

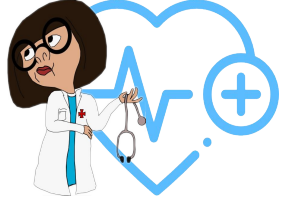
# Emergency Department Clinical Assistance Tool

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



DS  
4A

COLOMBIA



EDNA

*"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

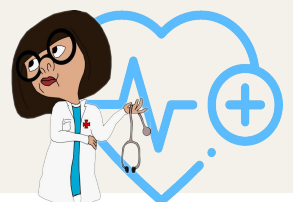
Architecture, hosting and dashboard

## 6 CONCLUSIONS

Future work and next possible steps

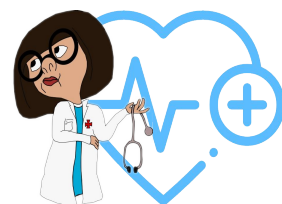
## 7 REFERENCES

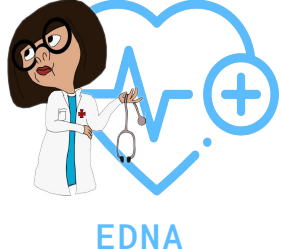
Team and credits



## 1

## SUMMARY





Medical emergency departments (ED) account for a considerable proportion of patient admissions at hospitals, and for that reason, efficiency processes at presentation are needed to avoid overcrowded rooms and long waiting times that could affect negatively the quality of service.

With the purpose of improving capacity management in emergency department and be used as a complementary technology to triage and other clinical methodologies, we developed **EDNA**, a clinical decision support tool based on predictive models that identifies patients to be hospitalized and their expected time in emergency room.


The results of this project are presented in this document.

## 2

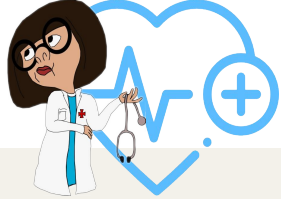
# INTRODUCTION

Problem overview and scoping

Team member's effort split:



MA	JB	DC	JG	CR	LS
15%	20%	20%	15%	15%	15%



## 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 emergency department (ED) event 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 hospitals.

## 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, often referred as ED coordinator.

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 that will 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.



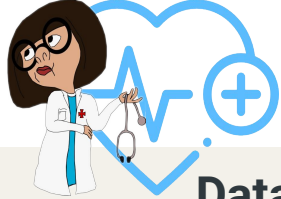
## 3

# DATASET

Description, wrangling and cleaning

Team members effort split:

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



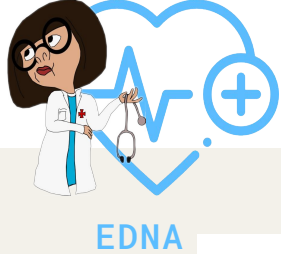
## Dataset description

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 1.** Data dictionary.

