COVID-19 SIMULATION PROJECT PLAN

DATA PROCESSES ASSIGNMENT

Authors: María Ayuso Luengo, Doga Cengiz, Javier Gallego Gutiérrez and Pablo Hernández Carrascosa
Degree: Master's Programme in Data Science
Date: January 9, 2022

Document History

Version	Date	Issued by	Description
V1	December 22, 2022	Whole group	First draft
V2	January 2, 2022	Javier Gallego	Revision first draft
V3	January 9, 2022	Whole group	Final version

Contents

1 Executive Summary

For the last two years, COVID 19 pandemic has been a great part of the activity of hospitals. Sometimes even, due to the seasonality of the virus, it has been a reason for the lack of personal and financial resources. As researchers we are fully committed on helping the health system to obtain the maximum success to a reasonable level of profitability. In this report we will identify some of the main problems the hospital might face when fighting the biggest peaks of the pandemic.

As this pandemic has hit all of us, many papers and investigations have been developed focusing on several faces of this crisis. However, we think this report covers many mayor problems to give a simple solution.

2 Introduction

2.1 Motivation

At the beginning of 2020, COVID-19 showed up and everybody's life has changed since then. It was firstly noticed in China, but quickly spread around the globe. In March 2020 a lot of countries (including Spain) had to establish mobility restrictions, close their borders and take some many other decisions in order to protect their population and health care systems. A lot of people died in these first weeks and hospitals had to attend much more people than they were prepared to. After two years, the consequences in personal, labor and economic terms have been big and we are still fighting the virus.

As we said, the health care system is not ready for this disease and keeps on following the same procedures as before the pandemic. To avoid the collapse of the system we will focus on finding essential information about the virus and its effects on different types of patients.

2.2 Purpose and scope of the document

On this project we have run different statistical and machine learning procedures using a simulated data set of patients suffering COVID in the last two years. The data, which is considered to come from the time these patients were admitted to the hospitals, will be described in more depth in following sections.

On the other hand, we consider different challenges that the hospitals might face and try to find a solution from a data mining perspective.

3 Goals

In this section, we will focus on different goals we want to achieve with this project. Our practical application will be envisioning a hospital as our prospective customer and what problems they might have in respect to COVID 19. This problems will be presented in the Business objectives, with its respective procedure from a data mining point of view.

The final products will be both the project plan and a technical report in addition to the prediction model.

3.1 Business goals

Business objective	Description	Indicator of success
BO1: Reduce triage's time	The first steps in the ER is taking basic diagnosis measures, like heart rate or oxygen saturation. When the ER is overcrowded due to COVID these measures should take as little time as possible, therefore the hospital asks for a solution to this problem.	Find measures that are not critical for determining if a person is at risk due to COVID.
BO2: Predict bed occupation	When a COVID wave occurs is important to have a planning strategy in order to prevent as long as possible the lack of beds for patients.	Achieve a high accuracy in a prediction model for the days in hospital.
BO3: Identify population candidate for ICU	Evidently COVID has different effects on different people. The hospital would find useful to have specific qualities for patients they have to look after more.	Be able to give clear specifications about patients that have more probabilities to end in the ICU.

Table 1: Business goals

3.2 Data Science goals

Data mining objective	Description	Indicator of success
DMO1: Decide if there is any feature that is not useful for the hospital purposes	In order to categorize patients, some measures are taken when they get to the ER. Features will be studied through statistical analyses, mainly through correlation plots.	Find features that provide the same information or that have little to no statistical relationship with the target variables.
DMO2: Predict bed occupation	Build a prediction model where the target variable will be "Days in hospital".	Achieve a high accuracy in a prediction model for the days in hospital.
DMO3: Find patterns for people that go to the ICU	In order to find specific characteristics for people that are more affected by COVID, we will apply exploratory analysis tools and survival curves.	Find clear differences between patients that are and are not at high risk.

Table 2: Data Science goals

It is clear that data mining objectives DMO1, DMO2 and DMO3 contribute to the business objectives BO1, BO2 and BO3, respectively.

4 Workplan

We used the CRISP-DM approach [1] to organize and implement the different parts of our project. This methodology differentiates six phases:

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

While the first phase was explained in the previous sections, the rest will be analyzed in depth in the technical report.

