mixed-effects logistic regression modelling.

The results were that ED crowding was not associated with hospital admission within 7 days of the ED visit or mortality in children after 14 days. On the other hand, they found a relation between ED crowding and hospital admission at ED visit among high-acuity visits, as well as return visits within 7 days among low-acuity visits with increasing ED crowding and so concluded that pediatric ED crowding also has implications for use of health services but not as in adults.

2.5 Comparison between researches.

The model elaborated by the author differs from the other similar studies already carried out shown on Table 1 due to the fact that these focus on the relationship between crowding and LoS [25] and demographic factors and LoS [26], while the author will look for an eventual influence between factors such as the disease presented by the patient, the weather conditions (occurrence of precipitation), the time of the year (season), existence of events in the city of the hospital under study (concerts and football matches) and the LoS and waiting times. The author will also apply a prediction algorithm based on these factors to predict the LoS and develop a user interface for an easier use of that model.

Date	Study focus	Research
2019	Relation between crowding and LoS.	[26]
2017	Impacts of demographic factors (patient's age and gender) on the LoS.	[27]
2018	Predictive analyses applied to waiting times.	[24]
2015	Predictive Analysis applied to pneumonia's mortality rate.	[21]

Table 1 – Most similar researches overview.

Chapter 3 – Model Building.

In this chapter, the used methodology in this thesis will be explained and will be divided into 4 different steps: Data Collection, Data Preparation, Descriptive Analyses and Predictive Analyses.

The first step is Data Collection, where the author collected the data needed for the research from Hospital Garcia da Horta in Almada, Lisbon (further described on section 3.1).

The second step is the Data Preparation, which is divided into two distinct parts, the first is the Data Cleaning, in which the author has deleted incomplete or inconsistent data from the dataset, and the second part is Additional Data, in which the author has added new fields to the database, some of those fields were calculated from existing fields and others added from external databases (further described on Section 3.2).

The third step is the Descriptive Analyses, that will allow an overview of numerous hospital metrics and find patterns that might help on the decision-making process (further described in Section 3.3).

The fourth and final step is the predictive analyses where the author will compare two DM algorithms and create a prototype prediction model to estimate the patient's LoS considering some factors (further described in Section 3.4).

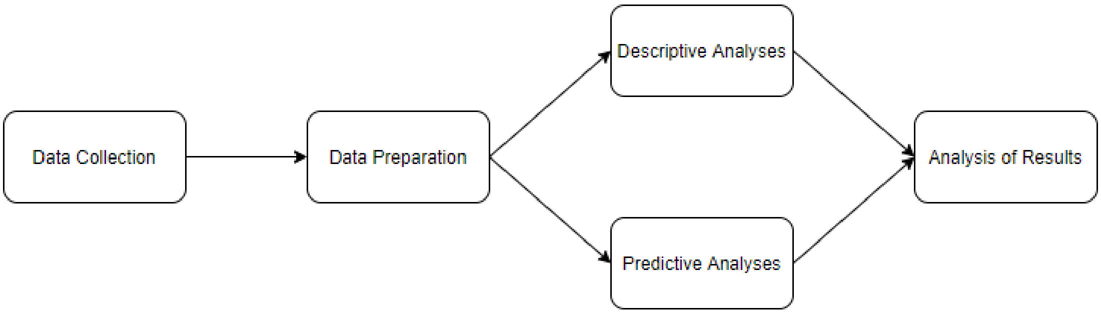


Figure 3 - Research's workflow.

3.1 Data Collection.

The first step is Data Collection, where the data that will be used to perform the descriptive and predictive analyses was collected from Hospital Garcia da Horta in Almada, Lisbon.

Every time a patient enters the ED, there is a set of procedures to follow, and can be observed on Figure 4:

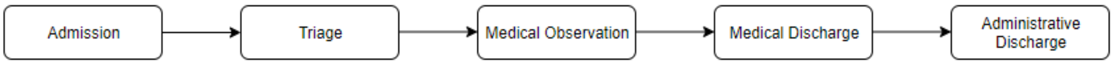
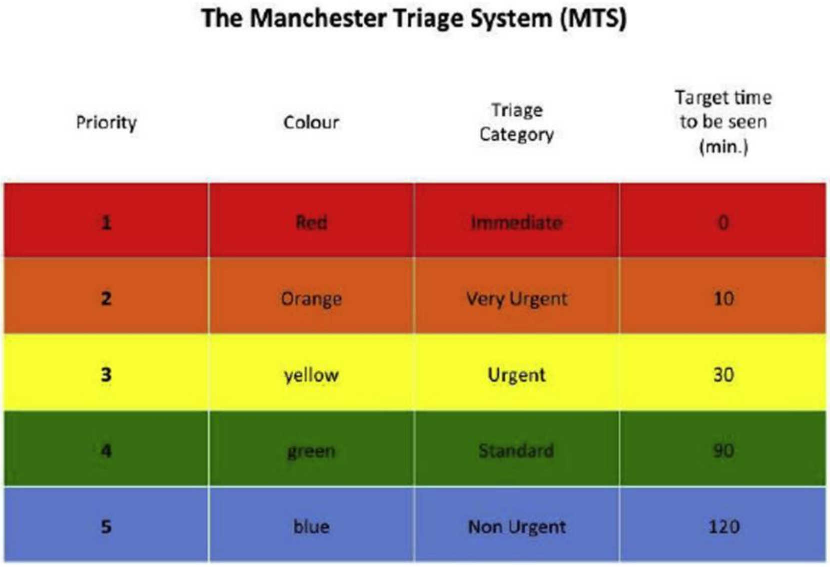


Figure 4 - Emergency Department's Procedures flow.

Whenever a patient enters the ED a new record is created and stored in the hospital database and the patient proceeds to the screening process, in which the patient is examined by a nurse and assigned a colour according to the MTP (Manchester Triage Protocol). There are 5 different colours, each with an associated priority level and a recommended waiting time as represented in Figure 5.



The Manchester Triage System (MTS)

Priority	Colour	Triage Category	Target time to be seen (min.)
1	Red	Immediate	0
2	Orange	Very Urgent	10
3	yellow	Urgent	30
4	green	Standard	90
5	blue	Non Urgent	120

Figure 5 - Manchester Triage Protocol [30].

After medical observation, if necessary, some treatment will be carried out, otherwise, the patient will receive the medical discharge and subsequently the administrative discharge.

The data extracted has 108295 cases and 18 parameters each and ranges from 01/01/2017 to 31/12/2017. The parameters presented in the dataset are displayed in Figure 6:

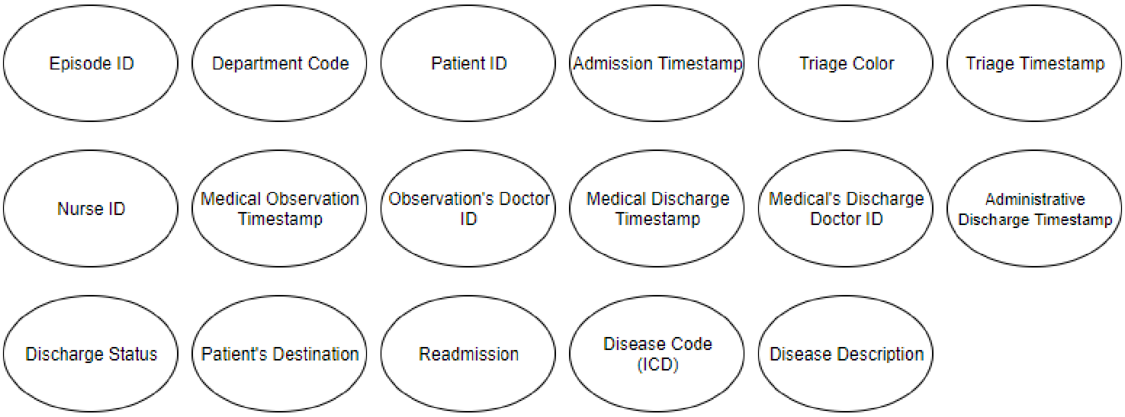


Figure 6 - Dataset's variables.

As illustrated in Figure 6 , there are several fields regarding identifications, such as the patient ID that allows the hospital to identify the patient and track his/her medical history within the hospital, the nurse ID that identifies the nurse who performed the triage (step 2 of the ED procedures as shown in Figure 4), the observer doctor ID that identifies the doctor responsible for the patient's medical observation (step 3 of the ED procedures), and the discharge doctor ID that identifies the discharge doctor (step 4 of the ED procedures).

To respect European privacy laws, based on the General Data Protection Regulation (GDPR), these fields, which allowed the identification of patients, nurses and doctors, were submitted to an anonymization process, so medical records can be used for analysis but can never reveal people's identifications in order to respect their privacy.

There are also several fields regarding the timestamps of the patient’s admission, triage, first medical observation and medical discharge as well as administrative discharge. This fields are essential and will allow to perform some analyses and get conclusions about some ED’s metrics like waiting times and LoS.

The triage colour field is also present, as well as the disease code according to the International Classification of Diseases (ICD) managed and published by the World Health Organization (WHO) [31], this allows to classify diseases by codes which facilitates analyses, and the diseases are given the code according to similarity, making possible to group diseases by categories. There is also a disease description field having the name of the disease corresponding to the ICD code.

3.2 Data Preparation

In order to allow the analysis of the data, and according to the working methodology presented in Chapter 3, and based on Figure 3 of the same chapter, after collecting the data they should go through a "cleaning" phase where incomplete or inconsistent data will be eliminated so as not to jeopardize the results of the analysis and this will be explained in detail on section 3.2.1.

To enrich the data and have more results to analyze and compare, some external factors will also be added to the dataset, like weather conditions, and big events in Lisbon, this process will also be described with more detail in this chapter on section 3.2.2.

3.2.1 Data Cleaning

The aim of this section is to lose inconsistent or incomplete data that can compromise the accuracy of the results of the analyses.

In order to do that, a python open-source data analysis and manipulation tool named Pandas will be used. The data that is considered corrupted is an episode (any row) that has any empty field. From this operation, the dataset went from 108295 to 90352 cases, which means that 16.6% of the original dataset had missing information.

The author noted some inconsistencies between the admission timestamps and the first medical observation timestamps, which led to negative waiting times and also between the admission timestamps and the administrative discharge timestamps, which was generating negative LoS, so the solution was to remove these corrupted data as well, which resulted in a reduction of the dataset to 87414 cases.

The next step was to convert the timestamps columns to the Pandas Datetime64 format, to make it possible to manipulate the timestamps, to do this the parse_dates function.

The author also dropped some of the attributes present on the original dataset as they are not relevant for this work, in Figure 7 are represented the remaining 6 attributes that will also be used to calculate new metrics that will be later used on Sections 3.3 and 3.4.

