Lab

Assignment 1

Bana Sadi

ID: 206611477

Almog Bin Simon

ID: 206571135

Link for the GitHub:

https://github.com/almog2065/Sepsis

1 Executive summary

Sepsis is a potentially life-threatening condition when the body's response to an infection causes widespread inflammation. It can lead to organ failure, septic shock, and even death if not treated promptly and appropriately. Although sepsis is currently the primary cause of death, its symptoms can be triggered by various illnesses, making it difficult to diagnose.
For this task, we have been given files for 20,000 patients, where each file contains medical and demographic information. Each row of the file corresponds to an hourly measurement of these attributes. Our purpose is to train a model in order to predict whether a patient is likely to suffer from sepsis in a 6 hours period.
In our analysis, we take into account medical and demographic measures deemed important for sepsis detection in the literature in addition to features found significant in the data analysis presented in this document. Three algorithms were used and compared for prediction purposes: random forest, neural network, and XG-Boost.

2 Exploratory Data Analysis

2.1 Features Description:

The data set contains 41 features, for each of those features, a brief explanation, classification as discrete or continuous, and value range are provided in the table below:

Feature	Description	Range	Continuous or Discrete
HR	Heart rate (beats per minute)	[20.0, 280.0]	Continuous
O2Sat	Pulse oximetry (%)	[20.0, 100.0]	Continuous
Temp	Temperature (Deg C)	[20.9, 50.0]	Continuous
SBP	Systolic BP (mm Hg)	[20.0, 299.0]	Continuous
MAP	Mean arterial pressure (mm Hg)	[20.0, 300.0]	Continuous
DBP	Diastolic BP (mm Hg)	[20.0, 300.0]	Continuous
Resp	Respiration rate (breaths per minute)	[1.0, 100.0]	Discrete

EtCO2	End tidal carbon dioxide (mm Hg)	[10.0, 100.0]	Continuous
BaseExcess	Measure of excess bicarbonate (mmol/L)	[-32.0, 49.5]	Continuous
HCO3	Bicarbonate (mmol/L)	[0.0, 55.0]	Continuous
FiO2	Fraction of inspired oxygen (%)	[-50.0, 4000.0]	Continuous
pH	N/A	[6.62, 7.78]	Continuous
PaCO	Partial pressure of carbon dioxide from arterial blood (mm Hg)	[10.0, 100.0]	Continuous
SaO2	Oxygen saturation from arterial blood (%)	[23.0, 100.0]	Continuous
AST	Aspartate transaminase (IU/L)	[5.0, 9961.0]	Continuous
BUN	Blood urea nitrogen (mg/dL)	[1.0, 268.0]	Continuous
Alkalinephos	Alkaline phosphatase (IU/L)	[7.0, 2528.0]	Continuous
Calcium	(mg/dL)	[1.0, 27.9]	Continuous
Chloride	(mmol/L)	[74.0, 145.0]	Continuous
Creatinine	(mg/dL)	[0.1, 41.9]	Continuous
Bilirubin_direct	Bilirubin direct (mg/dL)	[0.01, 35.0]	Continuous
Glucose	Serum glucose (mg/dL)	[10.0, 952.0]	Continuous
Lactate	Lactic acid (mg/dL)	[0.2, 31.0]	Continuous
Magnesium	(mmol/dL)	[0.6, 9.8]	Continuous
Phosphate	(mg/dL)	[0.3, 17.6]	Continuous
Potassium	(mmol/L)	[1.3, 27.5]	Continuous
Bilirubin_total	Total bilirubin (mg/dL)	[0.1, 49.6]	Continuous
TroponinI	Troponin I (ng/mL)	[0.01, 440.0]	Continuous
Hct	Hematocrit (%)	[8.8, 71.7]	Continuous
Hgb	Hemoglobin (g/dL)	[2.6, 25.0]	Continuous
PTT	partial thromboplastin time (seconds)	[17.1, 250.0]	Continuous
WBC	Leukocyte count (count*10$\hat{3}$/µL)	[0.1, 440.0]	Continuous

