




# A Data Driven Tool to Assess Clinical Outcome Risk

**Team 62 Project Report**  
Data Science for All Colombia 6.0  
Correlation One and MinTIC Colombia  
June 7, 2022



DS  
4A

COLOMBIA



*"The best way to learn data science  
is to do data science"*

**Chanin Nantasenamat**



## Team 62



María Paula **Álvarez** [MA]



Juan **Barrios** [JB]



Daniel **Chavarría** [DC]



Jeyson **Guzmán** [JG]



Cristian **Rodríguez** [CR]



Luis **Serna** [LS]

# Table of Contents

## **1 SUMMARY**

## **2 INTRODUCTION**

Problem overview and scoping

## **3 DATASET**

Description, wrangling and cleaning

## **4 DATA ANALYSIS**

Descriptive analysis and models

## **5 APPLICATION OVERVIEW**

Architecture, hosting and dashboard

## **6 CONCLUSIONS**

Future work and next feasible steps

## **7 REFERENCES**

Team and credits

# 1

## SUMMARY



This document presents the results of...





## 2

# INTRODUCTION

Problem overview and scoping

Team member's effort split:

| MA  | JB  | DC  | JG  | CR  | LS  |
|-----|-----|-----|-----|-----|-----|
| 18% | 16% | 16% | 18% | 16% | 16% |



## Problem overview

Emergency admissions, understood as unpredicted and unscheduled presentation at short notice because of clinical need according to Ismail et al (2017), account for a considerable proportion of hospital bed occupation nowadays, with its associated negative impacts on inpatients with chronic diseases and multimorbidity and high health care resource utilization.

The early determination of each presentation's outcome can help reduce this negative impacts, especially for conditions that are considered non-urgent and that could overwhelm hospital beds availability. Additionally, a preventative approach can help to bring an adequate response for each patient, better focused on its symptoms, functional status, and quality of life, as well as reduce costs.

With this, the development of a clinical decision support system based on predictive tools is needed, to identify patients with the highest risk of admission and help professionals' decision making on presentation.

## Problem scoping

With this project we attempted to assess factors associated with admissions following presentation to emergency departments, using an open access dataset gathered during a month in three different sites in London, and developed a data-based application able to classify the outcome of each presentation as hospitalized or not hospitalized and predict the time of hospitalization of a group of patients uploaded to the application by a health professional user.

The dataset used included workload and inpatient bed occupancy rates for emergency departments, besides the usual demographic data, which allowed to get insights from the results and extrapolated them as possible future work using data from patients in Colombia.

In this context, we tried to answer the question: ***how to support clinical decision making in an emergency department using predictive models to detect patients to be hospitalized and the time they will remain in the service?*** This, with the purpose of improve capacity management in emergency department and be used as a complementary technology to support triage and clinical methodologies.



## 3

# DATASET

Description, wrangling and cleaning

Team member's effort split:

| MA  | JB  | DC  | JG  | CR  | LS  |
|-----|-----|-----|-----|-----|-----|
| 18% | 20% | 16% | 18% | 14% | 14% |

## Dataset description

We used data from risk factors for admission at 3 emergency departments in London, which was originally used for a cross-sectional analysis of attendances in December 2013 (Ismail et al, 2017). The dataset is available as open access resource at Zenodo repository (Ismail et al, 2016) and contains 18 variables for **19,734** unique adult patients aged 16 and older, described as follows :