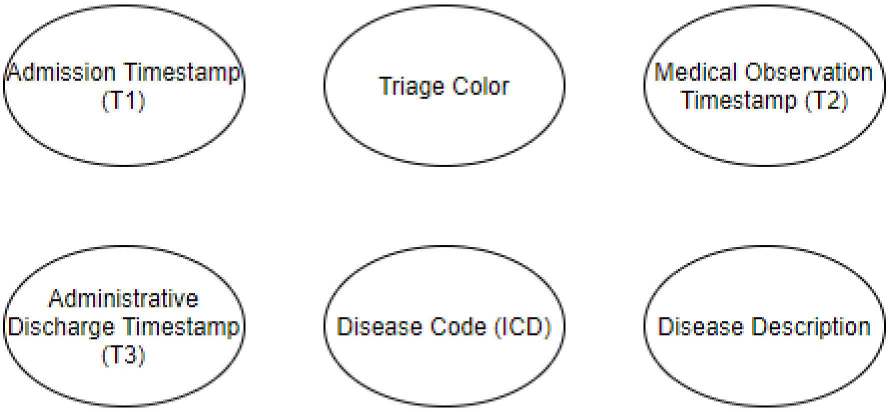

Admission Timestamp (T1)
Triage Color
Medical Observation Timestamp (T2)
Administrative Discharge Timestamp (T3)
Disease Code (ICD)
Disease Description

Figure 7 - Original dataset attributes after cleaning process.

3.2.2 Additional Data

After cleaning the original dataset, the author added new relevant data for the analyses, this additional data will be divided into 2 categories: the first one will be calculated data, adding two important times, the waiting times and the patient's LoS, which will be described on section 3.2.2.1 in detail, and the other one is external data, where the weather conditions will be added, as well as the occurrence of events and will be described in detail on section 3.2.2.2.

The objective of adding this new data is to enrich the results, which will allow them to take more conclusions about the influence of some factors on the hospital's metrics and management.

3.2.2.1 Calculated Data

The main goal in this section is to calculate the 2 "time metrics" that will be used for the Descriptive Analyses and the predictive analyses.

The first one is patient waiting times and consists in the time that the patient spends waiting to be evaluated by the doctor, and it is an important factor concerning patient satisfaction [32].

The waiting time can be obtained by the difference between the medical observation timestamp and the admission timestamp (T2-T1).

The minimum value and the maximum value obtained for the waiting times are displayed on Table 2.

Minimum waiting time	00:00:04
Maximum waiting time	21:39:41

Table 2 - Minimum and maximum waiting time values (HH:MM:SS).

Applying the same methodology used to obtain the waiting time, the author calculated the patient's LoS. As explained in Chapter 2.4, the LoS is the time spent since the patient checks in until he receives the discharge and goes home. To calculate the LoS the attributes used from the dataset were the administrative release and the admission timestamp (T3-T1).

The minimum and maximum values for the LoS are displayed on the following Table 3.

Minimum LoS	0 days 00:04:54
Maximum LoS	249 days 19:49:09

Table 3 - Minimum and maximum LoS (days HH:MM:SS).

The author also added a column that will give the weekday based on the admission timestamp, using the weekday() function provided by the date-time library, this function attributes the number 0 to Mondays and successively increases until Sunday (number 6) and a column also based on the admission timestamp with the respective season of the year (winter, spring, summer or autumn).

The last calculated data was the disease group. The author grouped the diseases based on their ICD codes [33] accordingly to the Table 4, in order to facilitate the analyses as there are too many different diseases.

ICD code	Disease group
001 – 139	Infectious and parasitic diseases
140 – 239	Neoplasms
240 – 279	Endocrin, Nutritional and metabolic Diseases, and immunity Disorders
280 – 289	Diseases of the blood and blood-forming organs
290 – 319	Mental Disorders
320 – 389	Diseases of the nervous system and sense organs
390 – 459	Diseases of the circulatory system
460 – 519	Diseases of the respiratory system
520 – 579	Diseases of digestive system
580 – 629	Diseases of the genitourinary system
630 – 679	Complications of pregnancy, childbirth and the puerperium
680 – 709	Diseases of the skin and subcutaneous system
710 – 739	Diseases of muscoloskeletal system and connective tissue
740 – 759	Congenital anomalies
760 – 779	Certain conditions originating in the perinatal period
780 – 799	Symptoms, signs and ill-defined conditions
800 – 999	Injury and poisoning
E and V codes	External causes of injury and supplemental classification

Table 4 - ICD codes and disease groups.

3.2.2.2 External Data

This section's purpose was to add information to the dataset from external fonts. The first information added was the rainy days in 2017, in order to do this the author gathered information about the weather conditions from a weather website (https://en.tutiempo.net/climate/2017/ws-85350.html), as the information was incomplete, the author resorted to another website (https://rp5.ru/) in order to complete the missing data. The rainy days are marked with "1" and the non-rainy days are marked with a "0". The total number of rainy days for 2017 was 40.

The author also added all the home games from the two main teams in Lisbon, Sporting Clube de Portugal and Sport Lisboa e Benfica. Sporting games are marked with an "S", Benfica games are marked with a "B", and the games between both teams are marked with "SB". The information about the games was retrieved from the website https://www.zerozero.pt/. There was a total of 51 days with games from these two teams in 2017.

The last addition to the dataset was the main concerts and festivals in Lisbon in 2017, to find this information the author resorted to two different websites, https://blitz.pt/ and https://sicnoticias.pt/cultura/2017-01-02-Os-concertos-a-ver-em-2017, marking with a "C" the days with the concerts. The whole days with shows for 2017 was 31.

The structure of the dataset created by the author containing information about external factors is shown in Figure 8.

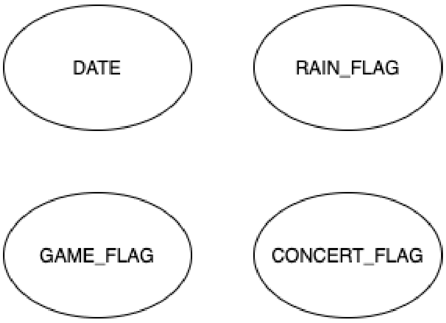


Figure 8 - External dataset attributes.

Finally, the author also loaded this external dataset to the python script, filled the NaN's present in the GAME_FLAG and CONCERT_FLAG with "No Event".

Then in order to merge both datasets, they had to have a standard column, so the author added a column to the original dataset also named DATE based on the ADMISSION TIMESTAMP column but without the time, only the date. That was done using the date function from the Datetime library. After this alteration, it was possible to use the Pandas library's merge function and join the two datasets. Figure 9 displays the final dataset's structure:

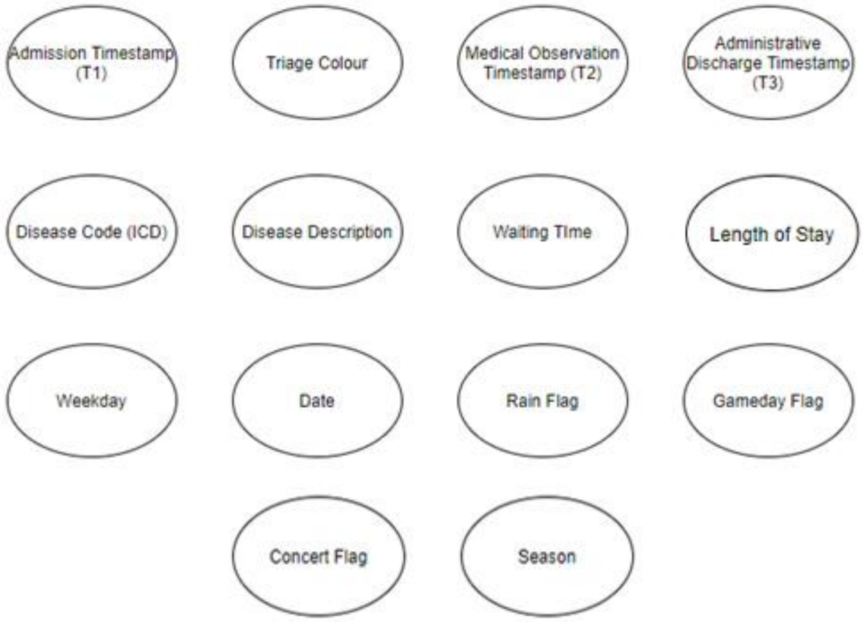


Figure 9 - Dataset after data preparation.

3.3 Descriptive Analyses.

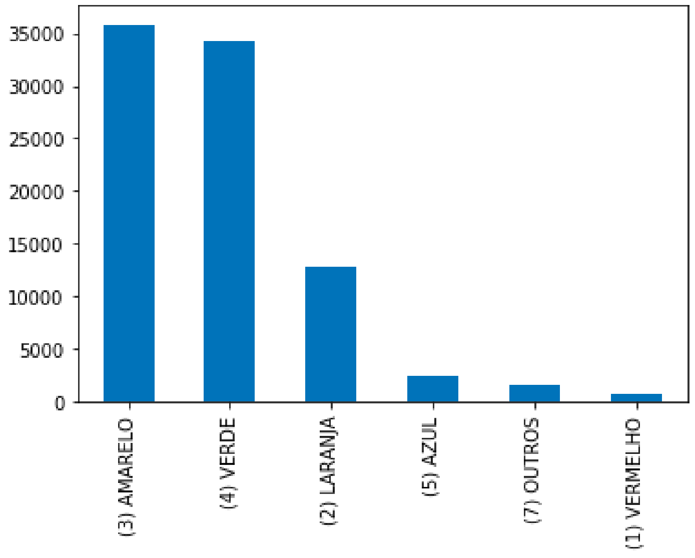
In this section, the author will perform a descriptive analyses on the data in order to extract useful information from the dataset to give an insight into hospital metrics and find patterns in order to draw conclusions with the goal of helping the decision-making proccess.

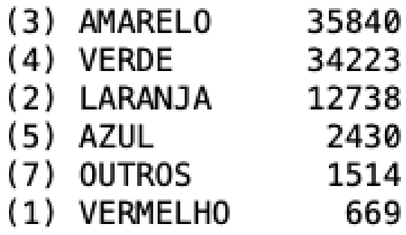
After the data collection (Subchapter 3.1) and the data preparation (Subchapter 3.2) the data is now ready to be analyzed. The descriptive analysis can be divided into two subchapters, in the first one (3.3.1) the author will do a general analysis of the data, i.e. analyze the distribution of diseases, the severity of diseases (through the colors of the bracelets assigned to the patients), the general waiting times and the general LOS by disease group.

In the second (3.3.2) the author will perform the analysis of waiting times and LoS taking into account the various external factors intorduced in subchapter 3.2 to see if they have an influence on these hospital metrics.

3.3.1 General Overview

The distribution of the triage colours throughout the year of 2017 can be observed on Figure 10:





```
(3) AMARELO     35840
(4) VERDE       34223
(2) LARANJA     12738
(5) AZUL         2430
(7) OUTROS       1514
(1) VERMELHO      669
```

Figure 10 - Triage Color distribution graph.

As shown in the figure above, the two most common colours are the yellow colour and the green colour, with the orange colour also having a good representation.

In order to compare the waiting times verified on the ED of this hospital with the recommended values from the MTP presented on Figure 5, the author calculated the mean average waiting time for each triage colours. The obtained values are displayed on Table 5:

Triage Color	Priority	Waiting Time
Red	1	00:20:30
Orange	2	00:36:14
Yellow	3	01:09:59
Green	4	01:34:26
Blue	5	02:05:11

Table 5 - Waiting times based on triage color.

By comparing the results in Table 5 for the waiting times with the recommended waiting times displayed in Figure 5, the following conclusions can be drawn:

The average waiting times for the Blue and Green colours are in-line with the recommended ones (recommended waiting time for Blue colour: 2h, recommended waiting time for the Green colour: 1h 30min).

