BEN-GURION UNIVERSITY OF THE NEGEV
FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF INDUSTRIAL ENGINEERING AND MANAGEMENT

Machine Learning and Data Mining Course

Algorithms Comparison for Reducing False Alarms in Intensive Care Units

FINAL REPORT

Lecturer: Prof. Boaz Lerner

Submitted by:
Liel Cohen
Gal Hever

28/02/2017

Abstract

Background: Over the past two decades, high false alarm (FA) rates have remained an important yet unresolved concern in the Intensive Care Unit (ICU). High FA rates lead to desensitization of the attending staff to such warnings, with associated slowing in response times and detrimental decreases in the quality of care for the patient (Sendelbach & Funk, 2013).

Method: A large multi-parameter patient database produces in Ichilov ICU was used to investigate the frequency of a set of 16 alarm definitions, namely: Tachycardia, Tachycardia-Hypotension, Bradycardia-Hypotension, Bradycardia, Ventricular tachycardia with hypotension, Obstructive shock 1, Obstructive Shock 2, Agitation 1, Agitation 2, LV shock 1, LV shock 2, Hypovolemia 1, Hypovolemia 2, Hypovolemia 3, SVT & Hypotension and SVT & Hypotension 2.

A Dataset of 73 patients records was extracted from a database containing a total of 8,273 patients and 781,229,536 records, using the associated 505 hours of simultaneous measuring of heart rate (HR), arterial pressure systolic (ArtBPS), arterial pressure mean (ArtBPM), central venous pressure (CVP), pulmonary artery pressure diastolic (PAPD), total respiratory rate (RR Total) mechanical respiratory rate (RR Mandatory), Spo2, ST and Fio2.

The data in this study was divided into train and test sets. The models were trained and tuned using CV-4 (cross-validation) over the train set, and the final models were evaluated over the test set. The evaluation metric we've used for comparison of the algorithms results was a weighted measure balancing sensitivity and specificity measures.

Results: The tested random forest (RF) full (12 variables) model including 11 variables for each tree and 50 trees has shown the highest sensitivity of 100%, specificity of 99.97% and weighted measure of 99.99%. This model was the best predictor for explaining the binary dependent variable, which indicated if a sample should have raised an alarm or not.

Keywords: False alarms; False alarm reduction; Machine learning; Statistical process control; Intensive care unit.

Tables of Contents

List of Terms and Abbreviations

Term	Description
ICU	Intensive Care Unit
Ichilov	Medical Information Mart for Intensive Care (previously Multiparameter Intelligent Monitoring in Intensive Care)
FA	False alarms
ML	Machine Learning
MSPC	Multivariate Statistical Control Chart
NN	Neural Network
RF	Random Forest
CV	Cross Validation
ArtBPS	Arterial Pressure Systolic
ArtBPM	Arterial Pressure Mean
CVP	Central Venous Pressure
PAPD	Pulmonary Artery Pressure Diastolic
HR	Heart Rate
ST1	ST Segment 1
ST2	ST Segment 2

Term	Description
ST3	ST Segment 3
RR Mandatory	Mechanical Respiratory Rate
RR Total	Total Respiratory Rate
Spo2	Peripheral Capillary Oxygen Saturation
Fio2	Fraction of Inspired Oxygen

Business understanding

Intensive care monitoring

Monitors are ubiquitous and found in all modern hospitals used at patient bedside in intensive care unit (ICU). These monitors has an alert mechanism in order to alert clinicians to deviations from a predetermined normal status when a patient's condition is deteriorating and when a device is not functioning well. Audible alarms are intended as an effective measure, while their main purpose is to enhance safety (Sendelbach & Funk, 2013). The reason clinical alarms are necessary is because they can attract operator's attention, when he or she is engaged in other tasks. Alarms tend to proliferate in almost all safety-critical areas where workload is very high (Kristensen, Edworthy, & Denham, 2015).

Critical Problems in Monitoring

Alarms have an important role in man-machine interface, however, in some cases alarms are the source for critical problems. Sometimes there may be so many different alarms at the same time that it make it difficult for the clinicians to quickly identify the underlying condition (Sorkin, 1998). In some cases, the auditory signal is too loud and has a shrill and aversive sound that interfere clinicians to communicate each other at the time they are most necessary (Kristensen et al., 2015). ICU alarms produce sound intensities above 80 dB that can lead to sleep deprivation and stress for both patients and staff. Furthermore, such disruptions depressed immune systems and been shown to have an important effect on recovery and length of stay in ICU, cortisol levels have been shown to be elevated (reflecting increased stress), and sleep disruption has been shown to lead to longer stays in the ICU (Berg, 2001).