**Table X.** Data dictionary.

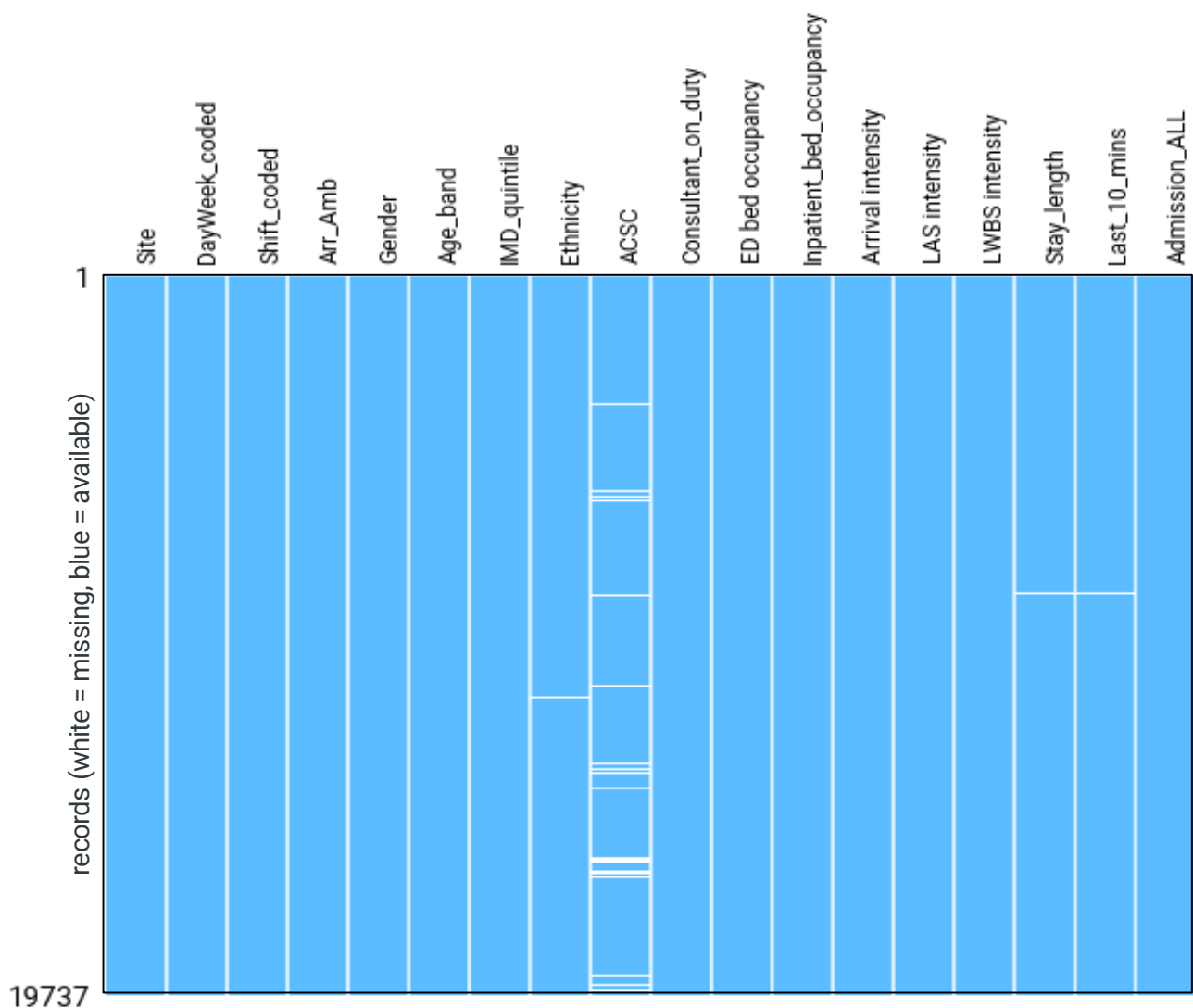
| Variable name  | Description  | Type           | Values   |
|--|--|----------------|--|
| Site_1   | Site of presentation   | int<br>[cat]   | 1 = site 1, 2 = site 2,<br>3 = site 3                                    |
| Study_day  | Day of the month on which the patient presented                                      | int<br>[num]   | 0 to 31  |
| DayWeek_coded  | Day of the week on which the patient presented                                       | int<br>[num]   | 1 = Monday to<br>7 = Sunday  |
| Shift_coded  | Shift during which the patient presented   | int<br>[num]   | 1 = day<br>0 = night   |
| Arr_Amb  | Arrival by ambulance   | int<br>[num]   | 1 = yes<br>0 = no  |
| Gender   | Gender of the patient  | int<br>[num]   | 1 = female<br>0 = male   |
| Age_band*  | Banded patient age   | int<br>[cat]   | 0 = 16-34, 1 = 35-64,<br>2 = 65-84, 3 = ≥85                              |
| IMD_quintile   | Index of multiple deprivation quintile   | int<br>[cat]   | 0 = no deprived<br>1 = least deprived to<br>5 = most deprived            |
| Ethnicity*   | Ethnicity code   | int<br>[cat]   | 1 = asian, 2 = black,<br>3 = mixed, 4 = other,<br>5 = unknown, 6 = white |
| ACSC*  | Diagnostic indicating presentation because of an ambulatory care sensitive condition | int<br>[cat]   | 1 = yes, 0 = no,<br>3 = unknown (imputed)                                |
| Consultant_on_duty   | Consultant on duty in the unit   | int<br>[num]   | 1 = yes<br>0 = no  |
| ED bed occupancy   | Emergency department bed occupancy rate for the preceding hour                       | float<br>[num] | 0.08 to 2.70   |
| Inpatient_bed_occupancy  | Inpatient bed occupancy rate for the day   | float<br>[num] | 0.82 to 1.00   |
| Arrival intensity  | Arrival intensity within the preceding hour  | float<br>[num] | 1.00 to 36.00  |
| LAS intensity  | Ambulance arrival intensity (proportion of arrivals presenting via ambulance)        | float<br>[num] | 0.00 to 1.00   |
| LWBS intensity   | Proportion of patients within the hour who leave without being seen by a doctor      | float<br>[num] | 0.00 to 1.00   |
| Stay_length*   | Length of stay in the Emergency Department in minutes                                | int<br>[num]   | 0 to 1516  |
| Last_10_mins*  | Patient disposition decision made in the last 10 minutes before the four-hour target | int<br>[num]   | 1 = yes<br>0 = no  |
| <b>Source:</b> modified from Ismail et al, 2016 (* contains missing data). |  |                |  |

The 18 variables contain information about patient, clinician and organizational features, such as demographic characteristics, emergency department workload and staffing, chronic pathologies diagnostic and inpatient bed occupancy rates.

