Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Máster Universitario en Ingeniería Informática

Data Processes

Course project - Technical report

2021-22 | SEPT-FEB

Author/s:

Miguel Alonso, Carlos
Num. matrícula: 170243

Fernández Molleda, Lucía
Num. matrícula: 170312

Leira García-Baamonde, Manuel
Num. matrícula: 170136

January 9, 2022

Table of Contents

1 Data Preparation

The first thing we've done, and the one that has consumed most of out time, was cleaning the dataset, to delete any non-useful data, including errors and missing values.

First, the column `DESTINATION` has been deleted, since most of the values are **null**, and we had no use for it. The column `GLUCOSE` has been deleted too, since all of its values are 0.0, so it's of no use for our analysis.

There were some extreme outliers that couldn't be the source of real measurements, such as body temperature (`TEMP`) lower than 20 degrees, blood pressure over human limits (`BLOOS_PRES_SYS` and `BLOOD_PRES_DIAS`), etc. All this values has been deleted, since they make no sense, and a person in this condition would be dead.

The next step was to split the dataset on *categorical* and *numerical* sets, to properly treat any missing values on each one, and turn *categorical* values into *numerical* values, so the prediction model could use them.

ID	SEX	EXITUS
1	FEMALE	NO
2	FEMALE	NO
3	MALE	NO
4	MALE	NO
5	MALE	NO
...	...	...
2046	FEMALE	YES
2047	FEMALE	YES
2048	FEMALE	NO
2049	FEMALE	YES
2052	FEMALE	NO

[1551 rows x 2 columns]

(c) Columns of categorical values

ID	AGE	DAYS_HOSPITAL	DAYS_ICU	TEMP	HEART_RATE	SAT_02	\
1	15.0	4	0	37.0	0	92	
2	18.0	4	0	37.3	105	97	
3	21.0	7	0	38.5	112	95	
4	21.0	10	0	39.2	113	97	
5	22.0	4	0	36.3	80	92	
...	...	...	...	...	...	...	
2046	101.0	2	0	36.8	84	95	
2047	102.0	5	0	36.5	83	94	
2048	105.0	4	0	36.4	74	98	
2049	106.0	5	0	38.2	89	98	
2052	NaN	6	6	36.8	190	98	

ID	BLOOD_PRES_SYS	BLOOD_PRES_DIAS
1	0	0
2	0	0
3	85	47
4	0	0
5	111	70
...	...	...
2046	110	65
2047	150	65
2048	169	97
2049	143	63
2052	0	0

[1551 rows x 8 columns]

(d) Columns of numerical values

Figure 1: Slices of the categorical (a) and numerical (b) datasets.

To transform each *categorical* value into a *numerical* value, first, we must treat any missing ones. To do this, we've used the module *SimpleImputer*, from the package *sklearn*, applying the strategy of the **most frequent**. Then, we've assigned the next *numerical* values to the *categorical* values, and, since they're both binary values, this assigned *numerical* values are either 1 or 0:

- **SEX** $\Rightarrow$ Female $=0$ | Male $=1$
- **EXITUS** $\Rightarrow$ No $=0$ | Yes $=1$

On the other hand, the *numerical* values require no transformations, but the missing and strange values must be treated as-well.

Any missing values has been replaced with the **mean** of the values of its column, so the data is not artificially modified in excess. The strange values, in this case, measurements that cannot be real, such as 0 arterial oxygen saturation (`SAT_02`) or 0 blood pressure ((`BLOOS_PRES_SYS` and `BLOOD_PRES_DIAS`).

Once all the pre-processing has been completed, we merged both the *numerical* and the *categorical* sets to analyse the dataset as a whole and extract any valuable information.

2 Data Analysis

To perform a thorough analysis, we divided it in different sections, to make the analysis easier and extract conclusions build on the data we've seen during each section of said analysis.

2.1 Univariate analysis

Using only one variable, we can see their distribution, and make some statements about both the data and the problem we're facing.

Using the data given by the Figure 2, we can say that:

- There are more men than women entering ER.
- The mortality rate is around 16%.
- The age of the patients lays between near 60 and 100 years old, which might indicate that COVID19 affects the elderly more severely.
- The great majority of the patients stays in the hospital less than 20 days.
- A very low number of patients require ICU time.
- Many patients enter ER with fever and high heart rate (the fever might be an effect of COVID19, but the high heart rate might be caused by stress).
- Many patients enter ER with low arterial oxygen saturation, which might be caused by COVID19 attacking the lungs first.