The picture is different when it comes to the other 3 priorities, the average waiting time for the yellow colour is twice of the recommended and the average waiting time for the orange colour is three times the recommended one. The patients with the red colour are supposed to be immediately seen by a doctor, but they have a waiting time of 20 minutes.

After analyzing this results, it is safe to say that the consequences of this high waiting times on the top priority colours can have drastic effects on patients' treatments, in fact reducing this waiting times can save lives. The author will evaluate if external factors have any influence on the waiting times and compare results on section 3.3.2.

The author also evaluated the causes of the ED entries registered over the year of 2017, on Table 6 are displayed the 10 causes of entry into the ED with most cases registered.

Cause of entry into the ED	Number of cases
Lumbago	3733
Urinary Tract Infection	2990
Abdominal Pain	2956
Chest Pain	1785
Head Injury	1562
Renal Colic	1493
Acute Bronchitis	1387
Headache	1346
Enteritis, Colitis, Gastroenteritis of Presumed Infectious Origin	1271
Acute Tonsillitis	1227

Table 6 - Top 10 diseases registered in 2017 on Hospital Garcia da Horta.

As there are too many different diseases/symptoms in order to take them all into consideration, the diseases are divided into 18 separate categories, as mentioned on section 3.2.2.1, and the distribution of those categories can be observed on Figure 11.

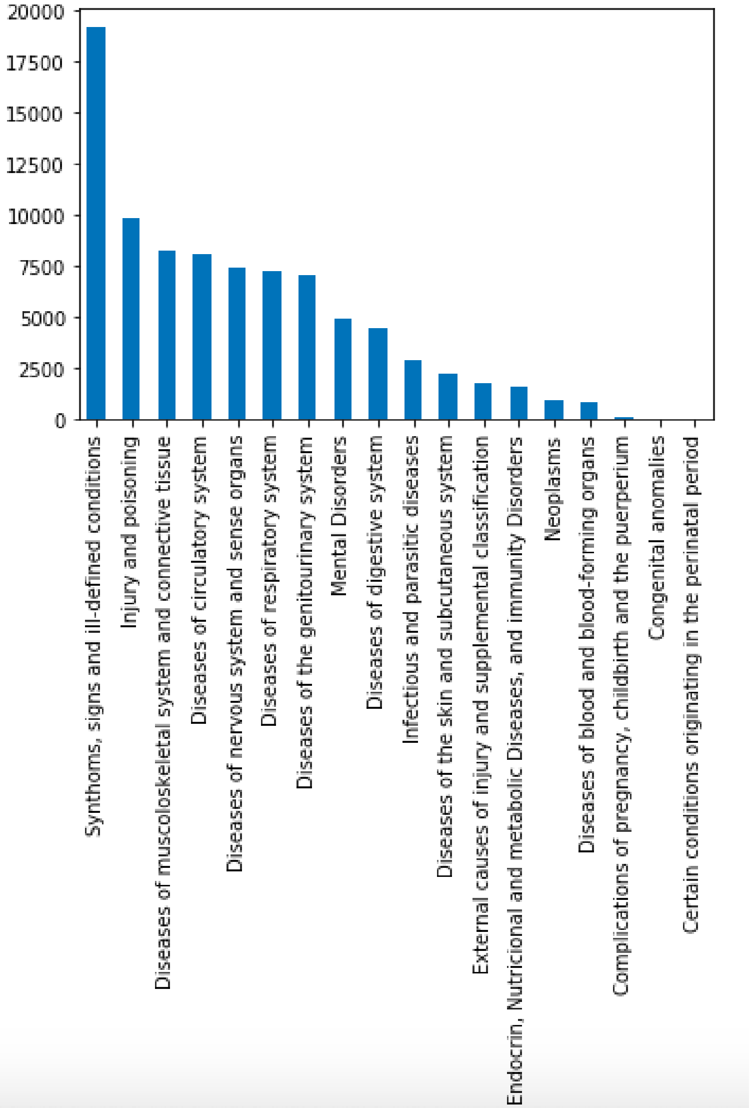


Figure 11 - Diseases groups distribution.

The group with more representation is the Symptoms, signs and ill-defined conditions (19202 cases), which includes Symptoms (ICD codes: 780–789), Nonspecific abnormal findings (ICD codes: 790–796) and ill-defined and unknown causes of morbidity and mortality (ICD codes: 797–799).

There are also 3 groups with almost no representation, Complications of pregnancy, childbirth and the puerperium (93 cases), Congenital anomalies (25) and Certain conditions originating in the perinatal period (6).

Another important metric to study is the LoS, to do that the author will study the LoS based on each disease group.

Disease group	**Average LoS**
Diseases of the blood and blood-forming organs	7 days 15:37:33
Neoplasms	5 days 06:16:15
Endocrin, Nutricional and metabolic Diseases, and immunity Disorders	5 days 18:43:52
Diseases of the genitourinary system	3 days 19:44:10
Diseases of the digestive system	3 days 09:35:00
Diseases of the respiratory system	3 days 04:09:09
Diseases of the circulatory system	3 days 00:21:36
Symptoms, signs and ill-defined conditions	2 days 22:59:39
Mental Disorders	2 days 18:16:30
External causes of injury and supplemental classification	2 days 15:23:41
Infectious and parasitic diseases	2 days 08:19:28
Injury and poisoning	2 days 07:16:25
Complications of pregnancy, childbirth and the puerperium	2 days 02:30:07
Diseases of the musculoskeletal system and connective tissue	1 days 18:13:35
Diseases of the skin and subcutaneous system	1 days 13:08:00
Diseases of the nervous system and sense organs	0 days 22:27:47
Congenital anomalies	0 days 07:55:01
Certain conditions originating in the perinatal period	0 days 07:20:57
General Average LoS	2 days 18:12:51

Table 7 -Average Length of Stay per disease group.

Observing the results obtained on Table 7, the LoS varies between 7h20mins and over 7 days, with the average being 2 days and 18 hours. Comparing the Table 7 with the results on Figure 11, it is visible that the more common diseases are the ones with a lower LoS, and the more severe diseases with more than 5 days of LoS are not as common.

3.3.2 Impact of External Conditions

On this section, the author will evaluate whether or not conditions like the time of the year, day of the week and events have an influence on healthcare as well as finding patterns that can help to improve hospital's decision-making.

3.3.2.1 Seasons of the year

In order to compare the four seasons of the year, the author divided the dataset into four other datasets, one for each season of the year (Spring, Summer, Autumn and Winter).

Starting with the Spring dataset, during this season there were a total of 22019 cases (approximately 25.2% of total cases), the waiting times during Spring are shown on Table 8.

Triage Color	Priority	Waiting Time
Red	1	00:18:22
Orange	2	00:36:34
Yellow	3	01:05:09
Green	4	01:29:22
Blue	5	02:02:31

Table 8 - Waiting times during Spring.

Comparing the results obtained on Table 8 with the general ones displayed on Table 5, it can be concluded that the waiting times in Spring were slightly shorter, but not by a relevant margin.

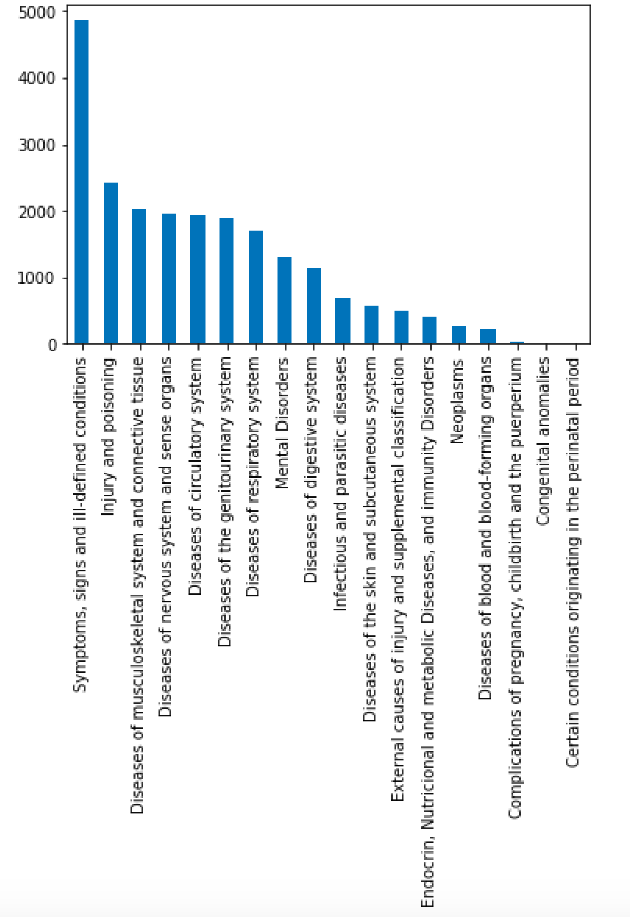


Figure 12 - Disease groups distribution on Spring.

Comparing the distribution of the disease groups in Figure 12 with the results obtained in Figure 11, the conclusions are that the incidence of the diseases remains in line with the overall results.

As for the LoS, it is higher in the spring as can be seen by the increase in the average LOS from 2 days 18:12:51 to 3 days 01:44:25. The results by disease group are presented in Table 9 and range from over 3 hours to over 8 days.

Disease group	Average LoS
Diseases of the blood and blood-forming organs	8 days 03:48:12
Endocrin, Nutricional and metabolic Diseases, and immunity Disorders	5 days 19:23:21
Neoplasms	4 days 23:29:42
Diseases of the genitourinary system	4 days 06:58:28
Diseases of the digestive system	3 days 21:11:57
Diseases of the respiratory system	3 days 13:11:34
Symptoms, signs and ill-defined conditions	3 days 06:00:50
Mental Disorders	3 days 06:06:16
Diseases of the circulatory system	3 days 04:31:26
Injury and poisoning	2 days 18:47:10
Infectious and parasitic diseases	2 days 18:28:02
Diseases of the skin and subcutaneous system	2 days 07:09:32
Diseases of the musculoskeletal system and connective tissue	2 days 02:48:39
External causes of injury and supplemental classification	2 days 00:43:57
Diseases of the nervous system and sense organs	0 days 23:49:06
Certain conditions originating in the perinatal period	0 days 12:05:34
Congenital anomalies	0 days 05:35:41
Complications of pregnancy, childbirth and the puerperium	0 days 03:47:59
General Average LoS	3 days 01:44:25

Table 9 - Average LoS per disease group on Spring.

During the Summer there were a total of 22427 cases (representing 25.7% of the cases) and registered the average waiting times on Table 10:

Triage Color	Priority	Waiting Time
Red	1	00:21:44
Orange	2	00:34:20
Yellow	3	01:04:46
Green	4	01:30:50
Blue	5	02:08:44

Table 10 - Average waiting time per triage color during summer.

The waiting times are similar to the ones verified on Table 5 and with the ones from Spring (Table 8) which makes sense because there were an identical number of cases, the disease group distribution (Figure 13) is also similar to the general one (Figure 11) and to the Spring's results (Figure 12).