## Wrangling and Cleaning

Missing data were identified for the dataset: 1 for Age\_band, 8 for Last\_10\_mins, 13 for Ethnicity, 21 for Stay\_length and 884 for ACSC (approximately 4.5% of missing data in this last variable).

**Figure X.** Missingness plot



**Source:** own elaboration.

We applied the following procedures:

- For ACSC variable, which indicates if the patient presents a sensitive care condition (i.e., chronicity), we decided to impute category 3 for the 887 missing records, to indicate that this field is not known and will be treated as a categorical variable. With this, when this variable turned out to be important for the prediction models, we could interpret it and indicate that patients for whom it is not known whether they came for treatment of a sensitive condition have a worse clinical emergency outcome.
- For Ethnicity variable the 13 missing records were imputed as 5, which corresponds to 'unknown' category in the data dictionary. This makes sense for patients with unknown information in this field.
- For Stay\_length variable the 21 missing records were deleted for the whole dataset, as this is one of the response variables and the size of missing data is small related to the size of the full dataset (more than 19000 records). With this last procedure the remaining missing data in Age\_band and Last\_10\_mins were also removed.

With this procedures we got a clean dataset, as the original resource had already gone through an extensive data transformation process the was performed prior to the analysis shown in Ismail et al (2017).

After this, we looked for hospitalization predictors in the dataset with two different approaches: a classification problem for 'hospitalized' or 'not hospitalized' categories and a regression problem for time of hospitalization, as shown in the next chapter.



## 4

# DATA ANALYSIS

Descriptive analysis and models

Team member's effort split:

| MA  | JB  | DC  | JG  | CR  | LS  |
|-----|-----|-----|-----|-----|-----|
| 16% | 16% | 24% | 16% | 14% | 14% |

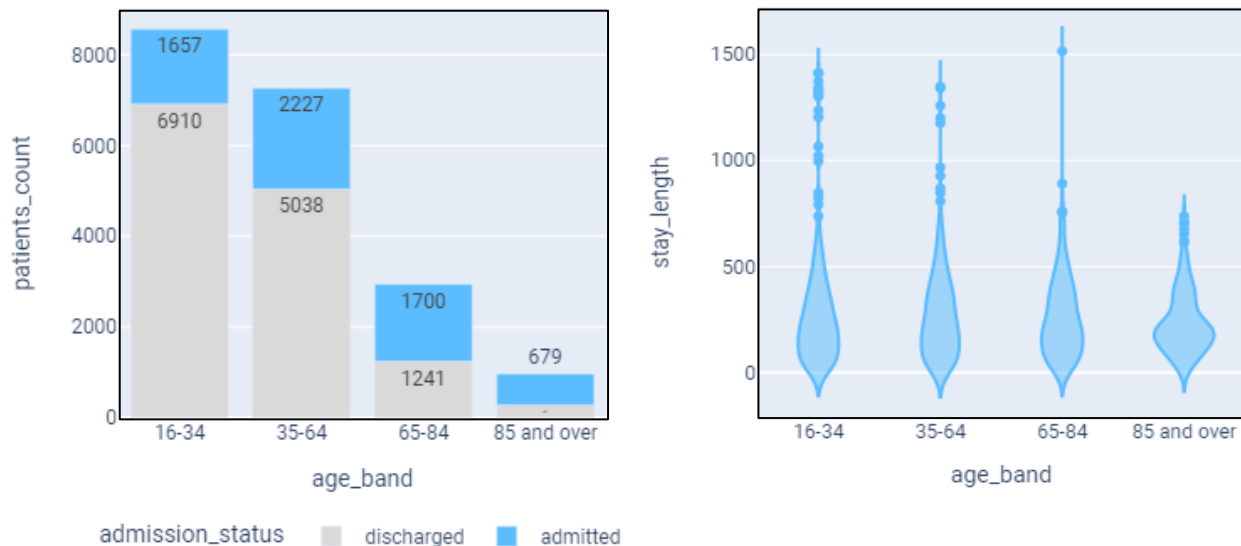
## Descriptive Analysis

There are two main variables in the dataset: the admission result (Admission\_ALL) and the stay length of patients (Stay\_length) that are prediction modeling targets. These variables can be examined beside demographic characteristics and inpatient bed occupancy rates in order to characterize the sample population represented by the dataset.