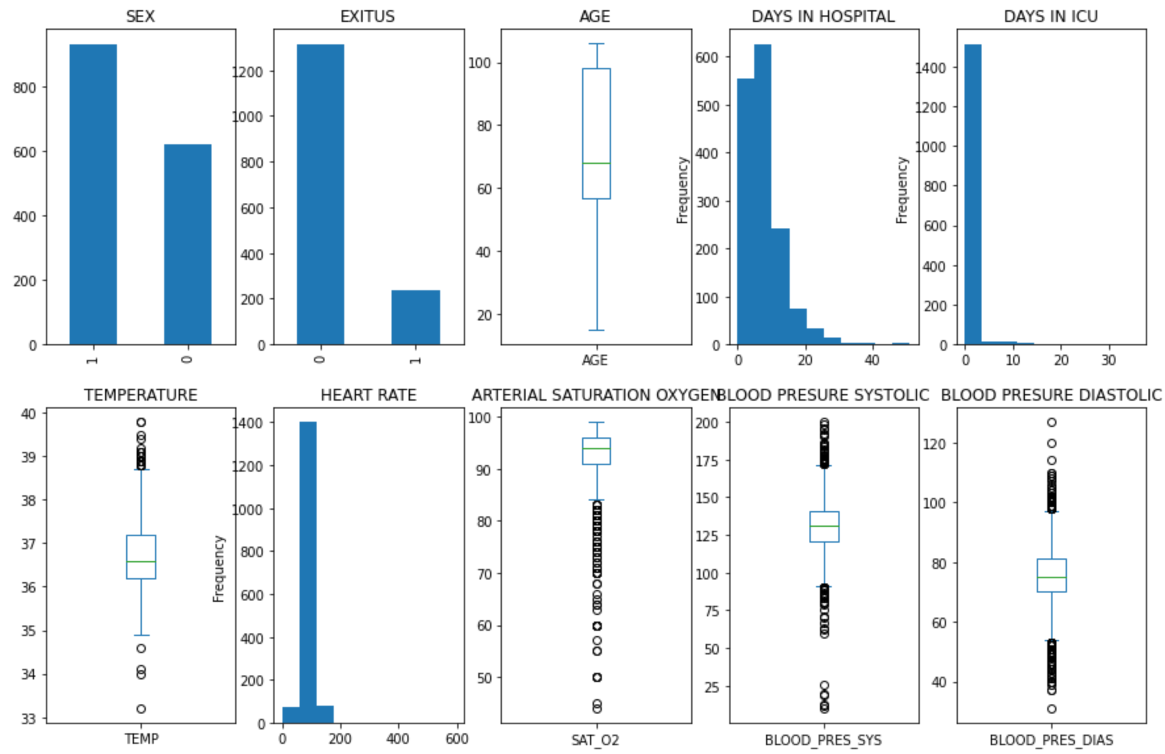

Figure 2: Plots of univariate analysis.

On Figure 3 we can see the great standard deviation of some variables related to the patients health condition, and that the age is a relevant factor on the exitus of a patient (older people are more prone to die). This can be observed on the percentiles, being the 25% the mark of 57 years old, which means that 75 % of patients that enter ER are 57 years old, or older.

	SEX	EXITUS	AGE	DAYS_HOSPITAL	DAYS_ICU	TEMP	HEART_RATE	SAT_O2	BLOOD_PRES_SYS	BLOOD_PRES_DIAS
count	1551.000000	1551.000000	1551.000000	1551.000000	1551.000000	1551.000000	1551.000000	1551.000000	1551.000000	1551.00000
mean	0.600258	0.152160	71.218065	7.948420	0.290135	36.735977	88.253385	92.613153	131.066502	75.09688
std	0.490003	0.359292	20.152968	5.433105	1.990532	0.813821	25.008765	6.422435	19.941586	11.27202
min	0.000000	0.000000	15.000000	0.000000	0.000000	33.200000	0.000000	44.000000	10.000000	31.00000
25%	0.000000	0.000000	57.000000	4.000000	0.000000	36.200000	78.000000	91.000000	121.000000	70.00000
50%	1.000000	0.000000	68.000000	7.000000	0.000000	36.600000	88.000000	94.000000	131.066502	75.09688
75%	1.000000	0.000000	98.000000	10.000000	0.000000	37.200000	100.000000	96.000000	141.000000	81.00000
max	1.000000	1.000000	106.000000	51.000000	36.000000	39.800000	593.000000	99.000000	200.000000	127.00000

Figure 3: Statistical analysis of each variable.