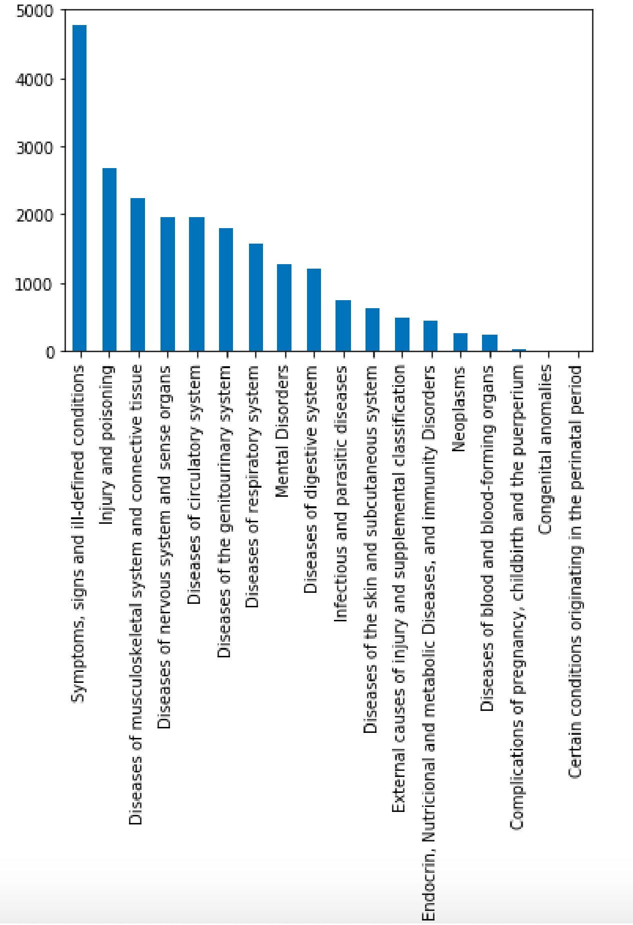


Figure 13 - Disease groups distribution on Summer.

The average LoS per disease group on Summer (Table 11) is in general similar to the general ones on Table 7 with some disease groups having a few more hours than the general results, it is also lower than the LoS registered during Spring. The summer's average LoS being 2 days 20:29:17 vs the overall average LoS of 2 days 18:12:51.

Disease group	**Average LoS**
Diseases of the blood and blood-forming organs	7 days 03:45:55
Neoplasms	5 days 23:16:39
Endocrin, Nutricional and metabolic Diseases, and immunity Disorders	5 days 11:43:03
Diseases of the respiratory system	3 days 21:58:39
Complications of pregnancy, childbirth and the puerperium	3 days 14:34:20
Diseases of the circulatory system	3 days 13:46:29
Diseases of the genitourinary system	3 days 08:34:37
Diseases of the digestive system	3 days 06:42:02
Symptoms, signs and ill-defined conditions	3 days 04:42:13
External causes of injury and supplemental classification	3 days 00:40:57
Mental Disorders	2 days 14:38:50
Infectious and parasitic diseases	2 days 02:41:49
Injury and poisoning	2 days 08:00:37
Diseases of the musculoskeletal system and connective tissue	1 days 17:25:59
Diseases of the skin and subcutaneous system	1 days 05:29:41
Diseases of the nervous system and sense organs	1 days 01:59:12
Congenital anomalies	0 days 08:53:19
Certain conditions originating in the perinatal period	0 days 01:48:36
General Average LoS	2 days 20:29:17

Table 11 – Average LoS per disease group on Summer.

During the Autumn, there were 21852 cases registered (equivalent to 25% of the total cases). The waiting times during this period of the year can be seen on Table 12 and comparing them to the general ones on Table 5, the green and yellow colours have a slightly higher waiting times and the other three priorities have the same waiting times as the general average.

Triage Color	Priority	Waiting Time
Red	1	00:20:40
Orange	2	00:36:05
Yellow	3	01:15:00
Green	4	01:40:23
Blue	5	02:05:01

Table 12 - Average waiting times on Autumn.

The disease groups with more cases during this season remain the same as the general, when comparing Figure 14 to Figure 11, the only difference being that no cases were registered for the "Certain conditions originating in the perinatal period" group.

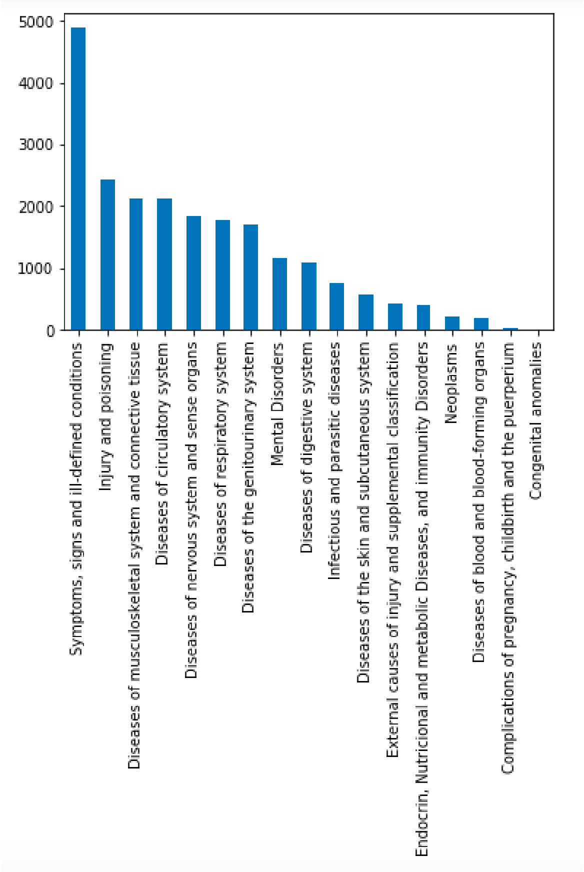


Figure 14 - Disease groups distribution during Autumn.

The LoS during the fall (Table 13) was considerably shorter in relation to the overall LoS from Table 7 being on average about a day shorter. The lowest LoS registered during fall was over 4 hours for the "Complications of pregnancy, childbirth and the puerperium" group and the longest approximately 5 days and 9 hours for the "Diseases of the blood and blood-forming organs" group.

Disease group	Average LoS
Diseases of the blood and blood-forming organs	5 days 08:54:10
Endocrin, Nutritional and metabolic Diseases, and immunity Disorders	4 days 14:37:31
Neoplasms	3 days 13:14:02
External causes of injury and supplemental classification	2 days 09:18:01
Diseases of the genitourinary system	2 days 07:13:08
Diseases of the respiratory system	2 days 05:33:52
Diseases of the digestive system	2 days 04:48:52
Mental Disorders	2 days 00:52:14
Diseases of the circulatory system	1 days 22:45:22
Symptoms, signs and ill-defined conditions	1 days 21:41:48
Infectious and parasitic diseases	1 days 15:44:47
Injury and poisoning	1 days 13:46:01
Diseases of the skin and subcutaneous system	1 days 10:17:11
Diseases of the musculoskeletal system and connective tissue	1 days 03:41:47
Diseases of the nervous system and sense organs	0 days 15:25:38
Congenital anomalies	0 days 09:17:53
Complications of pregnancy, childbirth and the puerperium	0 days 04:04:14
Certain conditions originating in the perinatal period	NaN
General Average LoS	1 days 20:28:32

Table 13 - Average LoS per disease group on Autumn.

The last season is Winter, having a total of 21116 cases (about 24% of total cases), the waiting times during this time of the year are exhibited on Table 14.

Triage Color	Priority	Waiting Time
Red	1	00:21:11
Orange	2	00:37:47
Yellow	3	01:15:06
Green	4	01:37:32
Blue	5	02:04:37

Table 14 - Average waiting times on winter.

As in other seasons, the values for the waiting times in winter are quite similar to the general ones. When it comes to disease distribution in winter, the “Diseases of the respiratory system” and “Diseases of circulatory system” which have more representation during this season of the year.

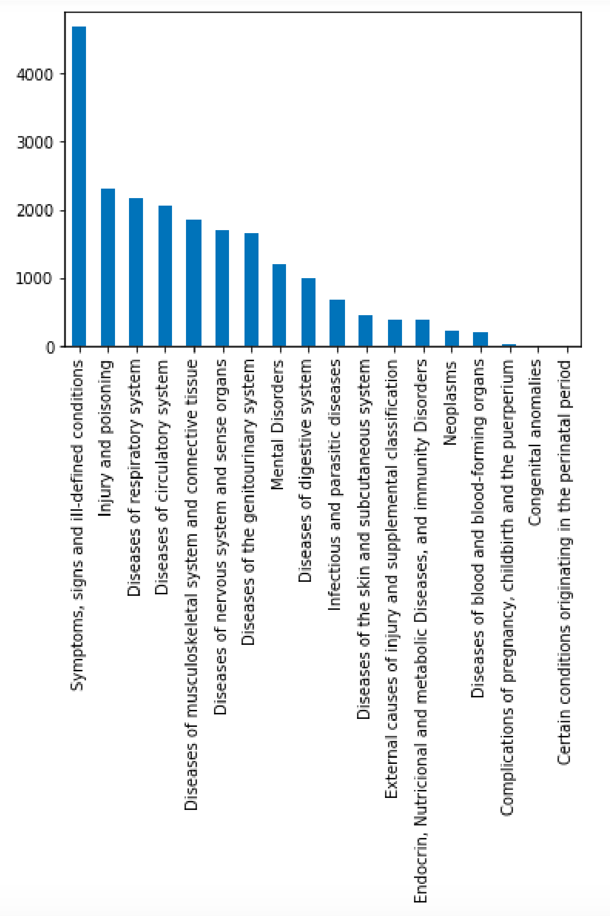


Figure 15 - Disease groups distribution on Winter.

Unlike autumn, winter recorded in general higher LoS, with values ranging between 6 hours and over 9 days and an average of 3 days 06:26:49, comparing to the general average of 2 days 18:12:51 registered in Table 7. The LoS per disease group can be seen in Table 15.

Disease group	Average LoS
Diseases of the blood and blood-forming organs	9 days 20:31:34
Endocrin, Nutricional and metabolic Diseases, and immunity Disorders	7 days 07:09:22
Neoplasms	6 days 09:35:21
Diseases of the genitourinary system	5 days 08:49:21
Diseases of the digestive system	4 days 07:40:46
Complications of pregnancy, childbirth and the puerperium	3 days 23:21:37
Symptoms, signs and ill-defined conditions	3 days 12:16:07
Diseases of the circulatory system	3 days 10:04:15
External causes of injury and supplemental classification	3 days 04:59:48
Diseases of the respiratory system	3 days 02:35:06
Mental Disorders	3 days 02:17:42
Infectious and parasitic diseases	2 days 22:43:03
Injury and poisoning	2 days 12:42:31
Diseases of the musculoskeletal system and connective tissue	2 days 02:25:27
Diseases of the skin and subcutaneous system	1 days 04:49:11
Diseases of the nervous system and sense organs	1 days 00:27:44
Congenital anomalies	0 days 08:34:49
Certain conditions originating in the perinatal period	0 days 06:01:59
General Average LoS	3 days 06:26:49

Table 15 - Average LoS per disease group on winter.

After analyzing all seasons of the year, the conclusions drawn are that the number of cases per season of the year is similar, this distribution can be seen on Figure 16, the waiting times are similar for all seasons. The diseases groups also have identical distributions throughout the year, except on winter where the respiratory system diseases have a more extensive representation than on the rest of the year.