### Demographic factors

Related to demographic factors, the age bands and genre of this sample population showed the greater portion of patients concentrated in 16-34 and 35-64 bands. As expected, when age band increases the admission does too. In general, the 32% of population was admitted in the emergency room. On the other hand, people from 85 and over presented lower variability and extreme values compared with other age band groups.

**Figure X.** Admission status (bars) and stay length (violins) by age band plots



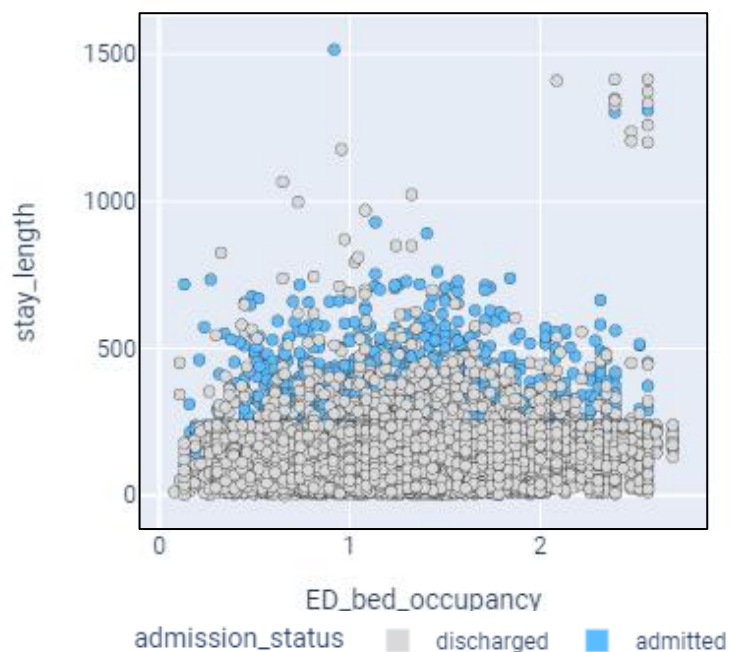
**Source:** own elaboration.

## Emergency department workload

Inpatient bed occupancy rates are an important indicator for emergency departments, because it shows use of installed capacity, that is, total number of patients in emergency department divided by total number of licensed beds. It is desirable to keep this rates as low as possible, ensuring place for new patients but without falling into underutilization.

The next figure shows that admitted patients tend to stay longer in emergency department from 250 to 600 minutes and in remarkably high and low emergency department bed occupancy the stay length is lower.

**Figure X.** Admission status (bars) and stay length (violins) by age band plots



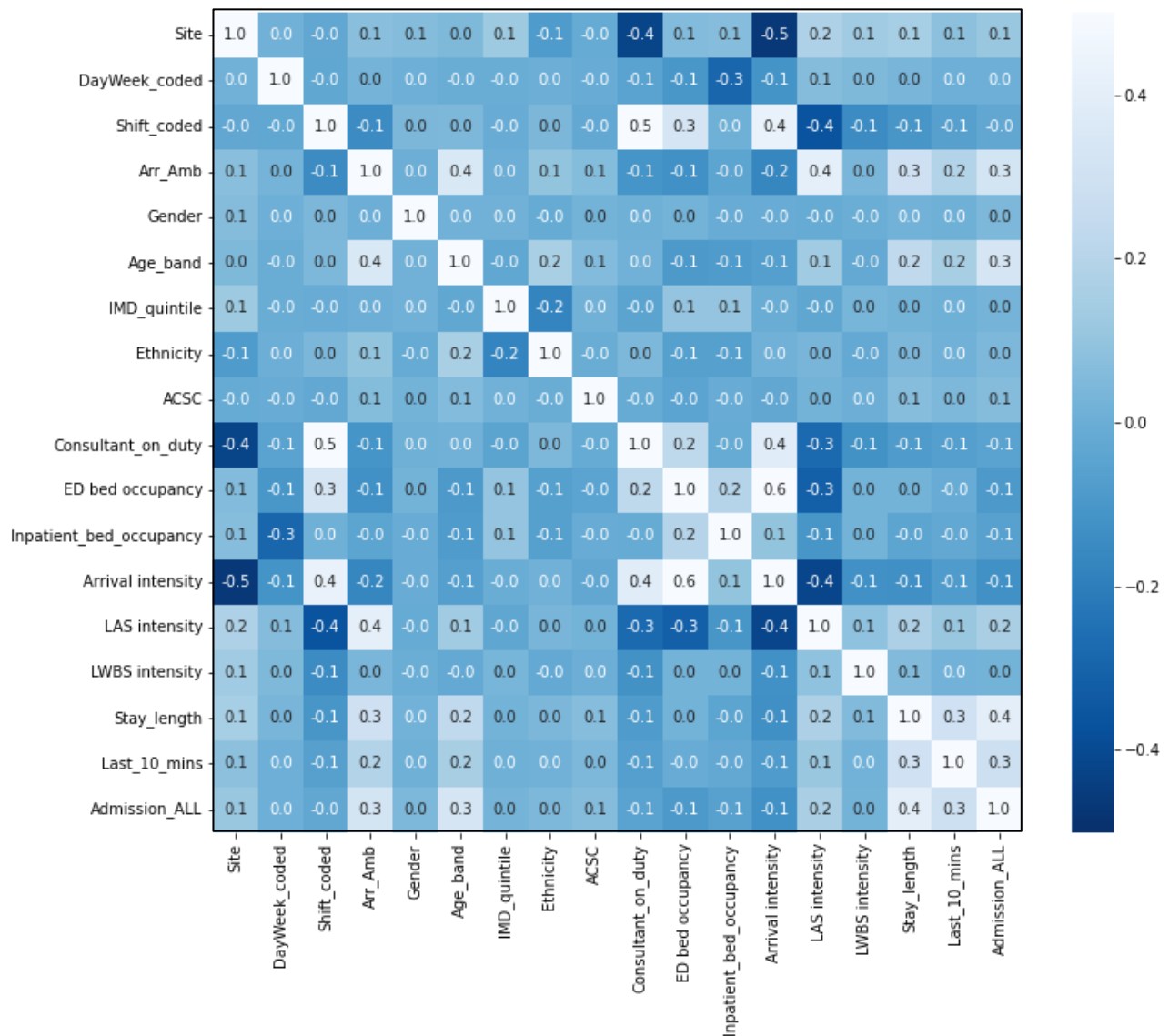
**Source:** own elaboration.