Looking at Figure 3 and Figure 2, we conclude that `HEART_RATE`, `BLOOS_PRES_SYS` and `BLOOD_PRES_DIAS`, might have no relation with the patient chance of death or ICU time.

2.2 Feature selection

Before proceeding with the bivariate analysis, we've use the *Chi2* and *KBest* (see Figure 4 and Figure 5) on each of the variables we are working on (`EXITUS` and `DAYS_UCI`) which chooses the best suited variables to analyse them.

```
             AGE  DAYS_HOSPITAL  HEART_RATE  SAT_02
ID
1      15.000000            4.0         0.0    92.0
2      18.000000            4.0       105.0    97.0
3      21.000000            7.0       112.0    95.0
4      21.000000           10.0       113.0    97.0
5      22.000000            4.0        80.0    92.0
...          ...            ...         ...     ...
2046  101.000000            2.0        84.0    95.0
2047  102.000000            5.0        83.0    94.0
2048  105.000000            4.0        74.0    98.0
2049  106.000000            5.0        89.0    98.0
2052   71.218065            6.0       190.0    98.0

[1551 rows x 4 columns]
```

Figure 4: feature selection for `DAYS_UCI`.

Through Figure 4 and Figure 5, we see that *Chi2* has decided that the best variables for analysing:

- `DAYS_ICU` are: `AGE`, `DAYS_HOSPITAL`, `HEART_RATE` and `SAT_02`
- `EXITUS` are: `AGE`, `HEART_RATE`, `SAT_02` and `BLOOD_PRES_DIAS`

We've used *Chi2* because it's the model we have more experience with and we think is the more useful for this particular dataset.

2.3 Bivariate analysis

Through the bivariate analysis we've been able to establish the relation or independence between some variables. To perform this analysis, we've chosen the variables `DAYS_UCI`, `BLOOS_PRES_SYS`

ID	AGE	HEART_RATE	SAT_02	BLOOD_PRES_DIAS
1	15.000000	0.0	92.0	75.09688
2	18.000000	105.0	97.0	75.09688
3	21.000000	112.0	95.0	47.00000
4	21.000000	113.0	97.0	75.09688
5	22.000000	80.0	92.0	70.00000
...	...	...	...	...
2046	101.000000	84.0	95.0	65.00000
2047	102.000000	83.0	94.0	65.00000
2048	105.000000	74.0	98.0	97.00000
2049	106.000000	89.0	98.0	63.00000
2052	71.218065	190.0	98.0	75.09688

[1551 rows x 4 columns]

Figure 5: Feature selection for EXITUS.

and BLOOD_PRES_DIAS on one hand, and AGE, EXITUS, SAT_02 and BLOOS_PRES_SYS, since that was the result of the feature selection.

On the Figure 6, we can see the correlation between DAYS_HOSPITAL and DAYS_UCI, and the weak or non-existen relation between DAYS_UCI and BLOOS_PRES_SYS and BLOOD_PRES_DIAS.

This relation should be evident, since patients that require time on the ICU, will require more time in a hospital to fully recover, mainly because COVID19 hit them harder than other patients.

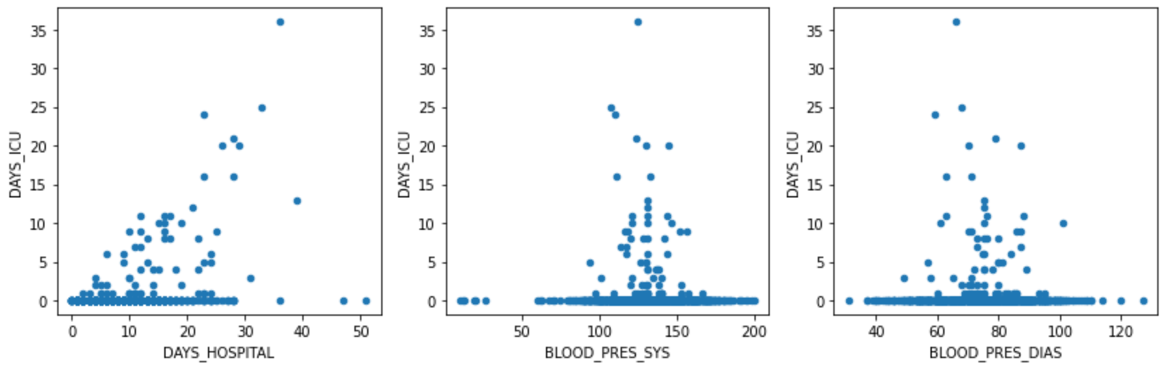

Figure 6: Bivariate analysis relevant variables related to DAYS_UCI.