Another problem with alarms are often criticized for generating an excessive number of false alarms (Aboukhalil, Nielsen, Saeed, Mark, & Clifford, 2008). By design, alarms are highly sensitive so that they do not miss an important event. However, this high sensitivity is achieved at the expense of specificity. As alarm limits become more sensitive and less specific, more false alarms are generated (Sendelbach & Funk, 2013).

False alarms can be caused either by noise and artifacts in signals or by inappropriate alarming criteria that are too generic and sensitive (Hu et al., 2012). In these cases, the accuracy depends on both limited sensitivity and the application of the clinician decision process. Contributions of these two components of performance based on signal detection theory (SDT). (Pashler, 2004)

Sendelbach & Funk (2013) have demonstrated that 72% to 99% of clinical alarms are false in their research. Another research of Salas-Boni, Bai & Hu (2015) reported of 88.8% false alarm rates. Frequent false alarms in the ICU have a negative impact on patients and staff. Multiple false alarms can led to phenomenon called alarm fatigue for bedside caregivers in ICU environments. Alarm fatigue is commonly defined as desensitization to alarm sounds. This happens when clinicians are exposed to an excessive number of alarms and have a sensory overload. It can cause to miss raising serious patient safety concerns and do not pay attentions to critical alarms (Hu et al., 2012). A repeated series of false alarms which eventually causes people to ignore the important alarms is called a 'cry- wolf' syndrome which leads to a lack of faith in the system and a casual attitude towards the constant presence of certain alarms (Sorkin, 1998). False alarms can lead also to care disruption, which impact both the patient and the clinical staff through noise disturbances (Aboukhalil et al., 2008).

Another problem with false alarms is additional burden on caregivers. ICU is a high workload environment, which means that caregivers will not have the time to treat each output from the alarm system, especially if most of the alarms indicate a normal condition (Sorkin, 1998).

The resulting of these problems being reported in the medical world and find expressions in different behaviors. A common sense analysis indicates that busy caregivers will adopt a strategy that ignores or discounts alarms from systems that have excessive false alarm rates (Sorkin, 1998). Another strategy is turn off audible alarms or turn down the volume of alarms. Some caregivers tend to delay in response and being desensitize to alarms leading to decreased quality of care (Kristensen et al., 2015). Such cases has resulted in sentinel events and patient deaths (Sorkin, 1998).

During the last decade, alarm hazards have been increasingly recognized as a major problem in the medical world. ECRI (Emergency Care Research Institute, US) recently announced its 2015 Top 10 Health Technology Hazards list. The top priority for the fourth year in a row was clinical alarm hazards. This fact demonstrates very well the severity of the problems associated with alarms at hospitals (Kristensen et al., 2015).

Data preparation

A large multi-parameter database of Ichilov ICU was used to investigate the frequency of a set of 16 alarm definitions, namely: Tachycardia, Tachycardia-Hypotension, Bradycardia-Hypotension, Bradycardia, Ventricular tachycardia with hypotension, Obstructive shock 1, Obstructive Shock 2, Agitation 1, Agitation 2, LV shock 1, LV shock 2, Hypovolemia 1, Hypovolemia 2, Hypovolemia 3, SVT & Hypotension and SVT & Hypotension 2.

The database contain 4,647 variables that were sampled of 8,273 patients between 2002-2015. Data selection was done by consulting with specialist, which selected the most valuable variables that caregiver look on when he or she diagnose patient's problem. The most important variables for caregivers that were found in the database are Heart rate, ArtBPS, ArtBPM, CVP, PAPD.

In addition, we selected 7 of the most frequent variables in the database (which were sampled for most of the patients): Spo2, ST1, ST2, ST3, Fio2, RR total and RR mandatory.

The database contained 70GB with 781,229,536 records in total. For extracting the relevant patients who have all the 12 variables sampled on the same time, we wrote a script in python that extract the relevant patients from the database. The script check each sampled minute and until now (7 weeks) pass over approximately 650 patients in total, while only 73 patients had the relevant 12 variables that were selected for initial checking for this research.