The LoS, on the other hand, varies throughout the seasons, reaching its highest on winter and spring (over 3 days) and lowest on autumn (over a day) according to Figure 17.

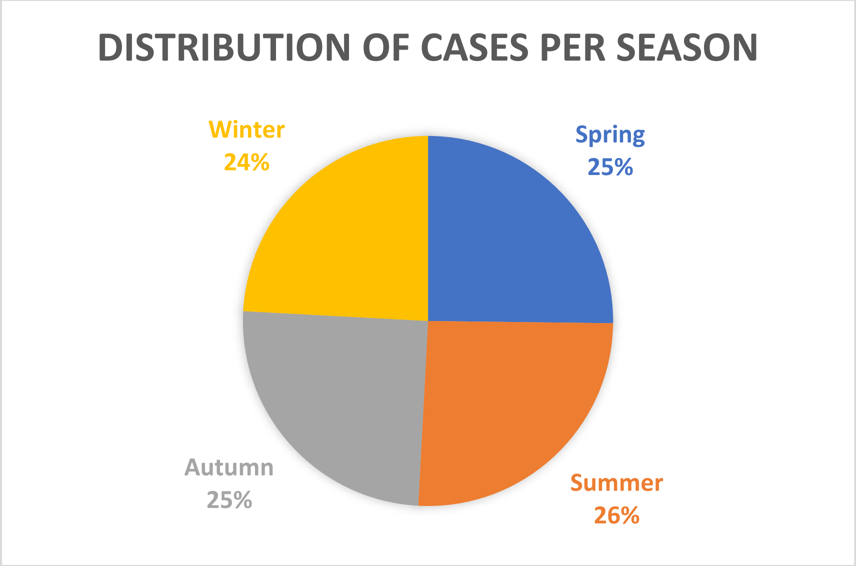


Figure 16 - Distribution of cases per season of the year.

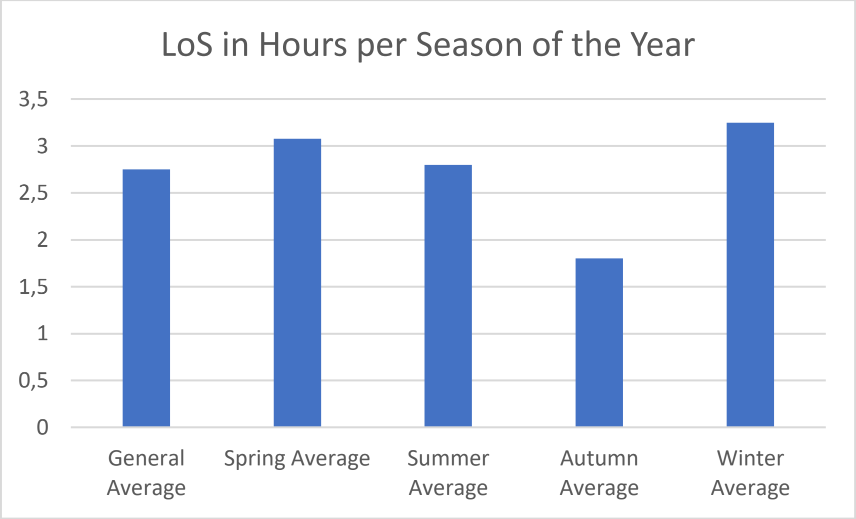


Figure 17 - Length of Stay in hours per Season of the year.

3.3.2.2 Weekdays vs Weekends

It is also important to analyze if the weekends have any influence on hospital functioning when compared to the weekdays, in order to do that, the first metric that will be examined are the waiting times.

Triage Color	Priority	Waiting Time
Red	1	00:20:00
Orange	2	00:36:16
Yellow	3	01:09:32
Green	4	01:34:32
Blue	5	02:02:48

Table 16 – Average waiting times during weekdays.

Triage Color	Priority	Waiting Time
Red	1	00:21:45
Orange	2	00:36:08
Yellow	3	01:11:08
Green	4	01:34:10
Blue	5	02:15:04

Table 17 – Average waiting times during weekends.

As seen in Tables 16 and 17, the waiting times for weekdays and weekends are similar with the biggest difference of 13 minutes for the blue colour and thus revealing that weekends are not a factor that has direct influence over the hospital's waiting times.

The other metric taken into consideration was the patient's LoS, and similarly to what was done for the waiting times the author calculated the LoS for the weekdays and the weekends and compared the results.

Disease group	**Average LoS**
Diseases of the blood and blood-forming organs	6 days 22:25:32
Endocrin, Nutricional and metabolic Diseases, and immunity Disorders	5 days 14:40:07
Neoplasms	5 days 04:06:49
Diseases of the genitourinary system	3 days 16:58:37
Diseases of the digestive system	3 days 04:28:30
Diseases of the respiratory system	2 days 23:56:59
Symptoms, signs and ill-defined conditions	2 days 20:39:17
Diseases of the circulatory system	2 days 16:47:57
Mental Disorders	2 days 13:05:44
External causes of injury and supplemental classification	2 days 10:52:14
Complications of pregnancy, childbirth and the puerperium	2 days 06:18:36
Infectious and parasitic diseases	2 days 04:49:43
Injury and poisoning	2 days 03:58:12
Diseases of the skin and subcutaneous system	1 days 16:50:18
Diseases of the musculoskeletal system and connective tissue	1 days 13:55:51
Diseases of the nervous system and sense organs	0 days 19:34:37
Congenital anomalies	0 days 08:35:22
Certain conditions originating in the perinatal period	0 days 07:20:57
General Average LoS	2 days 14:06:46

Table 18 - Average LoS during weekdays.

Disease group	Average LoS
Diseases of the blood and blood-forming organs	9 days 21:32:33
Endocrin, Nutricional and metabolic Diseases, and immunity Disorders	6 days 05:49:57
Neoplasms	5 days 11:42:35
Diseases of the genitourinary system	4 days 03:10:41
Diseases of the digestive system	3 days 22:56:24
Diseases of the circulatory system	3 days 21:07:41
Diseases of the respiratory system	3 days 14:51:14
Mental Disorders	3 days 09:38:50
Symptoms, signs and ill-defined conditions	3 days 05:19:01
External causes of injury and supplemental classification	3 days 02:07:14
Infectious and parasitic diseases	2 days 17:28:03
Injury and poisoning	2 days 15:41:06
Diseases of the musculoskeletal system and connective tissue	2 days 05:13:18
Complications of pregnancy, childbirth and the puerperium	1 days 13:26:42
Diseases of the nervous system and sense organs	1 days 10:45:14
Diseases of the skin and subcutaneous system	1 days 02:27:58
Congenital anomalies	0 days 05:47:16
Certain conditions originating in the perinatal period	NaN
General Average LoS	3 days 05:33:29

Table 19 - Average LoS during weekends.

Observing the obtained results in Tables 18 and 19, unlike the waiting times the LoS registered a significant difference with the average LoS during weekends being almost a day longer than during the weekdays, and especially for the "Diseases of blood and blood-forming organs" category which has an average LoS 3 days longer than the one registered during the weekdays. This can be a result of a more delayed treatment during the weekends, which will extend the LoS.

3.3.2.3 Precipitation Days

In this subchapter, the author will be comparing the results obtained in days with precipitation with the ones from non-rainy days as well as the overall results obtained on the section 3.3.1 both for the waiting times and LoS per disease group.

Triage Color	Priority	Waiting Time
Red	1	00:14:47
Orange	2	00:37:13
Yellow	3	01:15:13
Green	4	01:42:21
Blue	5	02:06:22

Table 20 - Waiting times on days with precipitation.

Triage Color	Priority	Waiting Time
Red	1	00:21:05
Orange	2	00:36:06
Yellow	3	01:09:17
Green	4	01:33:28
Blue	5	02:05:03

Table 21 - Waiting times for days without precipitation.

As evidenced by Tables 20 and 21, the waiting times are shorter by minutes on days without precipitation than on days with rainfall, except for the red colour that is shorter on rainy days for 7 minutes.

In regard to the patient's LoS the results are presented on Tables 22 and 23, and they show that the precipitation has a significant influence over the LoS, with the average LoS being over 9 hours long on days with the occurrence of precipitation.

Disease group	**Average LoS**
Diseases of the blood and blood-forming organs	7 days 16:16:27
Endocrin, Nutricional and metabolic Diseases, and immunity Disorders	6 days 21:08:45
Neoplasms	5 days 15:45:59
Diseases of the genitourinary system	5 days 04:20:03
Diseases of the digestive system	4 days 02:46:13
Mental Disorders	3 days 05:15:21
Diseases of the respiratory system	3 days 10:34:02
Diseases of the circulatory system	3 days 03:30:06
Symptoms, signs and ill-defined conditions	3 days 02:01:46
Infectious and parasitic diseases	2 days 22:03:41
Injury and poisoning	2 days 09:14:06
External causes of injury and supplemental classification	2 days 00:31:00
Diseases of the musculoskeletal system and connective tissue	1 days 22:04:10
Diseases of the skin and subcutaneous system	1 days 17:44:33
Diseases of the nervous system and sense organs	1 days 11:14:37
Congenital anomalies	0 days 06:11:23
Complications of pregnancy, childbirth and the puerperium	0 days 05:07:38
Certain conditions originating in the perinatal period	0 days 02:31:41
General Average LoS	3 days 02:41:37

Table 22 - Patient's LoS on days with precipitation.

Disease group	**Average LoS**
Diseases of the blood and blood-forming organs	7 days 15:32:47
Endocrin, Nutricional and metabolic Diseases, and immunity Disorders	5 days 15:28:14
Neoplasms	5 days 05:06:54
Diseases of the genitourinary system	3 days 15:42:28
Diseases of the respiratory system	3 days 03:10:25
Diseases of the digestive system	3 days 07:24:20
Diseases of the circulatory system	2 days 23:56:32
Symptoms, signs and ill-defined conditions	2 days 22:37:09
External causes of injury and supplemental classification	2 days 17:00:26
Mental Disorders	2 days 16:48:05
Complications of pregnancy, childbirth and the puerperium	2 days 07:21:48
Injury and poisoning	2 days 07:02:31
Infectious and parasitic diseases	2 days 06:31:35
Diseases of the musculoskeletal system and connective tissue	1 days 17:45:01
Diseases of the skin and subcutaneous system	1 days 12:37:31
Diseases of the nervous system and sense organs	0 days 20:45:49
Certain conditions originating in the perinatal period	0 days 08:18:48
Congenital anomalies	0 days 07:59:20
General Average LoS	2 days 17:08:03

Table 23 - Patient's LoS on days without precipitation.

3.3.2.4 Events

This section is where the author assesses whether major concerts in Lisbon and the football matches of the two "big" Lisbon teams have an influence on waiting times and LoS.

Starting with the waiting times for the days with concerts, the results obtained are presented in Table 24.

Triage Color	Priority	Waiting Time
Red	1	00:23:17
Orange	2	00:39:11
Yellow	3	01:13:14
Green	4	01:36:53
Blue	5	02:06:15

Table 24 - Waiting times for concert days.

Next the author calculated the waiting times for the gamedays and the results can be seen on Table 25.

Triage Color	Priority	Waiting Time
Red	1	00:16:03
Orange	2	00:35:52
Yellow	3	01:12:06
Green	4	01:33:05
Blue	5	02:02:44

Table 25 - Waiting times for gamedays.