On Figure 7, we see that, just as we said previously, older patients die more, maybe related to a more fragile health and additional health problems. Also, we see that systolic blood pressure has no effect on the death rate, just as we saw in the Figure 3.

Figure 7 also shows that arterial oxygen saturation SAT_02 has some relation with patients exitus EXITUS, since patients that end up dying enter ER with lower arterial oxygen saturation.

To confirm this suppositions, we have to see the results of a Pearson analysis, which will tell us the relation between different variables.

As we can see on the Figure 8, there are some important correlations between some variables:

- The age (AGE) of the patient has a strong relation with its exitus (EXITUS) (0.38), which confirms our assumptions, that the older a patient is, the more chances there are for it to die.

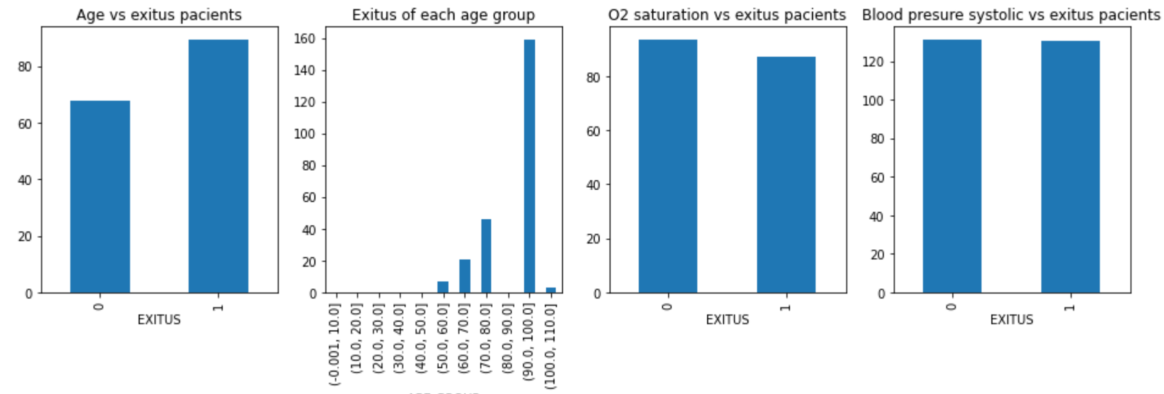

Figure 7: Bivariate analysis relevant variables related to `EXITUS`

	SEX	AGE	DAYS_HOSPITAL	TEMP	HEART_RATE
SEX	1.000000	-0.128772	0.084339	0.119245	0.025697
AGE	-0.128772	1.000000	0.014289	-0.141241	-0.231035
DAYS_HOSPITAL	0.084339	0.014289	1.000000	0.137942	0.004184
TEMP	0.119245	-0.141241	0.137942	1.000000	0.139281
HEART_RATE	0.025697	-0.231035	0.004184	0.139281	1.000000
SAT_O2	-0.069466	-0.183792	-0.137208	-0.043377	-0.072278
BLOOD_PRES_SYS	0.052269	0.135023	-0.007430	0.041930	0.011008
BLOOD_PRES_DIAS	0.115121	-0.135203	-0.010455	0.033861	0.179048
DAYS_ICU	0.064084	-0.052427	0.360571	0.065479	0.037584
EXITUS	0.041553	0.383271	-0.012502	-0.001965	-0.025259

	SAT_O2	BLOOD_PRES_SYS	BLOOD_PRES_DIAS	DAYS_ICU	EXITUS
SEX	-0.069466	0.052269	0.115121	0.064084	0.041553
AGE	-0.183792	0.135023	-0.135203	-0.052427	0.383271
DAYS_HOSPITAL	-0.137208	-0.007430	-0.010455	0.360571	-0.012502
TEMP	-0.043377	0.041930	0.033861	0.065479	-0.001965
HEART_RATE	-0.072278	0.011008	0.179048	0.037584	-0.025259
SAT_O2	1.000000	-0.011745	0.021552	-0.212458	-0.353319
BLOOD_PRES_SYS	-0.011745	1.000000	0.488299	-0.027271	-0.018859
BLOOD_PRES_DIAS	0.021552	0.488299	1.000000	-0.017646	-0.104115
DAYS_ICU	-0.212458	-0.027271	-0.017646	1.000000	0.104217
EXITUS	-0.353319	-0.018859	-0.104115	0.104217	1.000000

Figure 8: Pearson correlation coeficients

- There's also a strong correlation between the number of days a patient stays at the hospital (`DAYS_HOSPITAL`) and the number of days it stays on the ICU (`DAYS_ICU`), which also confirms our previous assumption, that the more days a patient stays on the ICU, the more days it will stays on the hospital.
- The arterial oxygen saturation (`SAT_O2`) and the exitus (`EXITUS`) also have a strong correlation, which also confirms our previous assumption, that, since COVID19 attacks the lungs first, the patients present a low O2 saturation, which is worse on the patients that are infected severely.