## Correlations

Finally, the linear correlation among all the variables, are shown in the next figures. Relation between ED bed occupancy rate and Arrival intensity is the higher value.

**Figure X.** Correlation matrix



**Source:** own elaboration.

# Model selection

Two training and test subsets were created from original dataset (at 80%-20% proportion) for two problems: a classification set for the Admission\_ALL and a regression set for Stay\_length variables.

It was ensured that neither of the two target variables were included in the training sets, since they are unknown during arrival at the emergency department.

One-hot encoding was used for Site, Age\_band, IMD\_quintile, Ethnicity, ACSC, Admission\_ALL and Stay\_length variables.

After building the first model with the training data, several comparison metrics were used, and k-fold cross validation were performed to select best model.



## **Classification problem**

Presentation to emergency department  
outcome prediction as hospitalized or  
not hospitalized.



## **Regression problem**

Presentation to emergency department  
outcome prediction as time of  
hospitalization in minutes.

## Classification model

Six models were initially tested with default parameters on the training set. These were Logistic Regression, Random Forest, K-Nearest Neighbors, Support Vector Machines, Gaussian Naive Bayes and XGBoost, with three different scores: ROC-AUC, Sensitivity and Specificity. 10 k-fold cross validation showed the following mean scores for each model:

**Table X.** Classification model k-fold cross validation output

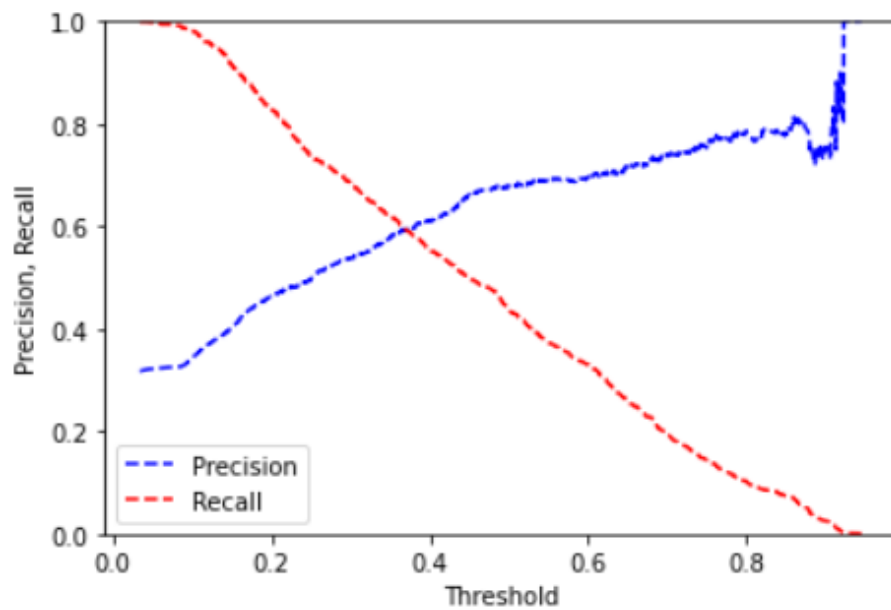
| Model                   | ROC-AUC | Sensitivity | Specificity |
|-------------------------|---------|-------------|-------------|
| Logistic Regression     | 0.781   | 0.446       | 0.904       |
| Random Forest           | 0.767   | 0.472       | 0.883       |
| K-Nearest Neighbors     | 0.674   | 0.320       | 0.890       |
| Support Vector Machines | 0.766   | 0.356       | 0.910       |
| Gaussian Naive Bayes    | 0.746   | 0.551       | 0.818       |
| XGBoost                 | 0.761   | 0.476       | 0.865       |