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

**Source:** modified from Ismail et al, 2016 (\* contains missing data).

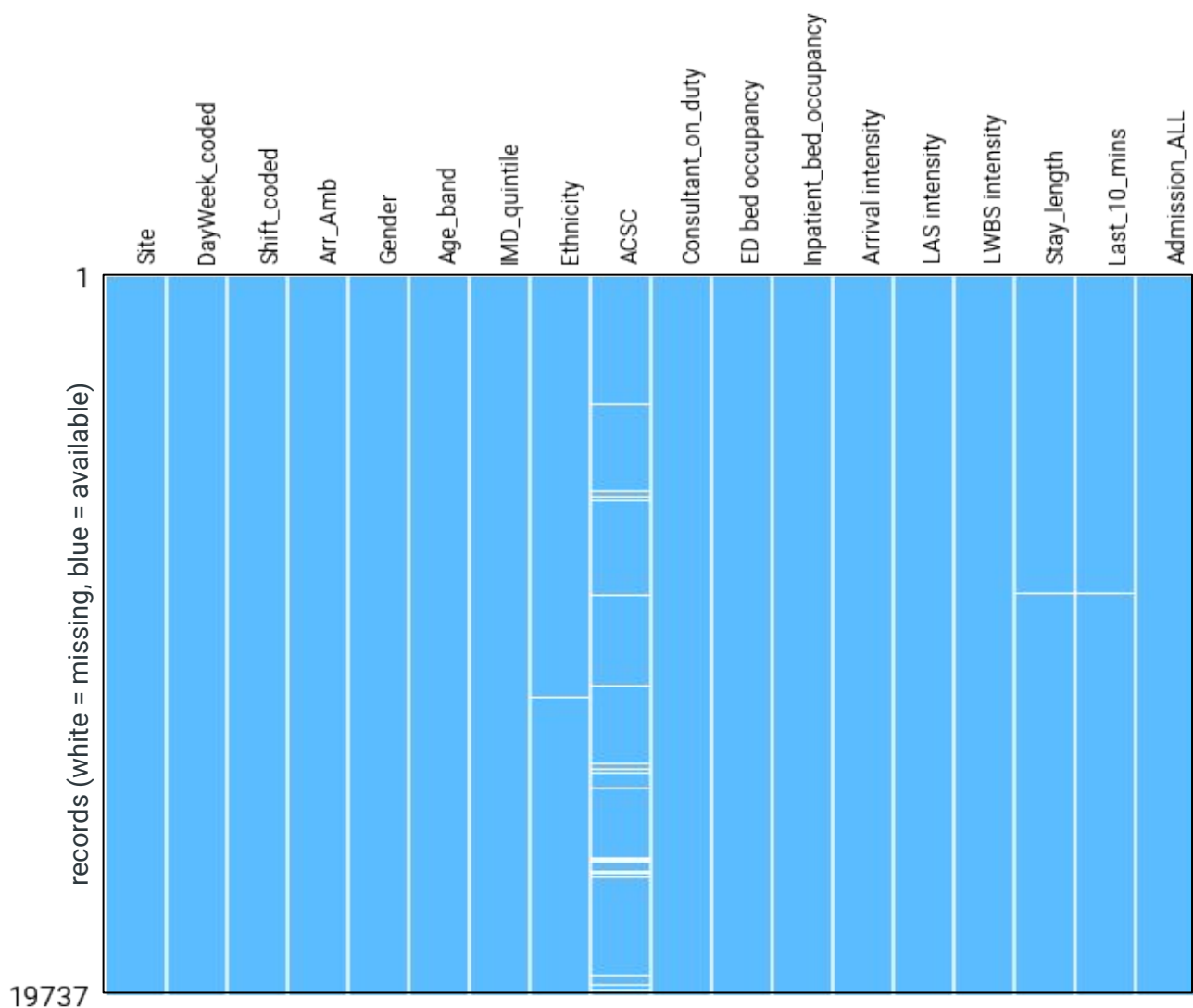


The 18 variables contain information about patients and the clinic's organizational features, such as demographic characteristics, emergency department workload and staffing, chronic pathologies diagnostic and inpatient bed occupancy rates.

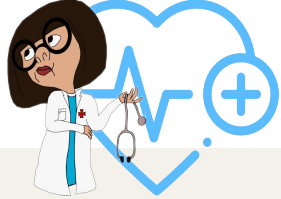
## Wrangling and Cleaning

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 1.** Missingness plot



**Source:** own elaboration.



We applied the following procedures:

- For the ACSC variable, which indicates if the patient presents a sensitive care condition (i.e., chronicity), we decided to impute a new category called “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 the variable turns out to be important for the predictive 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 the 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 the 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 by coincidence.

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

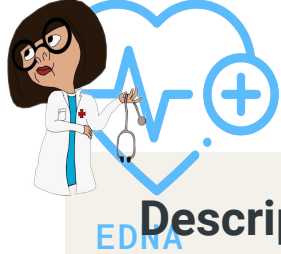
Descriptive analysis and models

Team member's effort split:



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





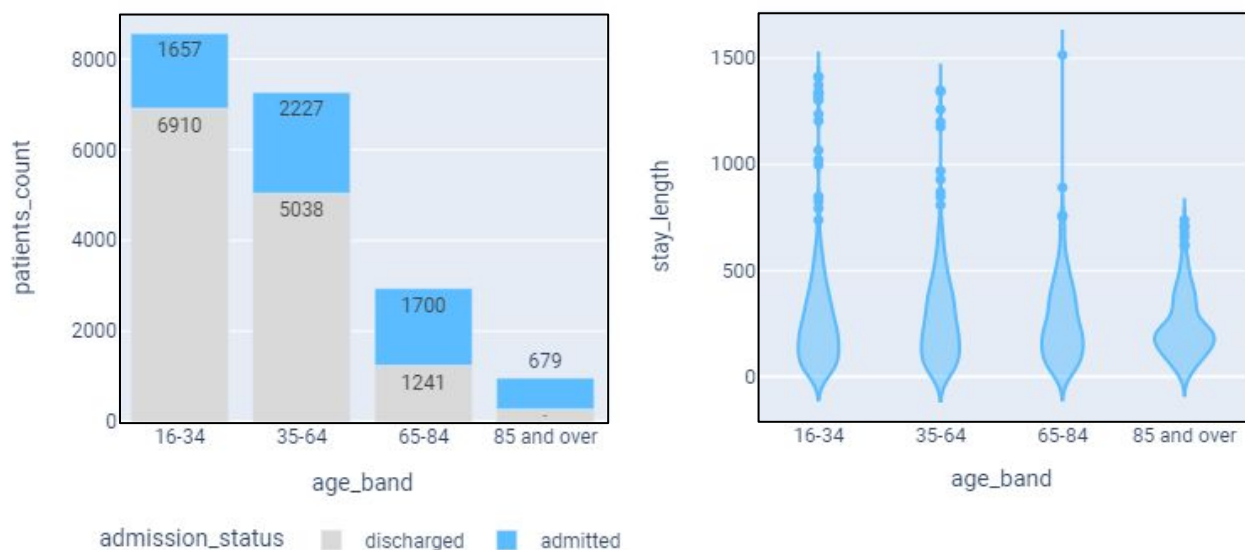
## Descriptive Analysis

There are two main variables in the dataset: the admission result (Admission\_ALL) and the patients stay length in the Emergency Department (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.