4.1 Workpackages

- Business Understanding
  - Examine the variables of the dataset.
  - Research the possible needs of the hospital.
  - Generate the business objectives.
  - Propose solutions.
- Data understanding
  - Study outliers.
  - Analyze different variables distribution.
  - Analyze correlations.
  - Build and study survival curves.
- Data preparation
  - Data collection.
  - Treat missing values.
  - Descriptive analysis, studying maximums, minimums, etc.
  - Treat abnormal values.
- Modeling
  - Select appropriate algorithms.
  - Create different workflows in KNIME.
  - Add and configure nodes.
  - Add additional nodes to improve the models.
- Evaluation
  - Select measures for the evaluation of the model.
  - Interpret the results.

  - Tune hyperparameters in order to get better results.

- Deployment
  - Create a report of the project.
  - Create a README.txt explaining the steps to follow to replicate the process.

4.2 Deliverables

- Report with the data analysis performed.
- A clean data set.
- A set of models that could satisfy the goals.
- The final model.
- A report and a text file explaining the project.

4.3 Milestones

Different milestones represent the end of the different work phases. They are the following;

- Get business goals and solution proposals (end of business understanding).
- Obtain consistent data (end of data preparation).
- Find patterns and variables relationships (end of data understanding).
- Create a prediction model (end of modeling).
- Get the best model (end of evaluation).
- Get the technical report (end of deployment).

4.4 Gantt and Pert

For the development of Gantt and Pert charts we used GanttProject. The planning spans from December 22nd, day we started with the project, to January 9th, deadline for the delivery, and it is detailed in Figures 1 and 2.

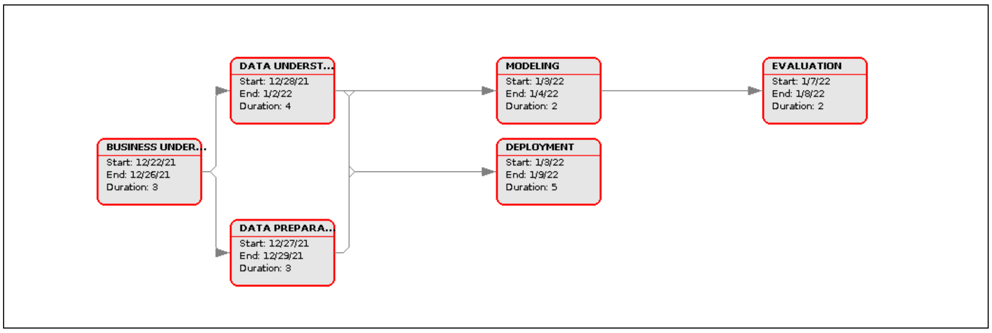
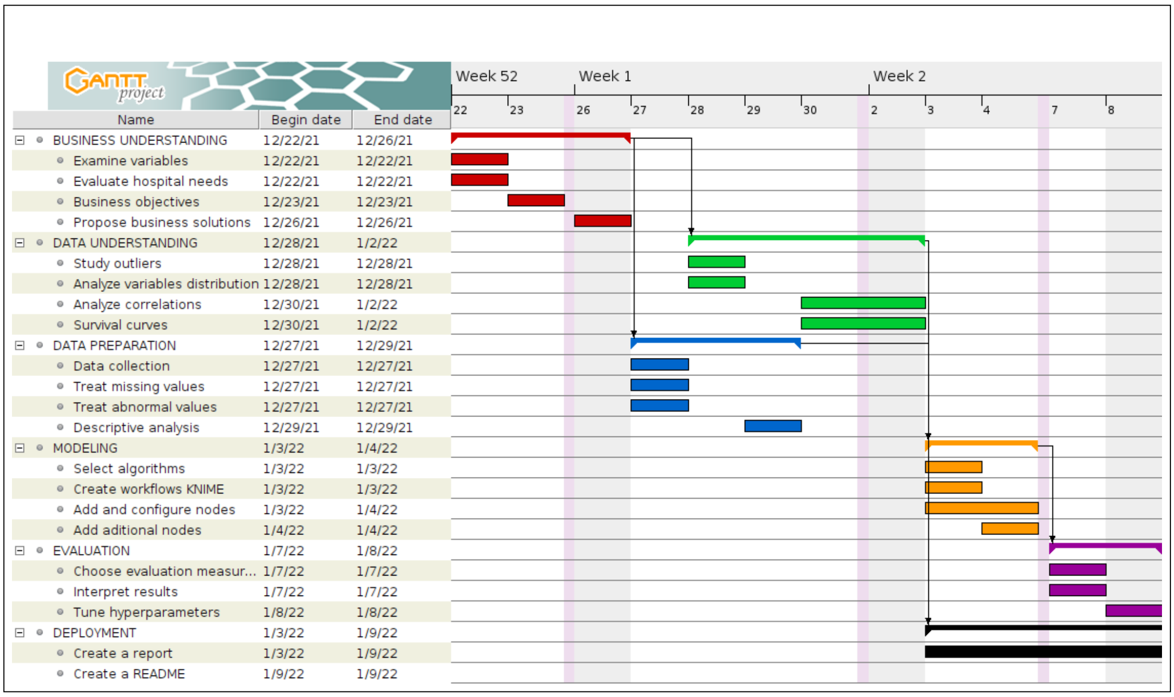