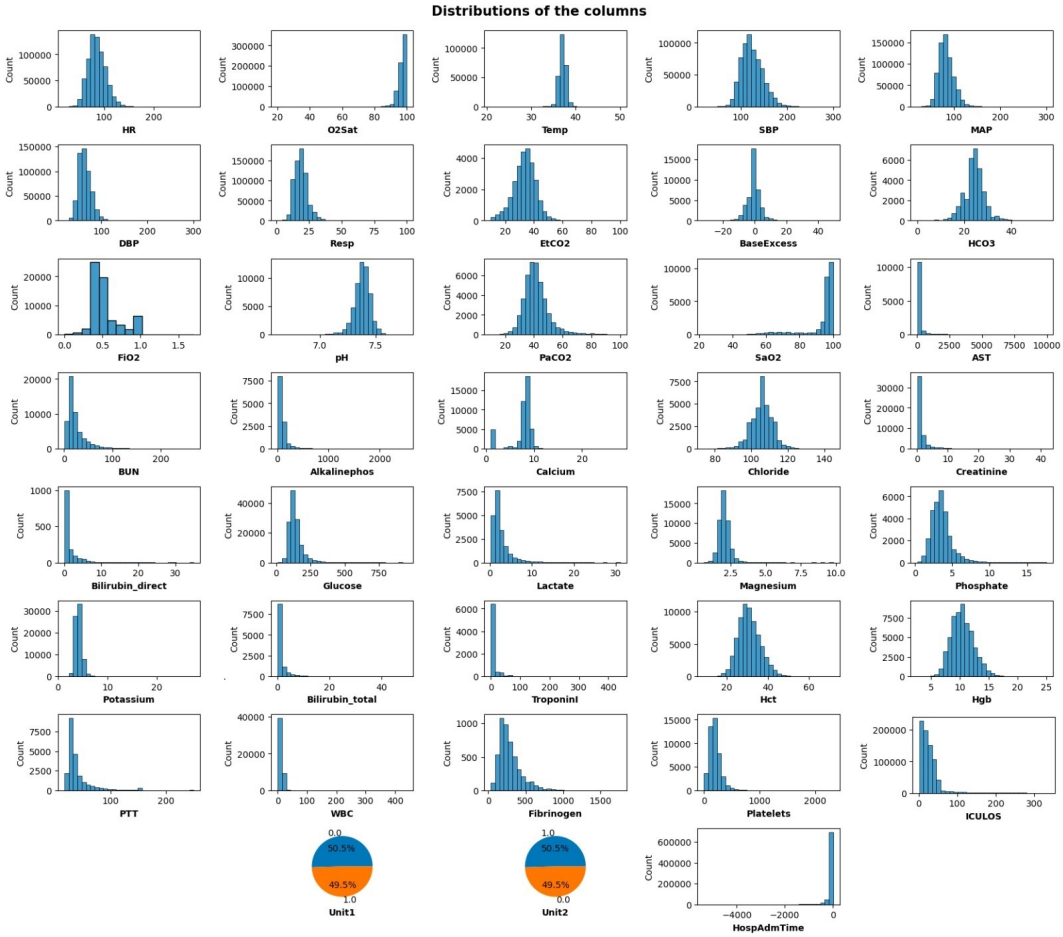
Fibrinogen	(mg/dL)	[35.0, 1760.0]	Continuous
Platelets	(count*10$\hat{3}$/µL)	[2.0, 2322.0]	Continuous
Age	Years (100 for patients 90 or above)	[15.0, 100.0]	Discrete
Gender	Female (0) or Male (1)	{0, 1}	Discrete
Unit1	Administrative identifier for ICU unit (MICU)	{0, 1}	Discrete
Unit2	Administrative identifier for ICU unit (SICU)	{0, 1}	Discrete
HospAdmTime	Hours between hospital admit and ICU admit	[-5366.86, 17.34]	Continuous
ICULOS	ICU length-of-stay (hours since ICU admit)	[1.0, 336.0]	Discrete
SepsisLabel	For sepsis patients, SepsisLabel is 1 if t $\leq$ tsepsis 6 and 0 if t $<$ tsepsis 6. For non-sepsis patients, SepsisLabel is 0	{0,1}	Discrete