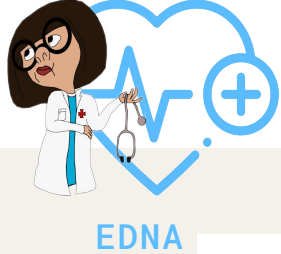
### Demographic factors

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



**Source:** own elaboration.

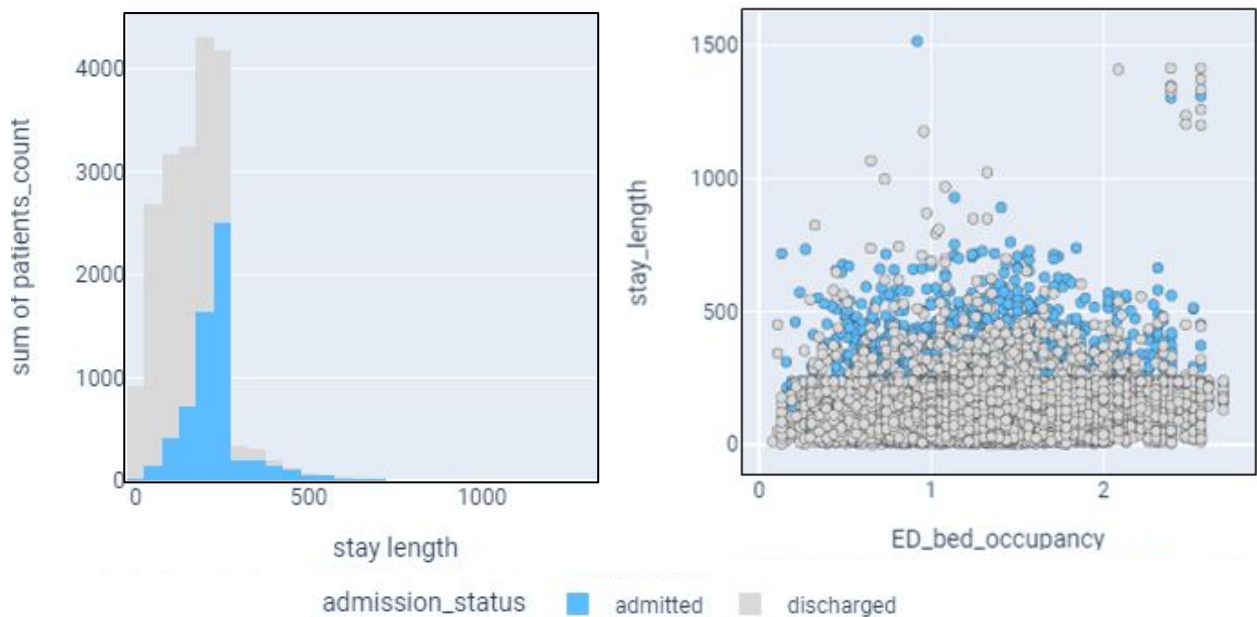


## Emergency department workload

Bed occupancy rates are an important indicator for emergency departments, it shows the installed capacity usage, 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 the emergency department from 175 to 274 minutes while discharged patients from 75 to 275 and in remarkably high bed occupancy it is more likely to experience extreme long stays time.