Figure 1: Pert chart



Name	Begin date	End date
BUSINESS UNDERSTANDING	12/22/21	12/26/21
Examine variables	12/22/21	12/22/21
Evaluate hospital needs	12/22/21	12/22/21
Business objectives	12/23/21	12/23/21
Propose business solutions	12/26/21	12/26/21
DATA UNDERSTANDING	12/28/21	1/2/22
Study outliers	12/28/21	12/28/21
Analyze variables distribution	12/28/21	12/28/21
Analyze correlations	12/30/21	1/2/22
Survival curves	12/30/21	1/2/22
DATA PREPARATION	12/27/21	12/29/21
Data collection	12/27/21	12/27/21
Treat missing values	12/27/21	12/27/21
Treat abnormal values	12/27/21	12/27/21
Descriptive analysis	12/29/21	12/29/21
MODELING	1/3/22	1/4/22
Select algorithms	1/3/22	1/3/22
Create workflows KNIME	1/3/22	1/3/22
Add and configure nodes	1/3/22	1/4/22
Add aditional nodes	1/4/22	1/4/22
EVALUATION	1/7/22	1/8/22
Choose evaluation measur...	1/7/22	1/7/22
Interpret results	1/7/22	1/7/22
Tune hyperparameters	1/8/22	1/8/22
DEPLOYMENT	1/3/22	1/9/22
Create a report	1/3/22	1/9/22
Create a README	1/9/22	1/9/22

Figure 2: Gantt chart

5 Risk plan

As this project is not extremely big, the severity of possible risks is not critical. However, we have to take into account some of them and try to mitigate their damage. Table 3 shows what at first we thought were the greatest risks we were facing.

Risk	Probability	Mitigation actions
Low data quality	High	Preprocessing the data
Inadequate analysis	Low	Research various analytic methods
Planning estimation failure	Medium	Organize frequent meetings
Low accuracy in prediction	Medium	Try different algorithms
Medical reasons (COVID-19 infection of a member)	Medium	Advance work when possible

Table 3: Risk plan

6 Budget

A simple estimation of the budget needed to develop the project is shown in Table 4. Actually, we believe most of the expenses have to do with human resources.

Type	**Description**	**Fee**
Personal	Data analyst salary - Doga Cengiz	2000 €/month
	Data analyst salary - Javier Gallego Gutiérrez	2000 €/month
	Data analyst salary - Pablo Hernández Carrascosa	2000 €/month
	Data analyst salary - María Ayuso Luengo	2000 €/month
Travel	City transportation	100 €/month
	Gasoil	100 €/month
Material	Computational resources	500 €/month
	Software	0 €
Indirect costs	Invoices(Electricity, Internet)	100 €/month
	Office rent	1000 €/month

Table 4: Budget plan

7 Conclusions and future steps

In conclusion, data mining techniques are very useful to help mitigate the effects of COVID 19 on our health care system. The quick spread of the virus makes it much more important to see successful results in little time. In this paper we address several problems and explain how these could progress through the results presented on the technical report. It is important to mention that these results could have a high impact by the enlargement of the data set.

Focusing on future researches the study could advance into applying different machine learning techniques. On the one hand to find a group of risk would be interesting to apply clustering and compare the results. Additionally, other prediction models could be trained or more hyper-parameter tuning.

References

[1] P. Chapman, J. Clinton, R. Kerber, *et al.*, “The crisp-dm user guide,” in *4th CRISP-DM SIG Workshop in Brussels in March*, sn, vol. 1999, 1999.

[2] K. L. Petrocelly, “Project management: Survival and success,” in *Project Management: Survival and Success*. 2018, pp. i–xiv.

[3] In *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)*, ser. The Morgan Kaufmann Series in Data Management Systems, I. H. Witten, E. Frank, and M. A. Hall, Eds., Third Edition, Boston: Morgan Kaufmann, 2011, pp. xxi–xxvii, ISBN: 978-0-12-374856-0. DOI: https://doi.org/10.1016/B978-0-12-374856-0.00021-3.