Through this analysis, we've manage to extract the following conclusions related to our goals:

- The age and the arterial oxygen saturation of the patient are key factors on its survival rate.
- This factors have some relation with the amount of time the patient will stay at the hospital and at the ICU.

With this data, we have developed a predictive model, to predict, based on the patients condition, if it will survive and if it will require time on the ICU.

2.4 Kaplain-Meier curves

We have implemented two *Kaplain-Meier* curves of survival rate (see Figure 9 and Figure 10, but we couldn't read what the curves were showing us, since we have no timeline, we couldn't figure out how to read them.

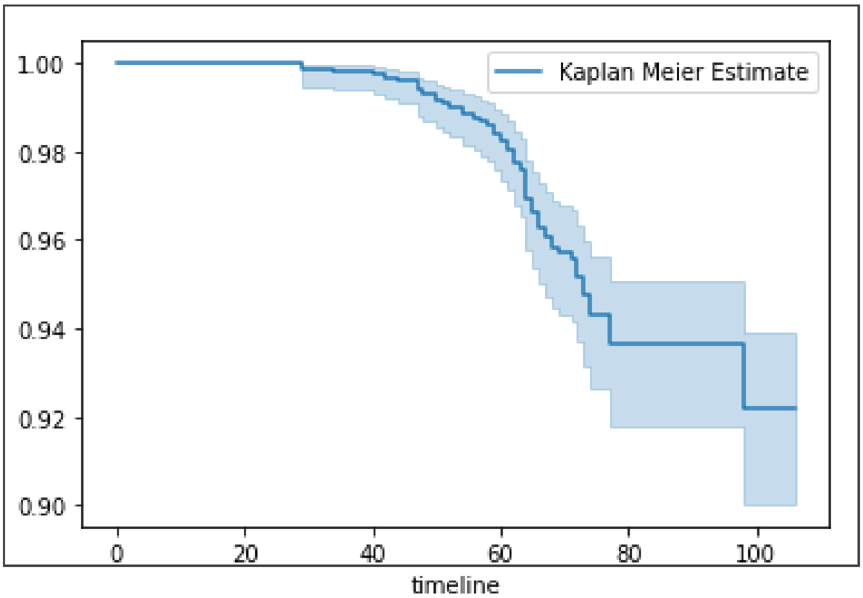

Figure 9: Kaplain-Meier curve for `ICU` and `AGE`

3 Prediction Models

For the prediction model, we've used a **Decision Tree** to predict the exitus (`EXITUS`) of a patient based on its condition, and a **Simple Regression Tree** to predict the number of days a patient will stay on the ICU (`DAYS_ICU`).

This models (Figure 12) will read the dataset produced as a result of the pre-processing phase, an the will divide it into two sets, one for training, and one for testing. On both models, the