**Figure 3.** Stay length (histogram) and stay length vs bed occupancy rates (scatterplot) by admission status

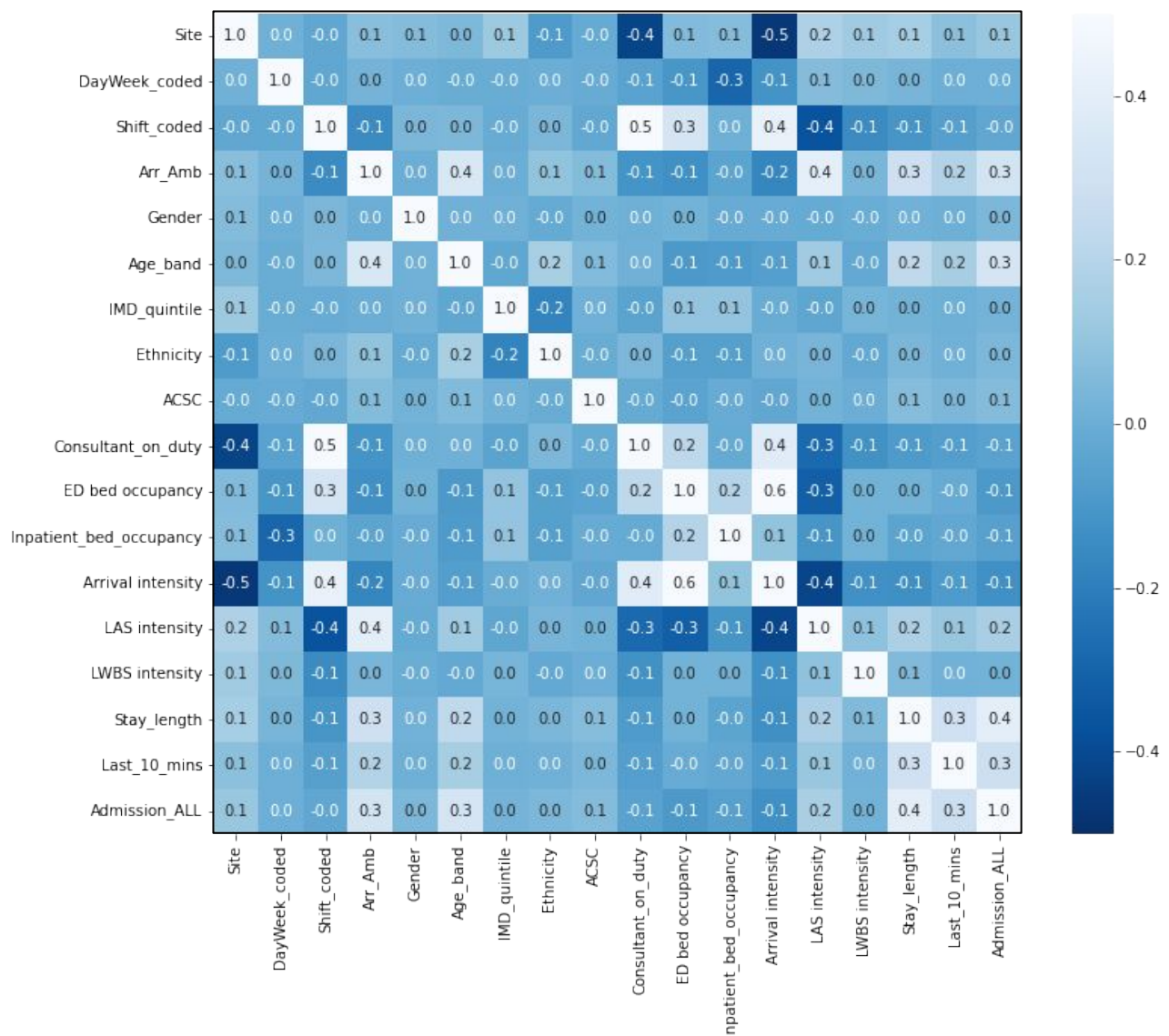


**Source:** own elaboration.

## 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 4. 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.

At first, 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.

Afterwards, it was identified that the result of the classification algorithm could be fed into the regression problem to improve performance.



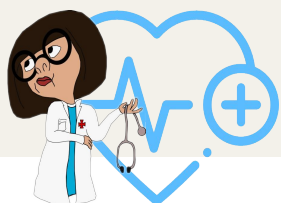
## Classification problem

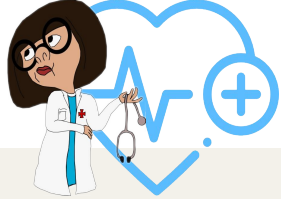
Presentation to emergency department  
outcome prediction as hospitalized or not  
hospitalized.



## 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-fold cross validation showed the following mean scores for each model:

**Table 2.** Classification model 10-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

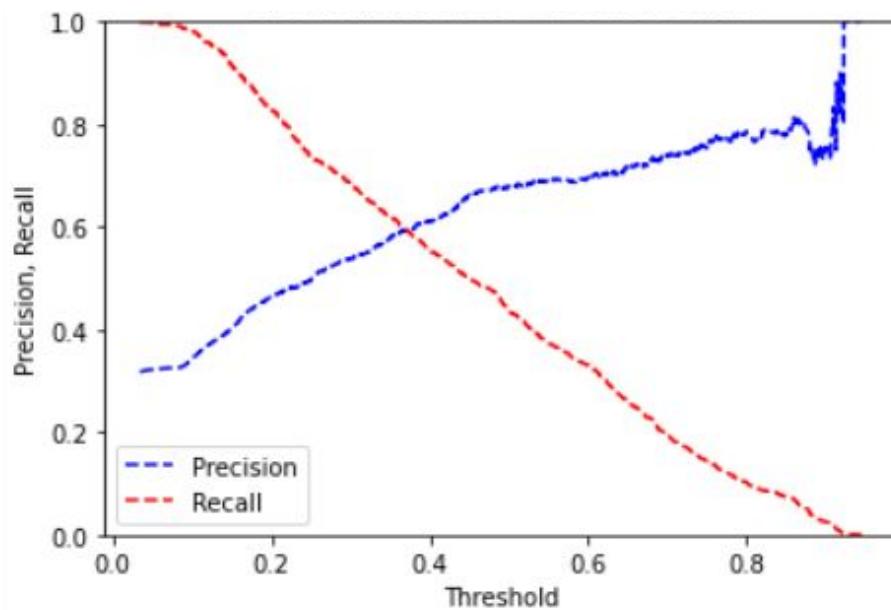
**Source:** own elaboration.

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 was performed using ElasticNetCV 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 that we validated with an expert clinician.