2.2 Inspecting the features distribution, comparative analysis between features

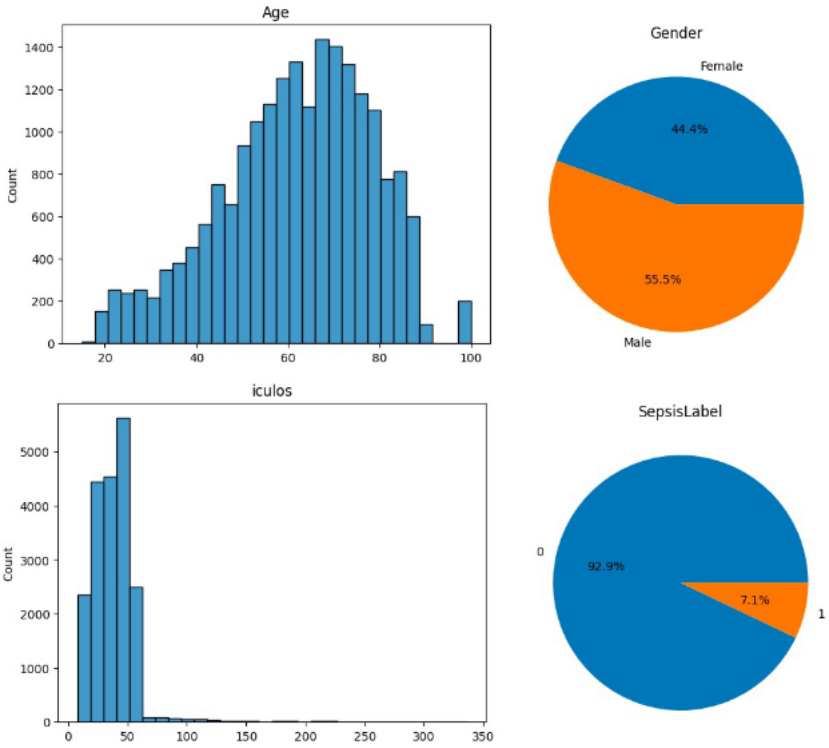
In this section, we will display histograms, different analyses, and hypothesis tests that helped us navigate the data and determine which features to use.

2.2.1 Big Picture

First, in order to get a general picture of the distribution of each feature we concatenated the data of all the patients into one data frame. The following plots show the distribution of each feature:



This method is not very informative for features Gender, Age, SepsisLabel (i.e. number of patients with sepsis) and ICULOS (number of hours spent in the ICU). In the following plots we counted each patient only once:



2.2.2 Medically Important Features

We chose to focus on features that are medically significant for sepsis early identification in hospitals and are easily obtained from a simple blood test. According to medical literature, if two of the following features are within a certain range, it could indicate a greater risk of sepsis:

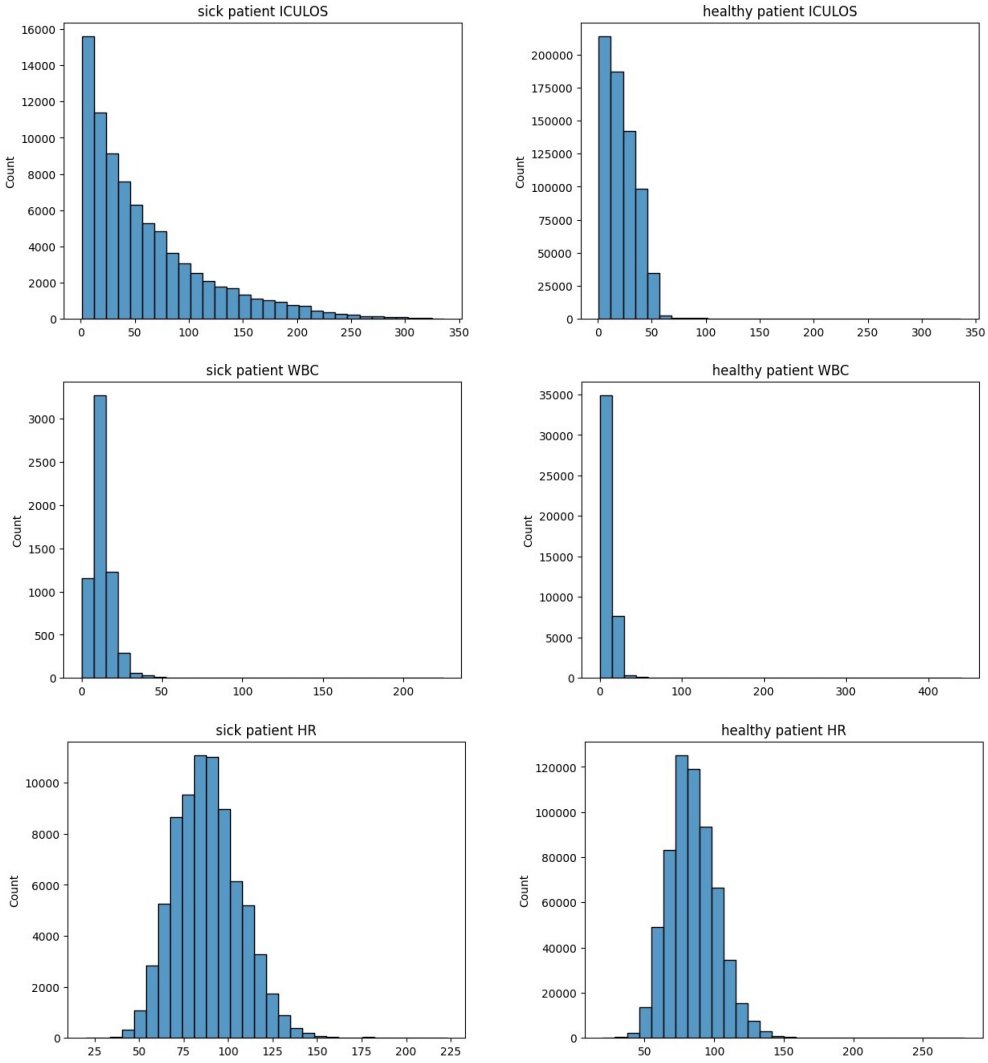
Temperature	$< 36C\,(96.8F)\,or\, > 38C\,(100.4F)$
Heart rate	$> 90/min$
Respiratory rate	$> 20/min$
WBC	$< 4x109/L\,(< 4000/mm3),\, > 12x109/L\,(> 12,000/mm3)$

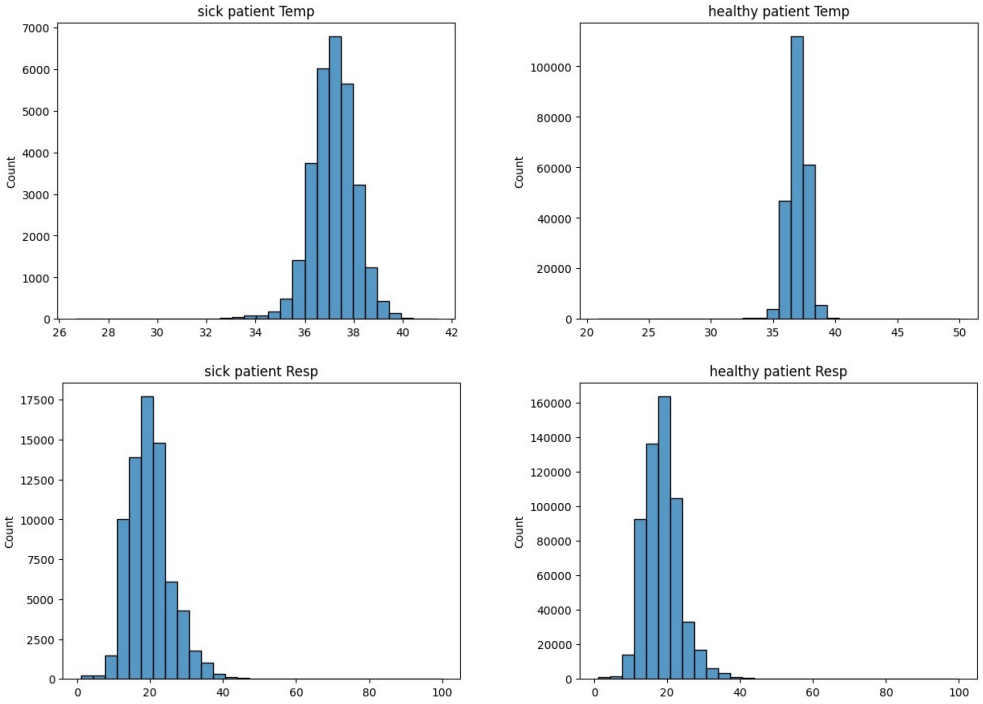
[1]