**Source:** own elaboration.

The main interest of the clinical decision support system based on predictive models is to optimize the detection of well-classified hospitalized patients, even sacrificing some non-hospitalized patients with a greater error. For this purpose, the models with better mean scores were Logistic Regression and Gaussian Naive Bayes, with Random Forest and XGBoost showing a good Sensitivity too.

Elastic Net, Lasso and Ridge regularization were tried, and model tuning were performed using Elastic Net with a classification threshold of 0.25 that was identified through iteration to maximize Sensitivity without sacrificing too much Specificity. Even when an optimum balance point around 0.4 was identified analyzing the model precision and recall, we decided to continue with 0.25 threshold, since this has a response in accordance with the purpose defined for the decision tool.

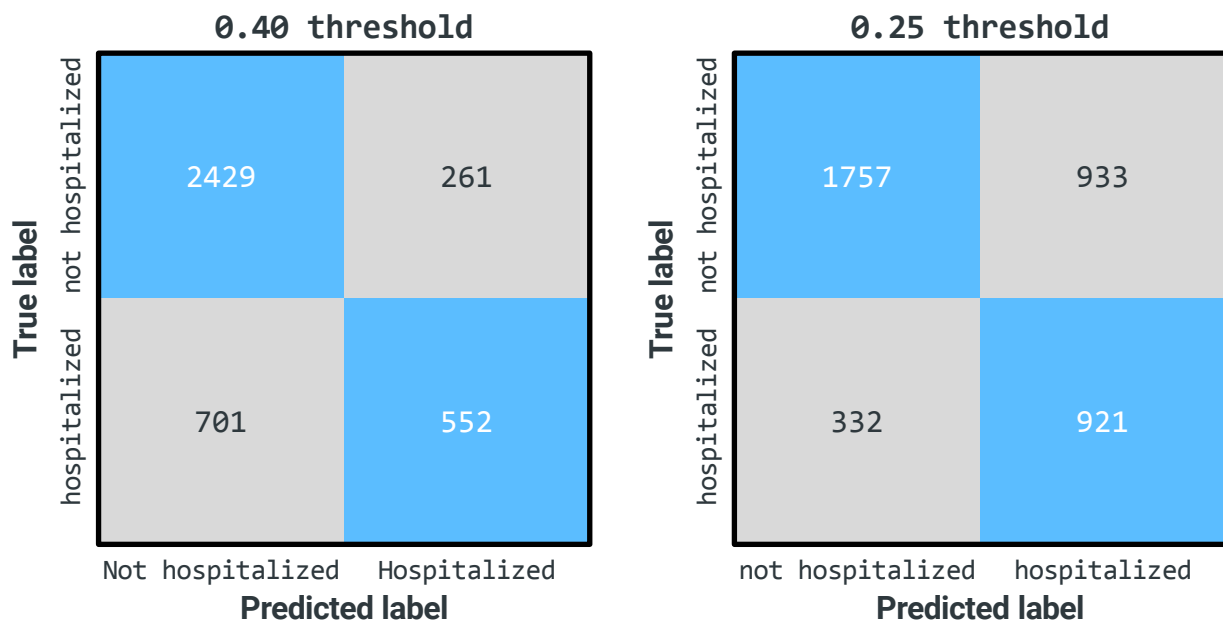
**Figure X.** Precision-Recall vs Threshold Chart



**Source:** own elaboration.

Confusion matrix for the chosen tuned model are shown next, with and without adjusted threshold, emphasizing in...

**Figure X.** Confusion matrix for classification tuned models



**Source:** own elaboration.

## Regression model

Regression model.