Each row in the dataset represents a minute of sampling for each variable separately. The data were reshaped that each row will represent all the variables that were sampled on the same time for each patient.

The dataset contain 10,154 alarms (33.57%) from a total of 73 patient records while the algorithms were made using the associated 505 hours of simultaneous Heart Rate (HR), Arterial Pressure Systolic (ArtBPS), Arterial Pressure Mean (ArtBPM), Central Venous Pressure (CVP), Pulmonary Artery Pressure Diastolic (PAPD), Total Respiratory Rate (RR Total) Mechanical Respiratory Rate (RR Mandatory), Spo2, ST and Fio2.

Data understanding

Variables

Dependent variable

Variable Name	Description	Type
Tag	1 – True Alarm 0 – No Alarm	Binary

Covariate variables

Category	Variable Name	Description	Type	Normal Range
Blood Pressure	ArtBPS	Arterial Pressure Systolic	Numeric	90 - 140 mmHg
	ArtBPM	Arterial Pressure Mean	Numeric	70 - 105 mmHg
	CVP	Central Venous Pressure	Numeric	3–8 mmHg
	PAPD	Pulmonary Artery Pressure Diastolic	Numeric	8 - 15 mmHg
Heart Beat	HR	Heart Rate	Numeric	60–100 bpm
	ST1	ST Segment 1	Numeric	~-0.5
	ST2	ST Segment 2	Numeric	~-0.5

	ST3	ST Segment 3	Numeric	~-0.5
Respiration	RR Mandatory	Mechanical Respiratory Rate	Numeric	12-18-breaths per minute
	RR Total	Total Respiratory Rate	Numeric	12-18-breaths per minute
	Spo2	Peripheral capillary oxygen saturation	Numeric	> 92%
	Fio2	Fraction of Inspired Oxygen	Numeric	30%-50%

Type	Amount	%
Alarm	10,154	33.57
No Alarm	20,093	66.43
Total	30,247	100

Precise amount of each type of alarm attached in Appendix 1.

Covariate Description

ArtBPS

Systemic blood pressure refers to the pressure exerted on blood vessels in systemic circulation, and is often measured using arterial pressure, or pressure exerted upon arteries during heart contractions.

ArtBPM

The Mean Arterial Pressure (MAP) calculates mean arterial pressure from measured systolic and diastolic blood pressure values. It is defined as the average arterial pressure during a single cardiac cycle.

CVP

The central venous pressure (CVP) is the pressure measured in the central veins close to the heart. It indicates mean right atrial pressure and is frequently used as an estimate of right ventricular preload. The CVP does not measure blood volume directly, although it is often used to estimate this.

PAPD

Pulmonary arterial pressure is generated by the right ventricle ejecting blood into the pulmonary circulation, which acts as a resistance to the output from the right ventricle. With each ejection of blood during ventricular systole, the pulmonary arterial blood volume increases, which stretches the wall of the artery. As the heart relaxes (ventricular diastole), blood continues to flow from the pulmonary artery into the pulmonary circulation. The smaller arteries and arterioles serve as the chief resistance vessels, and through changes in their diameter, regulate pulmonary vascular resistance.

RR total

Respiratory rate means the rate at which breaths occur, usually measured in breaths per minute. Human respiration rate is measured when a person is at rest and involves counting the number of breaths for one minute by counting how many times the chest rises. Respiration rates may increase with fever, illness, or other medical conditions.

RR mandatory

Mechanical ventilation helps patients breathe by assisting the inhalation of oxygen into the lungs and the exhalation of carbon dioxide. Depending on the patient's condition, mechanical ventilation can help support or completely control breathing.

Spo2

Peripheral oxygen saturation is an estimation of the oxygen saturation level usually measured with a pulse oximeter device.

ST

In electrocardiography, the ST segment connects the QRS complex and the T wave and has a duration of 0.080 to 0.120 sec (80 to 120 ms). It starts at the J point (junction between the QRS complex and ST segment) and ends at the beginning of the T wave. However, since it is usually difficult to determine exactly where the ST segment ends and the T wave begins, the relationship between the ST segment and T wave should be examined together.