As our prediction occurs six hours before diagnosis, it's important to verify if those measures are applicable to patients at that specific time. To determine whether there are significant differences in certain features between patients with and without sepsis, we need to conduct hypothesis testing.

[1] table was taken from Wikipedia and confirmed with other medical papers

Initially, we will display histograms that compare the four features between the group of individuals diagnosed with sepsis and the group that is not diagnosed with sepsis. Additionally, testing and training algorithms showed that ICULOS has a significant impact on the algorithms' performance thus we will be considering it here as well.





It is evident that each of the features displays different statistical behavior for the two populations. The following table summarises means and SD:

value	ICULOS	Temp	HR	WBC	Resp
sick mean	60.095187	37.139844	88.398086	12.912542	19.909332
sick sd	56.909826	0.907249	18.722675	9.600269	5.855446
healthy mean	22.513677	36.956169	84.197185	11.247779	18.607007
healthy sd	18.284782	0.756490	17.166226	7.115355	4.991442

We want to use statistical hypothesis testing, to check whether the expected value and variance of the different features displayed above are indeed different among the two different categories (patients with sepsis and without sepsis). This is important because we want to check differences over time not only at the time of septic shock. Since we are not sure that our data is normally distributed we will use the Mann-Whitney test in order to decide whether each of the features comes from a different distribution in each of the populations.

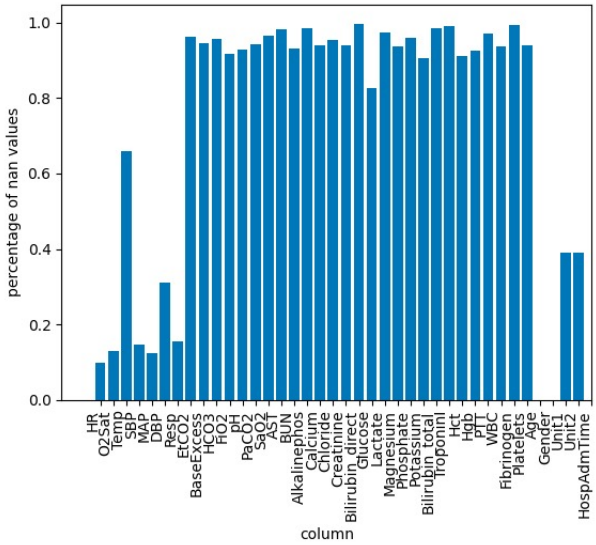
We separated the data of the two populations and conducted the Mann-Whitney U two-sided test. The null hypotheses for each feature is that the underlying distribution is identical in both populations, the alternative hypothesis is that the distribution of the values for each of the features is different between the populations. For each feature, the p-value for is presented in the table below.

Feature	p-val
Temperature	$6.33e-05$
Heart rate	$9.83e-24$
Respiratory rate	$3.80e-24$
WBC	0.085
ICULOS	$2.04e-23$

We can easily see that with 95 % confidence level all null hypothesis are rejected with an exception of the WBC where the null hypothesis can be rejected only with 90% confidance level.