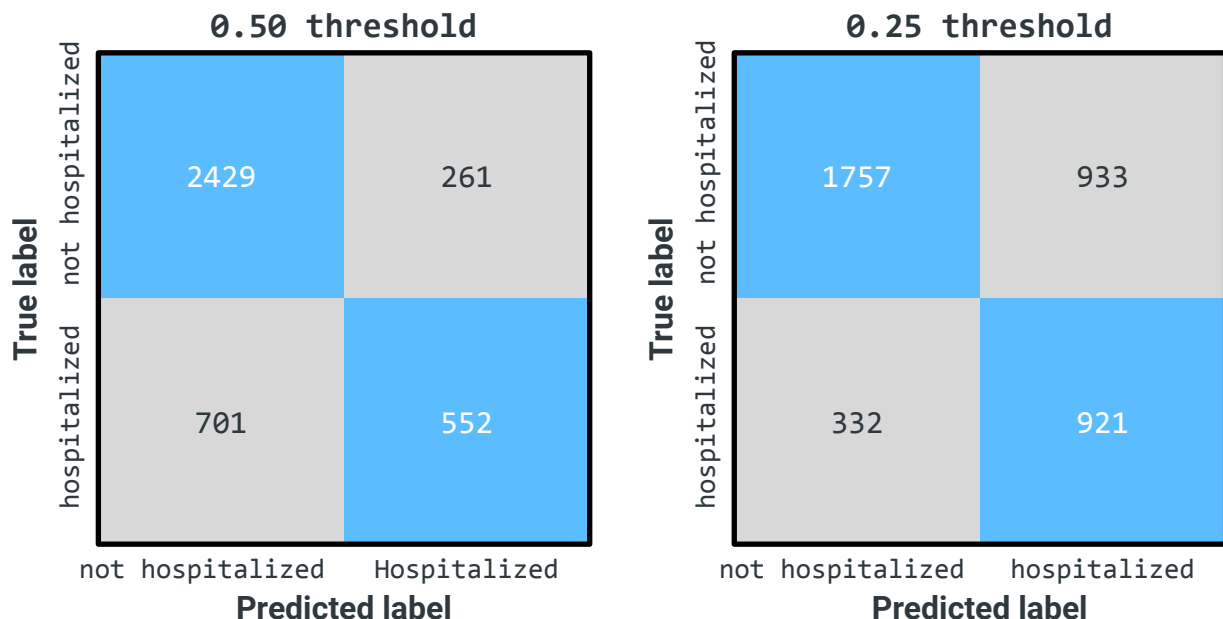
**Figure 5.** Precision-Recall vs Threshold Chart



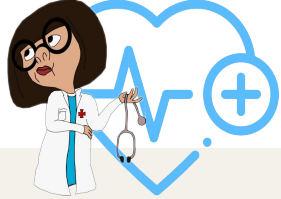
**Source:** own elaboration.

On figure 5 we can see that the precision and recall tradeoff intersection occurs around 0.4, but since the errors of not identifying a hospitalized patient are way more costly, we decided on 0.25 through iteration and a clinician's advice shown in figure 6 with the adjusted threshold.

**Figure 6.** Testing confusion matrices for the classification tuned model



**Source:** own elaboration.

**Table 3.** Final classification model coefficients

Variable	Coefficients
DayWeek_coded	-0.005116
Shift_coded	0.010057
Arr_Amb	0.960790
Gender	0.143640
Consultant_on_duty	0.152715
ED bed occupancy	-0.220076
Inpatient_bed_occupancy	-0.126819
Arrival intensity	-0.000614
LAS intensity	-0.060602
LWBS intensity	0.703586
Last_10_mins	1.312060
Site_2.0	0.492834
Site_3.0	0.446439
Age_band_1.0	0.440978
Age_band_2.0	1.234670
Age_band_3.0	1.643139
IMD_quintile_1.0	0.999944
IMD_quintile_2.0	0.960572
IMD_quintile_3.0	1.014671
IMD_quintile_4.0	0.970671
IMD_quintile_5.0	1.065084
Ethnicity 2.0	0.268959
Ethnicity 3.0	0.014554
Ethnicity 4.0	-0.045803
Ethnicity 5.0	-0.100549
Ethnicity 6.0	0.003477
ACSC_1.0	0.219609
ACSC_3.0	-0.439799

The coefficients of the tuned classification model (Elastic-Net Regression) can be seen.

It's worth noting that the top 3 variables that contribute to the hospitalization of a person are:

1. Being on the age band 3 (>85 years)
2. The amount of patients admitted in the last 10 minutes 4 hours before.
3. Being on the IMD\_quintile 5 (heavily sleep deprived).

The first variable makes sense, as people in the older ages tend to get hospitalized much more often due to their comorbidities and the natural aging process.

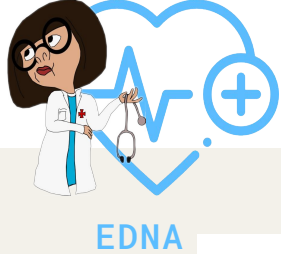
The second variable represents directly how busy the hospital is right now and whether there might be an emergency event (accident / surge of hospitalizations) in the last 4 hours.

Regarding the sleep deprivation, in Ismail et al (2017) it's explained why they decided to capture this data, as it's heavily related to hospitalization odds and could be a confounding variable.

#### Final metrics:

Sensitivity on testing: 73.50%  
Specificity on testing: 65.31%

**Source:** own elaboration.



## Regression model

Initially with the purpose of validate which regression model fit the best to the data from stay analysis we validate four (4) possible candidates to test, those were: Linear regression, Gradient boosting regression, Elastic net and XGBoost.

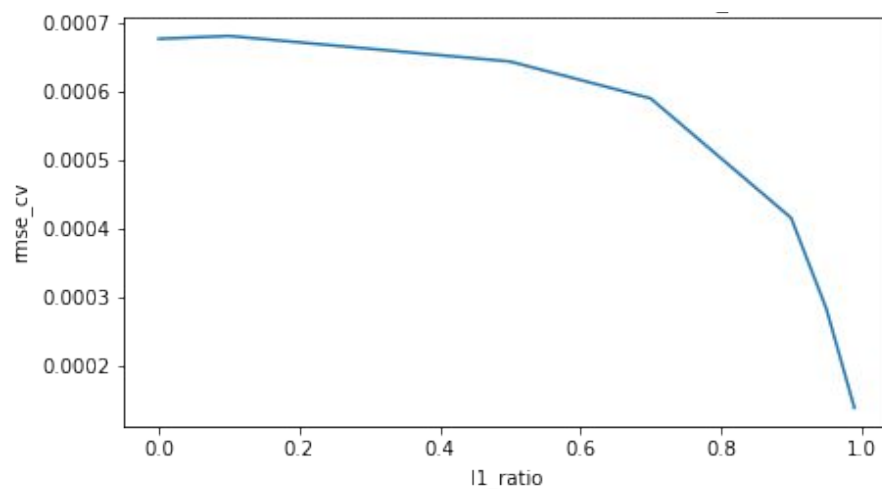
**Table 4.** RMSE for regression models

Reg Model	RMSE
Linear Regression	92,5144
Boosting Regression	87,9537
Elastic Net	87,9537
XGBoost	88,492