## 5

# APPLICATION OVERVIEW

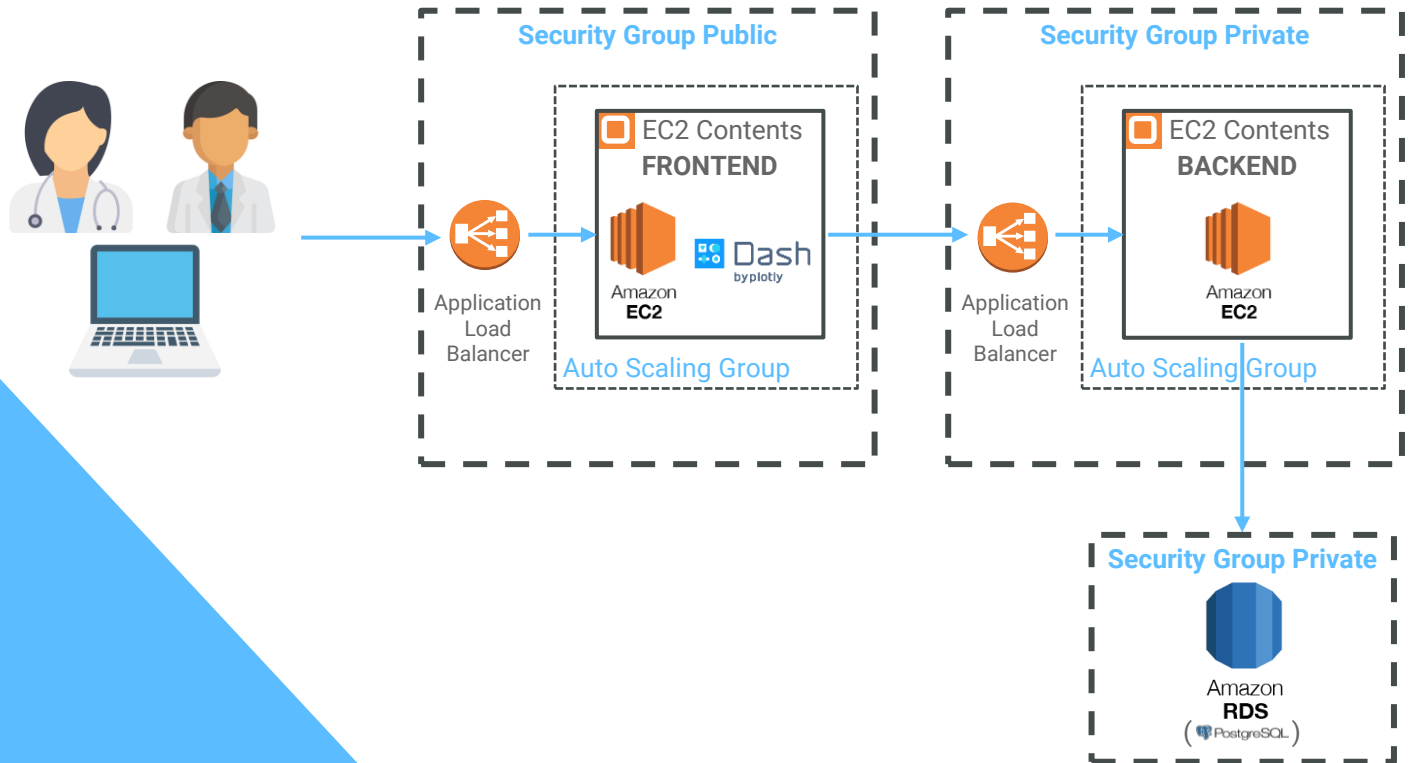
Architecture, hosting and dashboard

Team member's effort split:

| MA  | JB  | DC  | JG  | CR  | LS  |
|-----|-----|-----|-----|-----|-----|
| 12% | 13% | 14% | 13% | 24% | 24% |

# Solution architecture and hosting

**Figure X.** Web application architecture



**Source:** own elaboration.

Architecture is designed so that the user can only enter the frontend, and for this, access is provided to the entire public through the security group. The load balancer distributes the traffic of incoming requests via http or https and, if necessary, will tell the auto scaling group to generate new instances (max. 2 instances). In EC2 instance the project will be hosted in production mode.

For the backend, the load balancer, autoscaling group, and EC2 work the same way as in the frontend, but security group only allows traffic coming from the frontend to pass through, thus protecting data.


For the RDS, a security group was assigned which only allows traffic from the backend to protect the data hosted on it.