2.3 Missing Data

First, we created a histogram describing the percent of missing data per feature:



It is evident that a significant proportion of the features contain missing data. In an attempt to address this issue, we employed various methods to fill in the missing values. First, we tried to fill in all missing values with the same constant, this method gave good results. However, we wanted to fill the data in a more meaningful way while addressing the nature of the data therefore, in a different attempt, we used backward and forward filling for patients who had

at least one value in a column. For patients who had no data available for a given feature, we used mean imputation, where the mean was calculated based on data from all the patients in the dataset. For categorical data, 'Unit 1', and 'Unit 2' we set NaN columns to 0 and filled Nan's for patients with partial information based on back and forward filling as well.

3 Feature Engineering

3.1 Which Features we will be using:

We checked two approaches in feature selection, in the first approach we used all the available features to train each of the models. Our second approach, after reading several papers on the subject, was to only address direct symptoms of sepsis, the most important ones being Temperature, Heart rate, Respiratory Rate, and WBC in addition to one transformation based on those 4 features (see section 3.3).

3.2 Features transformations and Data enrichment

We added a SIRS (Systemic inflammatory response syndrome) categorical feature. This feature received values {0,1,2,3,4} depending on the number of medical symptoms observed in that row that fall within the SIRS range.

4 Prediction

We are faced with a binary classification problem, we tried building models using Random Forest XG-Boost and Neural-Network. We also tried using LSTM however that did not give significant results. For each algorithm, we will explain which features worked better and present several results.
A Note on the input: since the data is very imbalanced, there are a lot more people without sepsis than those with sepsis and for those with sepsis most rows have a 0 SepsisLabel We decided to represent each patient with only one row and not a whole table by taking the last row for patients with no sepsis and the first row with label 1 for patients with sepsis.

4.1 Algorithm 1: Random Forest

We used the sklearn RandomForestClassifier. A combination of different hyperparameters was tested we eventually chose the following parameters:

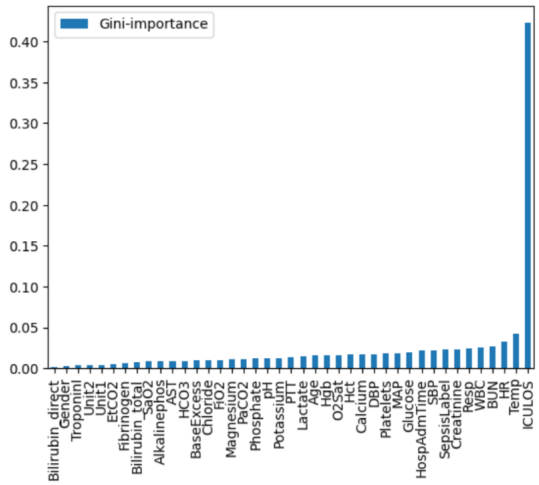
Number of trees: 300
Maximum depth of each tree: 40
Minimum number of samples required to split an internal node
Minimum number of samples required to be at a leaf node: 3
Random state : 2
The rest of the parameters were set to their default values.

4.1.1 Training and validation results and Post Analysis:

We will present a summary of important measures after training and testing the algorithm first on all available features and second on medically important features in addition to feature transformations.

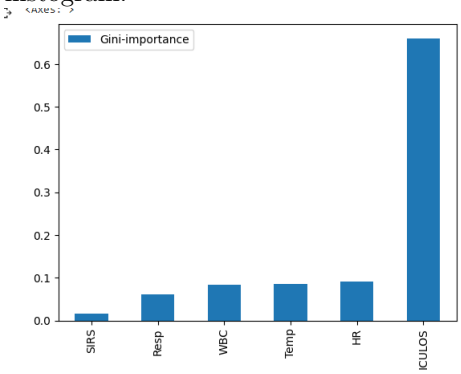
measure	all features	medically important features
F1	0.6616	0.6704
Precision	0.8828	0.838
Recall	0.5290	0.5587
Accuracy	0.9599	0.9593

After running the algorithms using all the features we extracted a histogram showing the importance of each feature:



It is evident that the four medically important features are among the leading features in the histogram. Additionally, we can't dismiss the high importance of ICULOS thus, we added to the four important features to check the models' performance.