**Source:** own elaboration.

Based on these preliminary results, we noticed that based on the metric RMSE (Root Mean Squared Error) for Elastic Net the metric is closer to zero among the other models, that signify that the model and it is predictions fits better.

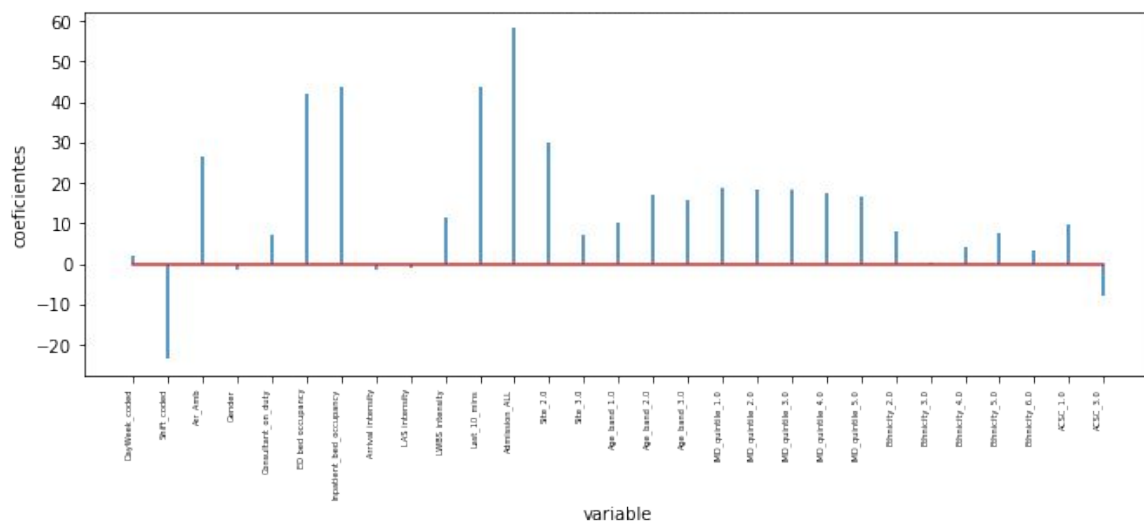
**Figure 7.** CV error vs L1 ratio



**Source:** own elaboration.

In the previous graph we noticed that the optimal level of the parameter  $\alpha$  that better works for Elastic Net regression model which it will be used to predict stay in urgency room.

**Figure 8.** Final regression model coefficients



**Source:** own elaboration.

If we go deeper on the previous graph, we observe how each of the variables affects the model, for example variable "Admission\_All" affects positively the stay prediction and by other hand we noticed that "Shift\_Coded" (during which the patient presented day or night).

## 5

# APPLICATION OVERVIEW

Architecture, hosting and dashboard

Team members effort split:

MA	JB	DC	JG	CR	LS
10%	12%	18%	12%	24%	24%



## Solution architecture and hosting

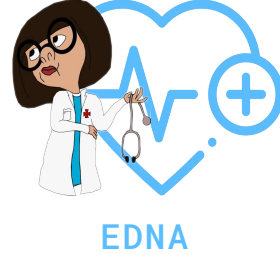
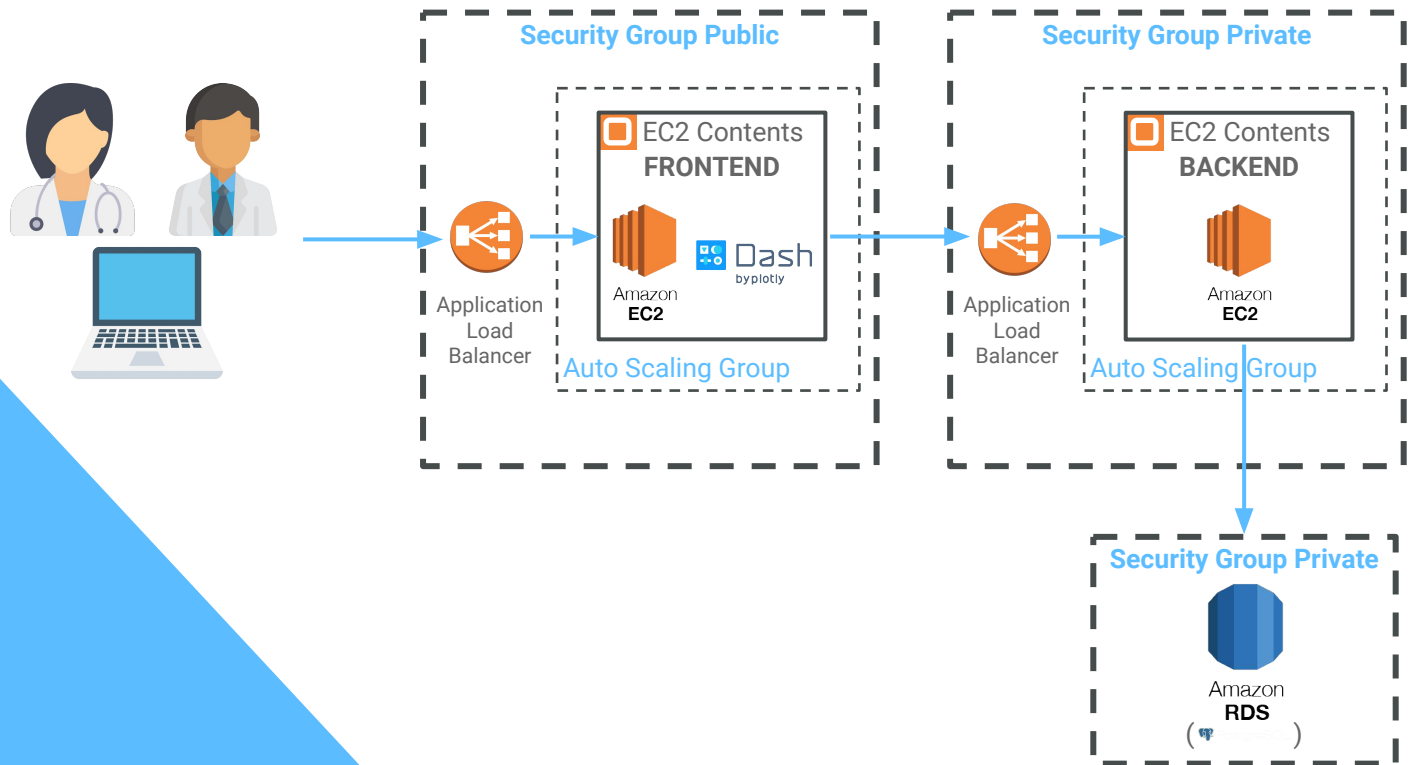


Figure 9. Web application architecture

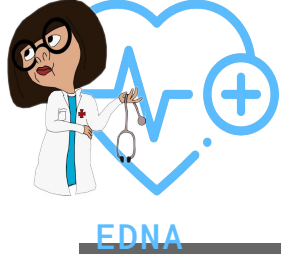


Source: own elaboration.

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, auto scaling 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.

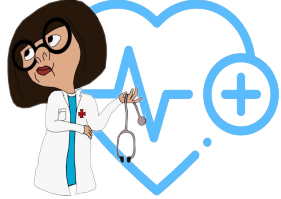
For the RDS, a security group was assigned which only allows traffic from the backend to protect the data hosted on it.



## 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.



EDNA

## **Back-end**

Back-end was created with the purpose of having access to the database in a secure way. For this, an application was created using Flask, which accesses the database through SQL queries and orders the data in a json file, so that the front-end can read them through API requests (this is possible through CORS).