In order to be able to compare the results obtained on Table 24 and Table 25, the author also calculated the waiting times for days without events and the results are shown on Table 26.

Triage Color	Priority	Waiting Time
Red	1	00:20:59
Orange	2	00:35:57
Yellow	3	01:09:13
Green	4	01:34:34
Blue	5	02:05:35

Table 26 - Waiting times for events-free days.

Comparing the three Tables (24, 25 and 26) the conclusions that can be drawn are that the events do not have a significant impact on the waiting times with the waiting times for gamedays being almost identical to the ones without events and the concert days having a slightly higher (1-4 mins higher) than the ones registered for days with no events.

Then, the author analyzed the patient's LoS starting with the days with concerts, the results of which are shown in Table 27.

Disease group	**Average LoS**
Endocrin, Nutritional and metabolic Diseases, and immunity Disorders	11 days 01:46:11
Diseases of the blood and blood-forming organs	10 days 22:34:59
Neoplasms	9 days 09:28:20
Diseases of the genitourinary system	7 days 12:18:26
Diseases of the circulatory system	5 days 15:20:38
Diseases of the digestive system	5 days 09:55:14
Symptoms , signs and ill-defined conditions	5 days 01:42:57
Diseases of the respiratory system	4 days 23:25:04
External causes of injury and supplemental classification	4 days 22:36:03
Mental Disorders	4 days 14:45:60
Infectious and parasitic diseases	4 days 13:10:11
Injury and poisoning	4 days 05:56:07
Diseases of musculoskeletal system and connective tissue	3 days 10:55:37
Diseases of the skin and subcutaneous system	2 days 11:46:23
Diseases of the nervous system and sense organs	1 days 14:50:39
Congenital anomalies	0 days 08:31:18
Complications of pregnancy, childbirth and the puerperium	0 days 05:35:53
Certain conditions originating in the perinatal period	0 days 03:04:27
General Average LoS	4 days 21:08:35

Table 27 - Average LoS on concert days.

Then the author obtained the results for the gamedays in Table 28 and the results for the days with no events in Table 29.

Disease group	Average LoS
Diseases of blood and blood-forming organs	7 days 04:11:43
Endocrin, Nutritional and metabolic Diseases, and immunity Disorders	3 days 12:15:55
Diseases of the digestive system	3 days 03:17:17
Diseases of the genitourinary system	2 days 21:34:08
Mental Disorders	2 days 21:07:35
Diseases of the circulatory system	2 days 20:13:26
External causes of injury and supplemental classification	2 days 16:14:30
Injury and poisoning	2 days 10:01:20
Diseases of the respiratory system	2 days 09:51:49
Symptoms , signs and ill-defined conditions	2 days 07:35:35
Neoplasms	2 days 06:00:22
Infectious and parasitic diseases	1 days 17:41:24
Diseases of the skin and subcutaneous system	1 days 03:08:22
Diseases of musculoskeletal system and connective tissue	1 days 02:19:39
Congenital anomalies	0 days 18:23:50
Diseases of the nervous system and sense organs	0 days 15:21:28
Complications of pregnancy, childbirth and the puerperium	0 days 03:56:10
Certain conditions originating in the perinatal period	NaN
General Average LoS	2 days 07:12:32

Table 28 – Average LoS for gamedays.

Disease group	Average LoS
Diseases of the blood and blood-forming organs	7 days 06:36:02
Endocrin, Nutricional and metabolic Diseases, and immunity Disorders	5 days 12:34:50
Neoplasms	5 days 04:19:05
Diseases of the genitourinary system	3 days 13:28:07
Diseases of the digestive system	3 days 04:48:25
Diseases of the respiratory system	3 days 02:14:47
Symptoms, signs and ill-defined conditions	2 days 19:27:40
Diseases of the circulatory system	2 days 17:09:21
Complications of pregnancy, childbirth and the puerperium	2 days 12:44:35
Mental Disorders	2 days 12:32:20
Infectious and parasitic diseases	2 days 05:20:29
External causes of injury and supplemental classification	2 days 08:17:56
Injury and poisoning	2 days 01:39:02
Diseases of the musculoskeletal system and connective tissue	1 days 16:18:56
Diseases of the skin and subcutaneous system	1 days 11:59:50
Diseases of the nervous system and sense organs	0 days 21:35:18
Certain conditions originating in the perinatal period	0 days 09:29:12
Congenital anomalies	0 days 07:19:54
General Average LoS	2 days 14:03:43

Table 29 - Average LoS on days without events.

Looking at the results on Tables 27, 28 and 29, the lower LoS was registered on days where the two big teams of Lisbon played (2 days and 7h) being around 7 hours shorter than the LoS for the days without events (2 days and 14h). On the other hand, the days with concerts registered the highest LoS (4 days and 21h) with around double the time of the days without events.

3.4 Predictive Model

Considering the flowchart represented in Figure 3, the author then proceeded to the elaboration of a prototype of a predictive model for the patient's LoS.

The idea of this prototype is described in Figure 18, where it can be observed that the nurse who performs the triage to the patient, will enter some information into the system, and the predictive model based on algorithms will predict the LoS, informing both the hospital and the patient of an estimate of the time the patient should spend in the facility in need of hospital care.

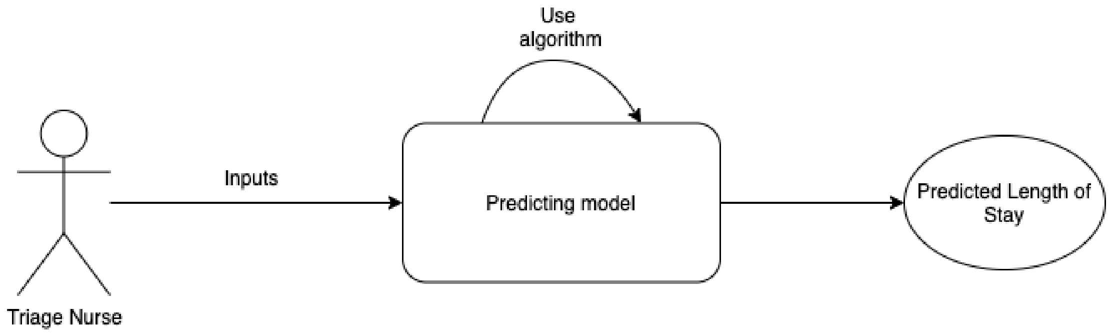


Figure 18 - Prototype schema.

Taking into account the dataset used by the author to perform the descriptive analysis (Figure 9), many of the attributes of this dataset are now dispensable for predictive analysis, so using the drop function of the Pandas library, the author removed from the dataset the columns will not be used, leaving only those necessary for this phase which are represented in Figure 19.

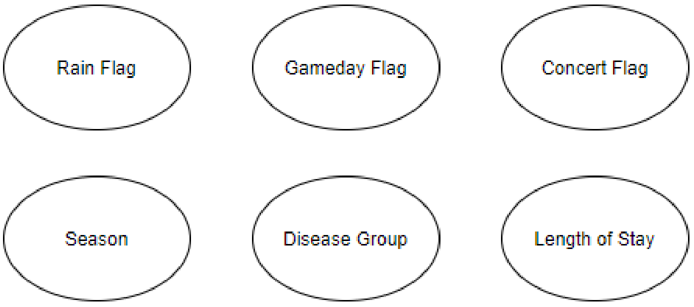


Figure 19 - Dataset used for predictive analysis.

Before being able to test the algorithms to evaluate which has the best performance and apply it to the model, the author had to make the data discrete, i.e. make all the variables "countable". Taking into account the dataset on Figure 19, the only discrete variable was the rain flag, the values of the remaining variables were changed according to the Table 30.

Variable	Old Values	Discrete Values
Gameday Flag	S,B, SB	1
	No Event	0
Concert Flag	C	1
	No Event	0
Season	Winter	1
	Spring	2
	Summer	3
	Autumn	4
Disease Group	Infectious and parasitic diseases	1
	Neoplasms	2
	Endocrin, nutricional and metabolic diseases, and immunity disorders	3
	Diseases of the blood and blood-forming organs	4
	Mental disorders	5
	Diseases of the nervous system and sense organs	6
	Diseases of the circulatory system	7
	Diseases of the respiratory system	8

Disease Group	Diseases of the digestive system	9
	Diseases of the genitourinary system	10
	Complications of pregnancy, childbirth and the puerperium	11
	Diseases of the skin and subcutaneous system	12
	Diseases of the musculoskeletal system and connective tissue	13
	Congenital anomalies	14
	Certain conditions originating in the perinatal period	15
	Symptoms, signs and ill-defined conditions	16
	Injury and poisoning	17
	External causes of injury and supplemental classification	18
Length of Stay	Short (Less than a day)	1
	Medium (Between a day and a week)	2
	Long (Between a week and a month)	3
	Very Long (More than a month)	4

Table 30 - Conversion to discrete variables.

Then, the author applied the two algorithms, NB and RF, to the dataset to see which one performed better.

The NB algorithm is a probabilistic classifier based on the Bayes' theorem (Equation 1). This algorithm calculates the probability of an event based on events that already occurred, applied to this particular dataset, the LoS will be calculated given the other four events, the existence of precipitation, football games or concerts and the disease group.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Equation 1 - Bayes' theorem.

The other algorithm that will be tested by the author is the RF and it is another popular machine learning technique used in regression and classification, it is based on decision trees, creating a number of decision trees using a bootstrap dataset built using randomly chosen samples from the original dataset, and for each node of each tree, it randomly chooses a predictor variable, which makes the generated trees different from each other, avoiding the overfitting problem that the Decision Tree algorithm has by using a unique tree and improving the accuracy of the RF model [34]. Then in order to give the final output, it calculates the mean or majority of the outputs from every tree of the "forest", this process is known as bagging.

Simplifying, the algorithm creates a dataset with randomly chosen samples from the original dataset (which may include duplicates) and then creates a chosen number of decision trees whose nodes are the dataset variables also randomly chosen, then each tree gives its output and the final output is the one most trees have chosen.

These two algorithms had a similar performance in several studies carried out by other authors as for example in the study to analyze and predict the results of cricket games [35], so the author decided to apply both algorithms in the dataset under study to see which of the two had the best performance in order to choose the most appropriate to implement in the model.

The author then using the the train-test split module from the sklearn library divided the dataset into a training dataset (70%) and a test dataset (30%) and test it with both algorithms to compare their scores.

For the RF algorithm the author used the Rain Forest module present in the sklearn library, and assigned the model 500 decision trees.

For the NB algorithm, the author also used the sklearn library using the NB module where there are 3 different models, BernoulliNB, indicated for binary data, GaussianNB appropriate for continuous decimal data and with normal distribution, and MultinomialNB which will be the one used by the author is ideal for discrete data such as the LoS groups (in between 1 and 4) used.

As each time you run the simulation the training and testing datasets are different because they are chosen randomly, the accuracy of the model shifts slightly between each simulation, so the author decided to do 5 simulations and measure the accuracy of both algorithms through the use of the score method also from the sklearn library that uses the test dataset inputs and compares the results obtained by the model with the actual results in order to obtain the accuracy of the model, to see which of the 2 algorithms has better performance in order to choose one of them to implement and the results can be seen in Figure 20.