## Interactive Front-end

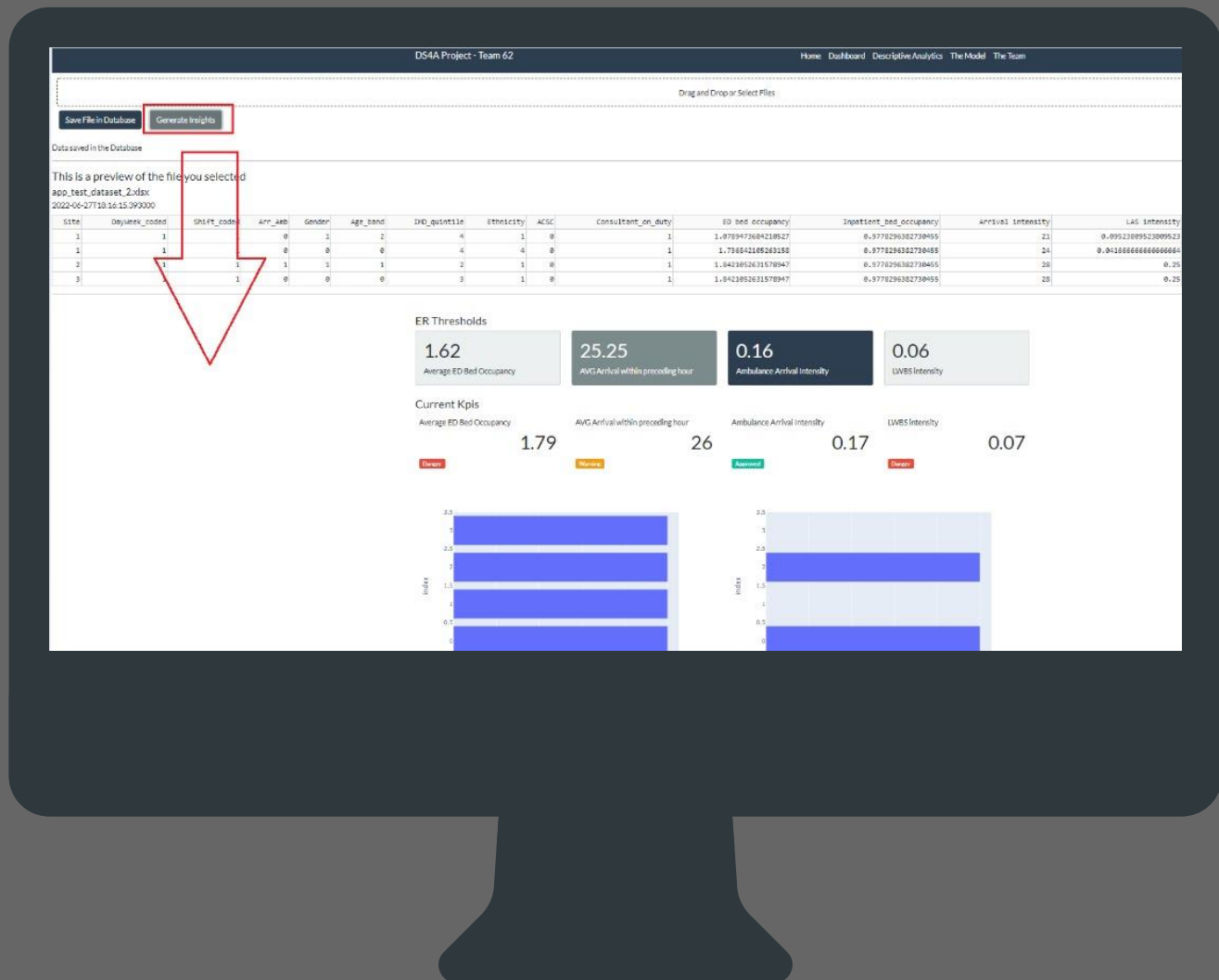
A functional web application was built using Dash library to show results of the project including 4 modules:

- **Dashboard:** in this module the user can upload a csv file with daily patients to run the classification and regression tuned models. The dashboard cards will show the uploaded file calculated KPIs compared with the thresholds defined by the training dataset.
  - **Descriptive analytics:** in this module different plots will show demographic characteristics of the population represented by the uploaded file.
  - **The Model:** in this module key information about classification and regression tuned models will be shown, for users interested in technical details of modeling performed.
  - **The Team:** this module presents the team members.
- 



**Database**





<https://finaldeployedurl.com>

Source code was made through collaborative work using  
GitHub and is available at repository:

 <https://github.com/Teett/ds4a-team-62>

## 6

# CONCLUSIONS

Future work and next feasible steps

Team member's effort split:

| MA  | JB  | DC  | JG  | CR  | LS  |
|-----|-----|-----|-----|-----|-----|
| 19% | 18% | 14% | 18% | 15% | 16% |

## Conclusions



# 7

## REFERENCES

**Ismail, S. A., Pope, I., Bloom, B., Catalao, R., Green, E., Longbottom, R. E., Jansen, G., McCoy, D., & Harris, T. (2017).** Risk factors for admission at three urban emergency departments in England: a cross-sectional analysis of attendances over 1 month. *BMJ open*, 7(6), e011547. <https://doi.org/10.1136/bmjopen-2016-011547>.

**Ismail, Sharif A., Pope, Ian, Bloom, Benjamin, Catalao, Raquel, Green, Emilie, Longbottom, Rebecca E., ... Harris, Tim. (2016).** Data from: Risk factors for admission at three urban emergency departments in England: a cross-sectional analysis of attendances over 1 month [Data set]. Retrieved from <https://zenodo.org/record/4946759>.

---

**Photo credits:**

Crop doctor sitting near devices in office  
Top view person writing laptop with copy space  
Close up doctor filling medical form with patient

**Template credits:**

This template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**  
(free license for personal and commercial purpose with attribution)