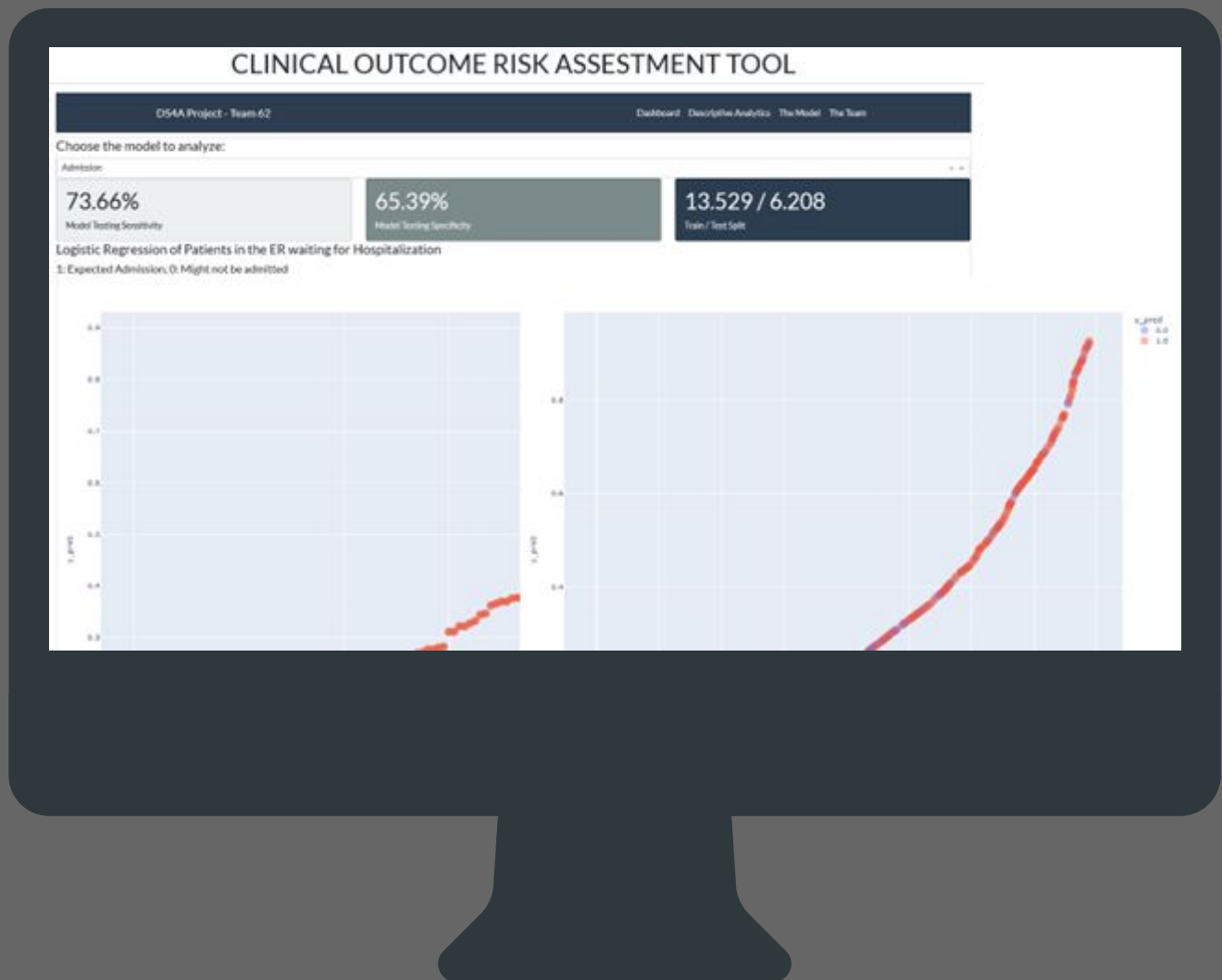
## **Relational Database Service**

For the RDS, a database was created in which all the records that the user wants to enter are saved, this to have consistency in the data, and to be able to read them constantly.

## **Elasticbeanstalk**

Deployment of the application was made using this AWS service, in which both the back-end and the front-end are deployed and communicate through http requests. This service was chosen due to its ease of implementation and advantages such as deploying the load balancer and auto scaling group, thus allowing connection from any device with internet access.

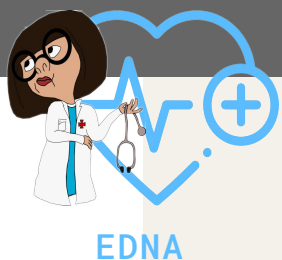
## Emergency Department Clinical Assistance Tool - EDNA is available as a functional web application in:



<http://ds4a-dev.us-east-2.elasticbeanstalk.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 possible steps

Team member's effort split:

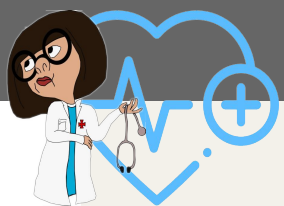


MA	JB	DC	JG	CR	LS
13%	20%	20%	18%	14%	15%



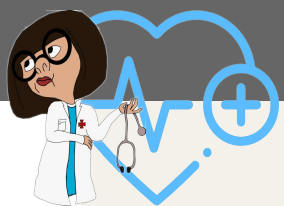
## Conclusions

- EDNA is a support tool for clinical and administrative decision making, whose intention is never to replace clinical concept or the TRIAGE system, but rather, to facilitate the operation of an Emergency Department featuring the use of AI to have better estimations of the capacity required for the next hour in hospitalization and whether or not the hospital will meet this quota.
- The classification model has been tuned for the highest sensibility without sacrificing the discrimination capabilities of the predictor, this was done with the aid of a field expert and iterating the different parameters through randomized search.
- EDNA works in real time. The ED coordinator can upload a file of the current patients in the ER and this is saved in our backend database. Today's patients predictions are shown right away.
- EDNA can be operated from anywhere in the world thanks to the design in our AWS based backend and it's programmed to scale automatically.



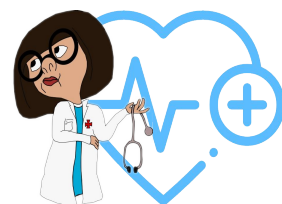
## Future work

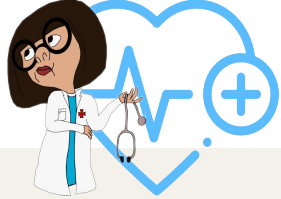
- The regression model could use improvement in performance, with an RMSE of 82.36 (minutes) the estimation of how long a patient will stay is not accurate enough for capacity decision making according to our clinical expert. One approach to improve this performance would be to include more variables of the patient, for example:
  - The ICD-10 diagnostic code of the patient as a category
  - The number of clinical events the patient had in the past month
  - A chronic disease multimorbidity index as described in Calderon et al (2016)
- For the classification models one of the improvements we suggest is to validate whether or not using data-balancing techniques improve the outcome of the classification as discussed [here](#). The hospitalization model could improve a lot as well from including the same variables suggested for the regression.
- The plots in the applications could be tailored further for the variables of interest of the ED coordinator.



## 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>.

**Calderón-larrañaga, A., Vetrano, L., Onder, G., Gimeno-feliu, L. A., Coscollar-santaliestra, C., Carfí, A., ... Fratiglioni, L. (2016).** Assessing and Measuring Chronic Multimorbidity in the Older Population : A Proposal for Its Operationalization. 00(00), 1–7. <https://doi.org/10.1093/gerona/glw233>.

---

#### **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  
EDNA logo illustration - Alejandro Guzmán Cano

#### **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)