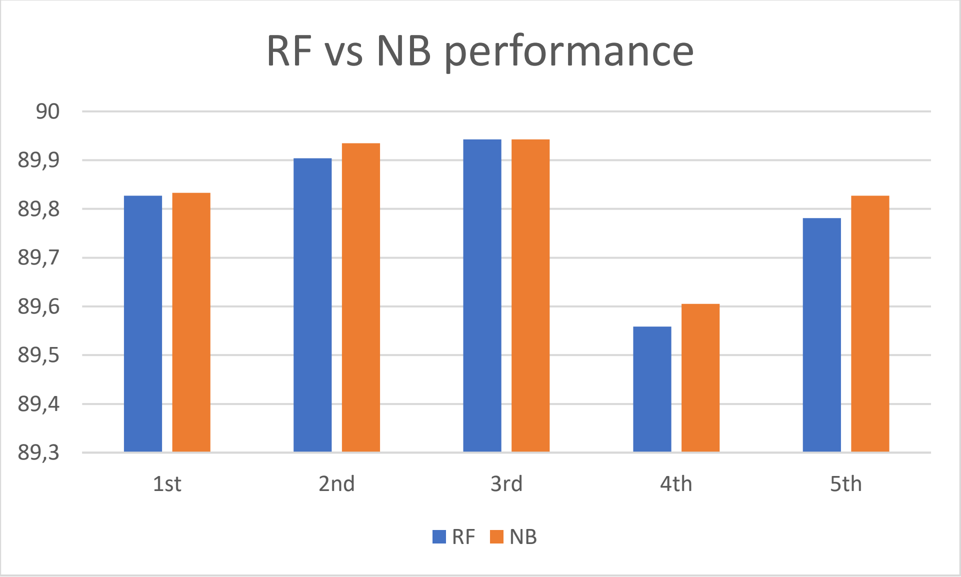
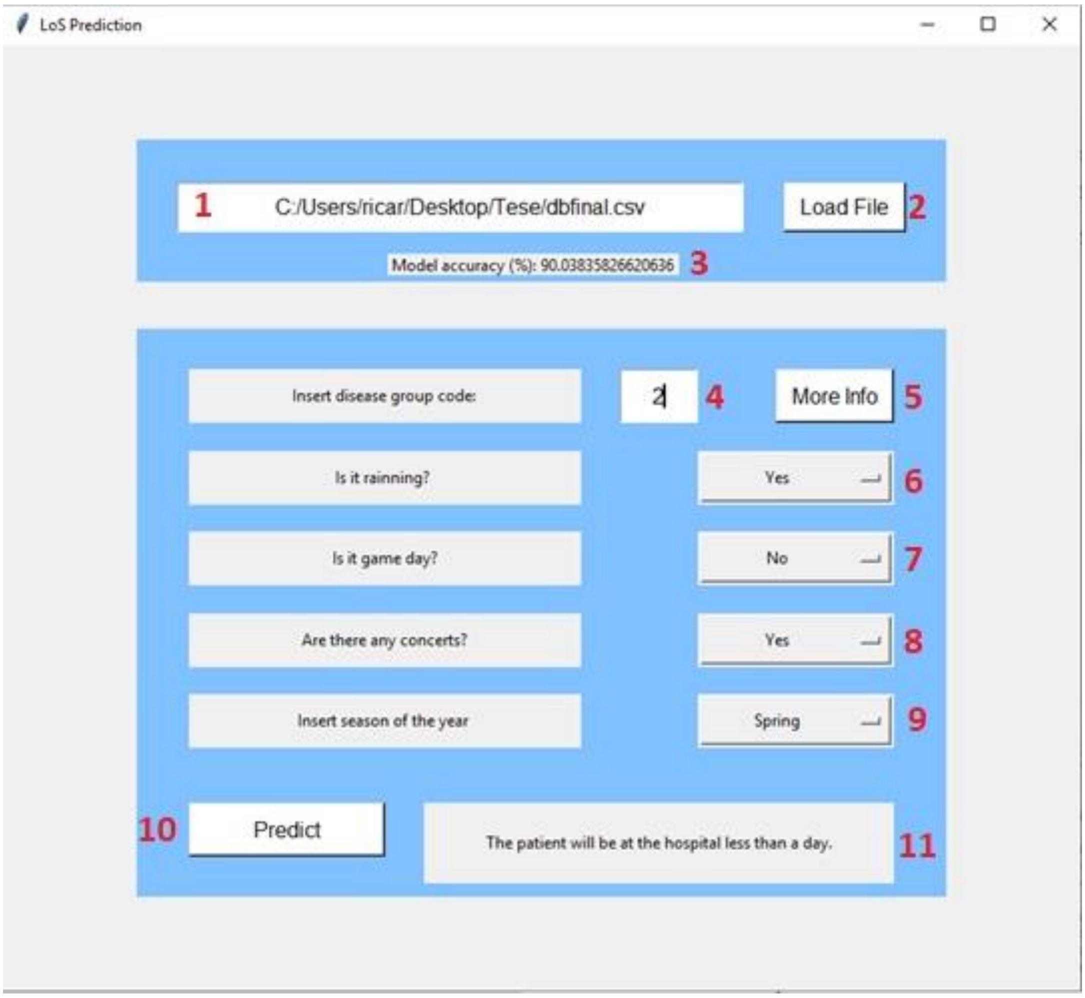


Figure 20 - Algorithms' accuracy (in %) by simulation.

Despite both algorithms having a good performance with scores between 89 and 90%, and NB had a slightly better performance in some of the simulations so it was the algorithm chosen to elaborate the predictive model.

After choosing the algorithm, the author decided to create a simple and easy to use user interface using PyCharm as the IDE and the tkinter library for the graphical components. Once the application is launched, the screen the user will see is the one in Figure 21.



C:/Users/ricar/Desktop/Tese/dbfinal.csv

Model accuracy (%): 90.03835826620636

Insert disease group code: 2

Is it rainning? Yes

Is it game day? No

Are there any concerts? Yes

Insert season of the year Spring

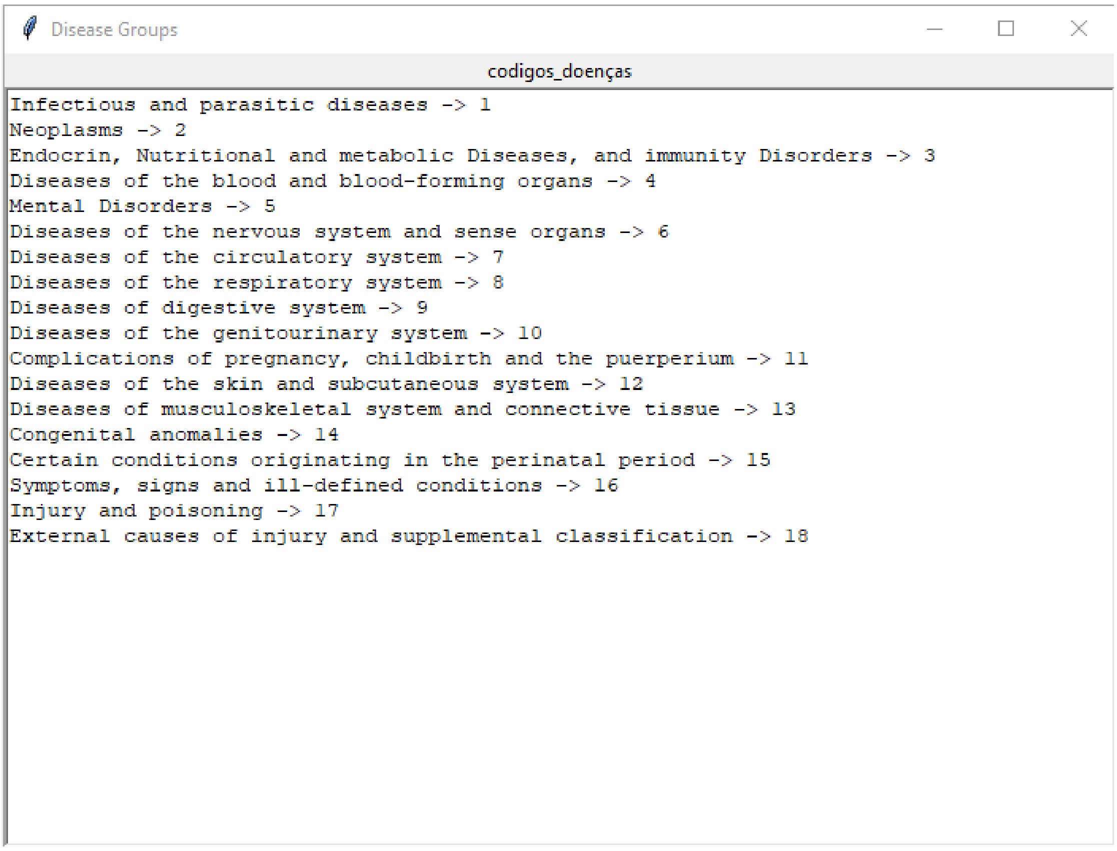
The patient will be at the hospital less than a day.

Figure 21 - Main screen.

To obtain the LoS prediction, the user must first insert the path to the database file in the space indicated with number 1 and press the button marked with number 2, the program then loads the data and creates the predictive model, showing its accuracy in the space marked with number 3.

As soon as it is ready, a popup window will appear indicating that the program is ready to make the prediction. If the path to the database file is not correct, a popup window with an error message will appear instead of the other one.

If everything is correct just fill in the fields corresponding to the disease group, precipitation, games, concerts and season (numbers 4, 6, 7, 8 and 9), and the fields marked with the numbers 6, 7, 8 and 9 are dropdowns where the user only has to choose the correct option, and the field number 4 receives a number that represents the disease group presented according to Table 30, and if the user presses the button marked with the number 5 the window of Figure 22 where the groups are and the corresponding number will appear.



```
Infectious and parasitic diseases -> 1
Neoplasms -> 2
Endocrin, Nutritional and metabolic Diseases, and immunity Disorders -> 3
Diseases of the blood and blood-forming organs -> 4
Mental Disorders -> 5
Diseases of the nervous system and sense organs -> 6
Diseases of the circulatory system -> 7
Diseases of the respiratory system -> 8
Diseases of digestive system -> 9
Diseases of the genitourinary system -> 10
Complications of pregnancy, childbirth and the puerperium -> 11
Diseases of the skin and subcutaneous system -> 12
Diseases of musculoskeletal system and connective tissue -> 13
Congenital anomalies -> 14
Certain conditions originating in the perinatal period -> 15
Symptoms, signs and ill-defined conditions -> 16
Injury and poisoning -> 17
External causes of injury and supplemental classification -> 18
```

Figure 22 - Disease groups and their corresponding codes.

After filling all the fields the user just need to press the button number 10 and the LoS prediction will be presented in the space marked with the number 11.

Chapter 4 – Conclusions.

In this research, the author performed a descriptive analysis of the data provided by the Hospital Garcia da Horta in Lisbon in order to observe the influence that external factors like the disease presented by the patient, the ocurrence of precipitation, the season of the year and the existence of events in the city of the hospital under study (concerts and football matches) could have on the LoS and on the waiting times for the ED and also developed a prototype to perform the LoS prediction based on these same factors.