After training the algorithm on the subset of the medically important features we can see an overall increase in the different measures, however the SIRS measure turned out to be less efficient than expected as we can see in the following histogram:



4.2 Algorithm 2: Neural Network

We build a Neural Network with torch library. First we build a dataset object with torch.dataset to train the model right, and we build this architecture of neural network:
linear layer of size (40, 200) relu activation linear layer of size (200, 100) relu activation linear layer of size (100, 1) sigmoid activation
We also used hyperparameters of:
Adam optimizer with learning rate of 0.001 20 epochs
The training process was with batch size 1 and the inference is when the network output was higher than 0.5, so the label is 1, and less than 0.5 it is labeled to 0. This architecture gives us the highest f1 score vs others architectures, we tried to change hyper-parameters, and the training process, and also deeper network but this one gives the highest score.

4.2.1 Training and validation results and Post Analysis:

The results show that with fewer features, and more important ones the neural network learns and predicts better, but with close f1 score. So the results show us that the features we think are important, are really important, and the other features are some noise.

measure	all features	medically important features
F1	0.5249	0.5313
Precision	0.9042	0.8396
Recall	0.3697	0.3886
Accuracy	0.9504	0.949

We also tried testing the neural network on two subgroups of the data - males and females in order to see if the network would perform differently. The results are displayed in the table below:

measure	male	female
F1	0.5565	0.6503
Precision	0.8902	0.7227
Recall	0.404	0.7227
Accuracy	0.9522	0.9529

Although the network gave different results we can see that the main difference is in the recall measure. We did not perform this analysis with other algorithms since Gender had very little importance in the their performance as we show in the histograms under each one.

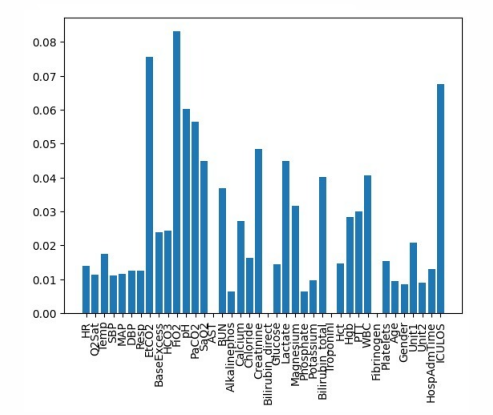
4.3 Algorithm 3: XG-Boost

We used xgboost algorithm from xgboost library. A combination of different hyper-parameters was tested we eventually chose the following parameters:
Maximum depth of each tree: 60
Minimum weight of child node in tree: 10
Fraction of samples to be used for each boosting round: 0.8
Fraction of features to be used for each boosting round: 0.8
learning rate: 0.1
Boosting rounds to perform: 1200
The balance of positive and negative weights: number of 0 / number of 1
evaluation metric: AUC
loss function: binary logistic
The rest of the parameters were set to their default values

4.3.1 Training and validation results and Post Analysis:

measure	all features	medically important features
F1	0.7131	0.6397
Precision	0.7720	0.6862
Recall	0.7131	0.5991
Accuracy	0.9605	0.95

Contrary to our expectation, this algorithm performed worse on the test set when trained only on the medically important features. We want to note that we received an even higher f1 score, **0.73**, when we trained the algorithm on the training data where we filled the Nan values with a constant.
We also extracted a feature importance histogram to see if the medical features stood out however the histogram shows that many other features have high importance that equals or even exceeds that of the four vital signs specified in section2.



Summary:

Overall, we tried to process the data of 20,000 patients, focusing on the medically important vital signs in order to train three different models using RandomForest, XGBoost and Neural Network. Although we are dealing with highly imbalanced data with a high percentage of missing data we managed to extract the important features and train algorithms that gave more than 0.5 F1 score.

5 References

we mainly used the papers proposed in the HW pdf.

- Wikipedia
- https://www.hindawi.com/journals/jhe/2019/5930379/conclusions
- https://pubmed.ncbi.nlm.nih.gov/26903338/