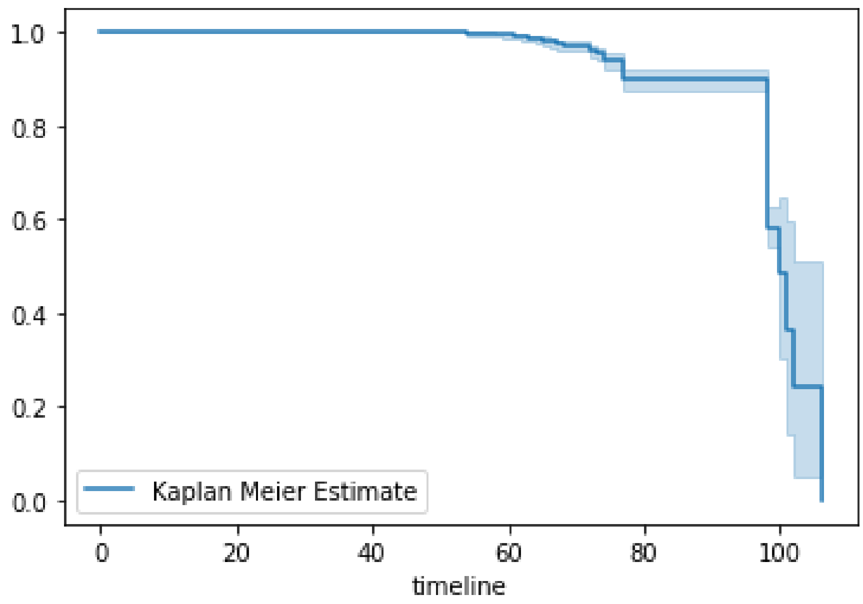

Figure 10: Kaplain-Meier curve for `EXITUS` and `AGE`

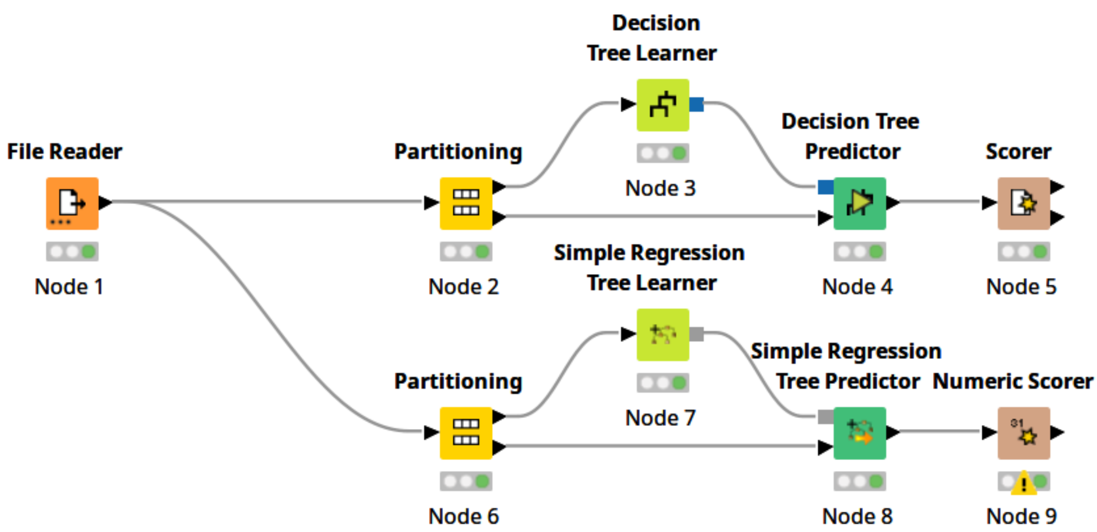

Figure 11: *KNIME* graph of the prediction models.

training set is chosen randomly (to avoid bias), and the remaining values will be used to test the predictions made by each model.

EXITUS \ ...	0	1
0	126	8
1	10	12

Figure 12: Confusion matrix of the `EXITUS` prediction.

The confusion matrix (see Figure 12) shows that this model has an **accuracy of 86 %**, classifying 138 elements correctly, which means that it might need more values to adjust its predictions, but proves that this model is correct and its predictions could be useful on a real situation.

On the other hand, the **Simple Regression Tree** used to predict the number of days a patient will be on the ICU (`DAYS_ICU`), after training and testing, has the results shown in Figure 13.

R^2:	-4.225
Mean absolute error:	0.346
Mean squared error:	2.718
Root mean squared error:	1.649
Mean signed difference:	0.128
Mean absolute percentage error:	NaN
Adjusted R^2:	-4.225

Figure 13: Statistics of the `DAYS_ICU` prediction

The result for **R2 of -4.225** tells us that the model is not properly adjusted to predict the intended result, which, in addition to this, a **mean squared error of 2.718**, which denotes the high rate of bad predictions of the model.

Due to our little-to-no knowledge of other predictive models, we've been unable to find another that adjusts properly to the intended predictions.

4 Results

The results of this project reflect the fact that the age of the patients, and their arterial oxygen saturation are key factors to their survival, since older patients, and patients with a low level of arterial oxygen saturation are the ones more likely to die. The predictive model for a patient exitus has shown, that, with a little more training, could predict this outcome based on the conditions of the patient when arrives at the ER.

Despite this good results, the same could not be said of the prediction of the number of days a patient might spend at ICU, since the predictive model does not adjust to out requirements. However, analysing the data has shown that older patients are more likely to end up on the ICU, due to the high rate of mortality between the patients of this age groups.