The author has observed through descriptive analysis that the distribution of cases over the year is similar, and therefore no season stands out in terms of the number of cases. However, the average LoS in winter and spring is higher than in summer and fall as can be seen in Figure 17. The LoS is also higher at weekends than on weekdays and is also longer on rainy days than on rain-free days. For events, LoS is higher on concert days than on non-event days and lower on match days of Sporting CP and SL Benfica.

The waiting times, on the other hand, did not undergo great oscillations with the factors analyzed, with the variation being only a few minutes. The author also concluded that the waiting times for this ED (Table 5) are much higher than the recommended ones (Figure 5) for priorities 3 (yellow), 2 (orange) and 1 (red), and for the lower priorities 4 (green) and 5 (blue) are in accordance with the recommended values.

The author also developed a prototype of a LoS prediction model, which can be used to predict a patient's stay in hospital based, within 4 categories:

- Short (less than a day);
- Medium (between one day and one week);
- Long (between one week and one month);
- Very Long (more than one month).

This prototype can be used by the hospital to find out how long a patient will be in the facility in need of services and to indicate to the patient how much time he/she will spend in the hospital.

During the elaboration of the predictive model the author compared two algorithms NB and RF, using each one of them to perform five simulations, and both had a very similar performance always between 89 and 90% of accuracy with a slight advantage for NB, and therefore was the algorithm chosen to apply in the model.

In order to facilitate the use of the predictive model, the author created a simple user interface and put a detailed description of how to use it so that users do not have to enter data directly into the script, which for those who are not familiar with it becomes quite complex, so the user simply provides the database file for the model to load and then enter the data according to the procedure explained in Chapter 3.4 to get the LoS prediction.

After this research, the author is now capable of answering the questions proposed in section 1.2, concerning question (1), the author verified that it is possible through past data to predict the LoS of patients with an accuracy of about 89%. In response to question (2), the author found that it is possible through the application of DM techniques to find patterns in the information and draw conclusions from these patterns that allow us to make decisions based on them. Finally in answer to questions (3) and (4), both the existence of events near the hospital and the existence of rainfall, the season of the year and day of the week have an influence on LoS and a not so pronounced influence on waiting times, as these did not undergo great oscillations.

In sum, when writing this dissertation, the author did a data analysis study going through all the phases that this process involves, described in Figure 2, understanding the area in which he is working, preparing the data, building the model, testing the model, and finally implementing the model, going to an earlier phase if necessary.

The results obtained were satisfactory, and after this analysis it is safe to conclude that both the time of the year, the day of the week and the disease presented by the patient have an influence on the LoS of the patients, but also the ocurrence of percipitation, the existence of football matches or concerts in Lisbon have an influence. Another conclusion to be drawn is that the analysis of past data makes it possible to analyze the services and find possible factors that are adversely affecting them so that they can be corrected in order to improve the services provided by the hospital and patient satisfaction.

This research has some limitations such as the fact that the data is from 2017 and does not take into account some diseases that have arisen in the meantime such as the case of COVID-19 that could alter the results of the study carried out, another limiting factor in this work is the fact that the data could be more complete such as having the timestamp of the end of medical observation, which would also allow the duration of medical consultations to be analyzed or the treatment times which would also be an important metric to study. The author also compared only two algorithms, and there may be more efficient algorithms than those tested and that would also improve the accuracy of the predictive model.

For future work, in order to generalize this analysis, it would be interesting to be able to analyse other ED datasets in order to create a uniform structure for hospital databases so that this type of analysis can be applied to all of them and not have to be carried out individually. As previously mentioned, the author could also test other algorithms in order to find out if there are any more efficient than those tested to apply to the model, add more factors to evaluate the influence they may have on hospital functioning, or even apply this analysis to more recent and more complete data.

Bibliografy.

[1] R. D. Todor and C. V Anastasiu, “a Future Trend in Healthcare: the Use of Big Data,” *Bull. Transilv. Univ. Brasov. Econ. Sci. Ser. V*, vol. 11, no. 1, pp. 119–124, 2018.

[2] “A model to predict length of stay in a hospital emergency department and en...: Sistema de descoberta para FCCN.” [Online]. Available: https://eds.b.ebscohost.com/eds/pdfviewer/pdfviewer?vid=1&sid=93899492-441a-47d0-a6f8-e9872ddfc185%40pdc-v-sessmgr01. [Accessed: 01-Jan-2020].

[3] M. Barad, T. Hadas, R. A. Yarom, and H. Weisman, “Emergency department crowding,” in *19th IEEE International Conference on Emerging Technologies and Factory Automation, ETFA 2014*, 2014.

[4] J. D. Sonis, E. L. Aaronson, R. Y. Lee, L. L. Philpotts, and B. A. White, “Emergency Department Patient Experience: A Systematic Review of the Literature,” *Journal of Patient Experience*, pp. 101-106, 2018.

[5] R. Kohli and S. S. L. Tan, “Electronic health records: How can is researchers contribute to transforming healthcare?,” *MIS Q. Manag. Inf. Syst.*, vol. 40, no. 3, pp. 553–573, 2016.

[6] A. Sharma and V. Mansotra, “Emerging applications of data mining for healthcare management - A critical review,” in *2014 International Conference on Computing for Sustainable Global Development, INDIACom 2014*, 2014, pp. 377–382.

[7] A. T. Janke, D. L. Overbeek, K. E. Kocher, and P. D. Levy, “Exploring the Potential of Predictive Analytics and Big Data in Emergency Care,” *Ann. Emerg. Med.*, 2016.

[8] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A design science research methodology for information systems research,” *J. Manag. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, 2007.

[9] W. Raghupathi and V. Raghupathi, “Big data analytics in healthcare: promise and potential,” *Heal. Inf. Sci. Syst.*, vol. 2, no. 1, pp. 1–10, 2014.

[10] B. Marr, “How Big Data Is Changing Healthcare,” *forbes.com*, 2015. [Online]. Available: https://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-

data-is-changing-healthcare/#2a54c9a82873. [Accessed: 17-Dec-2019].

[11] C. McDonld, “5 Big Data Trends in Healthcare for 2017 | MapR.”, 2017. [Online]. Available: https://mapr.com/blog/5-big-data-trends-healthcare-2017/. [Accessed: 14-Jul-2020].

[12] R. L. Wears and D. J. Williams, “Big Questions for ‘big Data,’” *Ann. Emerg. Med.*, vol. 67, no. 2, pp. 237–239, 2016.

[13] J. H. Ware, “The limitations of risk factors as prognostic tools,” *New England Journal of Medicine*, vol. 355, no. 25. pp. 2615–2617, 21-Dec-2006.

[14] S. Mishra and A. Misra, “Structured and Unstructured Big Data Analytics,” in *International Conference on Current Trends in Computer, Electrical, Electronics and Communication, CTCEEC 2017*, 2018, pp. 740–746.

[15] M. B. Ateya, B. C. Delaney, and S. M. Speedie, “The value of structured data elements from electronic health records for identifying subjects for primary care clinical trials Healthcare Information Systems,” *BMC Med. Inform. Decis. Mak.*, vol. 16, no. 1, pp. 1–8, 2016.

[16] D. L. Olson and D. Delen, “Data Mining Process,” in *Advanced Data Mining Techniques*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 9–35.

[17] G. Stiglic, P. Kocbek, N. Fijacko, A. Sheikh, and M. Pajnkihar, “Challenges associated with missing data in electronic health records: A case study of a risk prediction model for diabetes using data from Slovenian primary care,” *Health Informatics J.*, vol. 25, no. 3, pp. 951–959, 2019.

[18] M. M. Malik, S. Abdallah, and M. Ala’raj, “Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review,” *Ann. Oper. Res.*, vol. 270, no. 1–2, pp. 287–312, 2018.

[19] M. Srivathsan and A. K. Yogesh, “Health monitoring system by prognotive computing using big data analytics,” in *Procedia Computer Science*, 2015, vol. 50, pp. 602–609.

[20] R. Chauhan and R. Jangade, “A robust model for big healthcare data analytics,” *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference*. pp. 221–225, 2016.

[21] A. T. Janke, D. L. Overbeek, K. E. Kocher, and P. D. Levy, “Exploring the Potential

of Predictive Analytics and Big Data in Emergency Care," *Ann. Emerg. Med.*, vol. 67, no. 2, pp. 227–236, 2016.

[22] B. Van Calster, L. Wynants, D. Timmerman, E. W. Steyerberg, and G. S. Collins, "Predictive analytics in health care: how can we know it works?," *J. Am. Med. Informatics Assoc.*, vol. 26, no. August, pp. 1651–1654, 2019.

[23] H. Chennamsetty, S. Chalasani, and D. Riley, "Predictive analytics on Electronic Health Records (EHRs) using Hadoop and Hive," *Proc. 2015 IEEE Int. Conf. Electr. Comput. Commun. Technol. ICECCT 2015*, pp. 1–5, 2015.

[24] S. Gonçalves, "Predictive analysis in Healthcare," 2018.

[25] S. B. Brill, K. O. Moss, and L. Prater, "Transformation of the Doctor–Patient Relationship: Big Data, Accountable Care, and Predictive Health Analytics," *HEC Forum*, vol. 31, no. 4, pp. 261–282, 2019.

[26] I. Mentzoni, S. T. Bogstrand, and K. W. Faiz, "Emergency department crowding and length of stay before and after an increased catchment area," *BMC Health Serv. Res.*, vol. 19, no. 1, pp. 1–12, 2019.

[27] C. H. Chaou *et al.*, "Predicting length of stay among patients discharged from the emergency department-using an accelerated failure time model," *PLoS One*, vol. 12, no. 1, pp. 1–12, 2017.

[28] M. L. McCarthy *et al.*, "Crowding Delays Treatment and Lengthens Emergency Department Length of Stay, Even Among High-Acuity Patients," *Ann. Emerg. Med.*, vol. 54, no. 4, pp. 492-503.e4, 2009.

[29] Q. Dohan, H. Wong, G. Meckler, and D. Johnson, "The impact of pediatric emergency department crowding on patient and health care system outcomes: a multicentre cohort study," *Cmaj*, vol. 184, no. 3, 2019.

[30] D. Kothari *et al.*, "The Manchester Triage System (MTS): A Score for Emergency Management of Patients With Acute Gastrointestinal Bleeding-A Retrospective Analysi," 2016.

[31] C. for D. C. and Prevention, "ICD - ICD-9 - International Classification of Diseases, Ninth Revision," 2015. [Online]. Available: https://www.cdc.gov/nchs/icd/icd9.htm. [Accessed: 29-Jan-2020].

[32] R. T. Anderson, F. T. Camacho, and R. Balkrishnan, "Willing to wait? The influence of patient wait time on satisfaction with primary care," *BMC Health*

Serv. Res., vol. 7, 2007.

[33] I. Classification and O. F. Diseases, “Manual of the international statistical classification of diseases, injuries, and causes of death; 1975 revision (Volume 1),” *Who, Geneva*, vol. (733p.); U, 1977.

[34] G. Arha, “Reducing Wait Time Prediction In Hospital Emergency Room: Lean Analysis Using a Random Forest Model,” *Univ. Tennessee, Masters Thesis*, 2017.

[35] K. Kapadia, H. Abdel-Jaber, F. Thabtah, and W. Hadi, “Sport analytics for cricket game results using machine learning: An experimental study,” *Appl. Comput. Informatics*, 2019.