Fio2

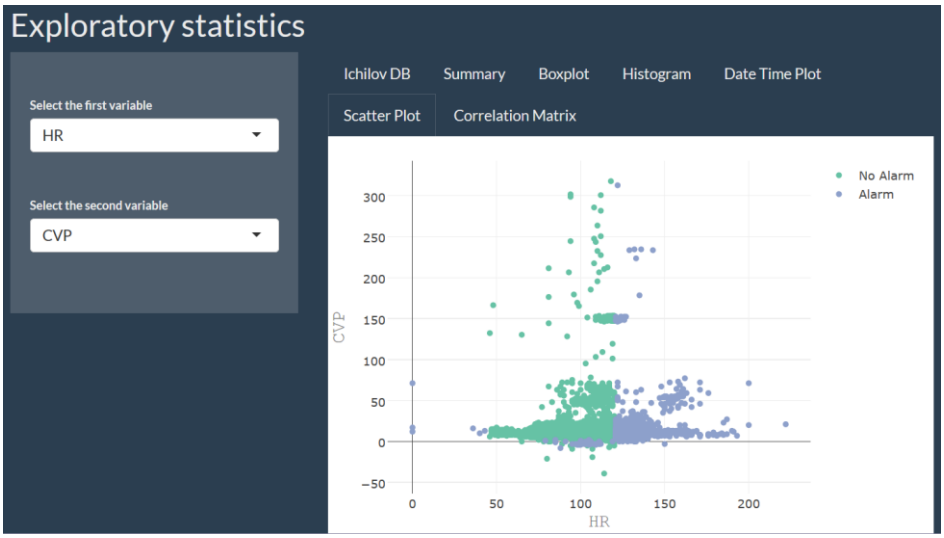
Fraction of inspired oxygen is the fraction or percentage stands for Fraction of Inspired Oxygen in the space being measured; it is a fraction of the amount of oxygen a patient is inhaling produced by an oxygen device such as a nasal cannula or mask. Different devices deliver different amounts of oxygen to the patient.

Data Visualization Tool – Description statistics

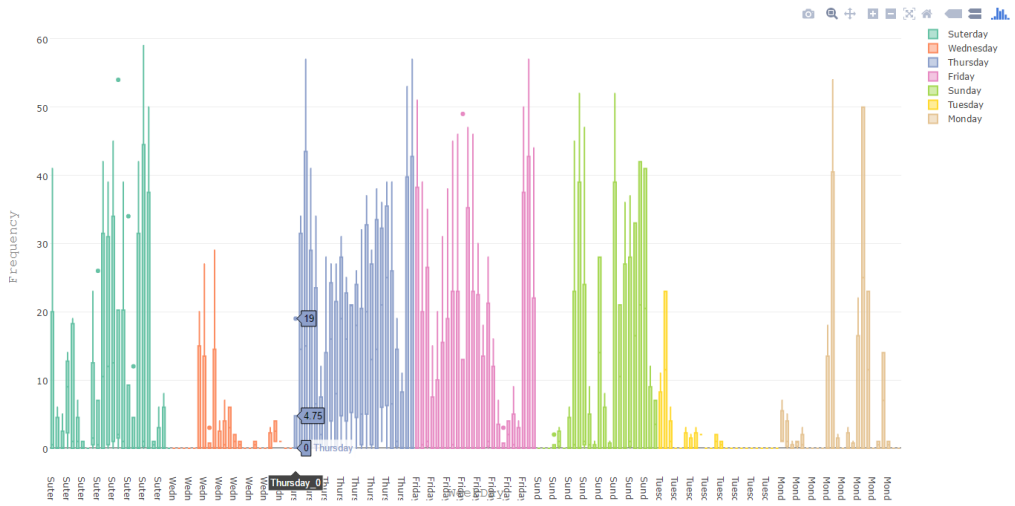
For understanding the abnormal behaviors of the data and to analyze the patterns in the dataset we built an interactive data visualization tool that allows easy access to Ichilov dataset and convenient way to analyze patient data. It provides the flexibility of selecting variables that can be visualized and present advanced visualization dimensions to detect patterns using drill down abilities on the digitally recorded signals.

The data description is the basis of the medical data evaluation and is the indispensable starting point for further methodological procedures such as statistical process control (SPC) and machine learning (ML).

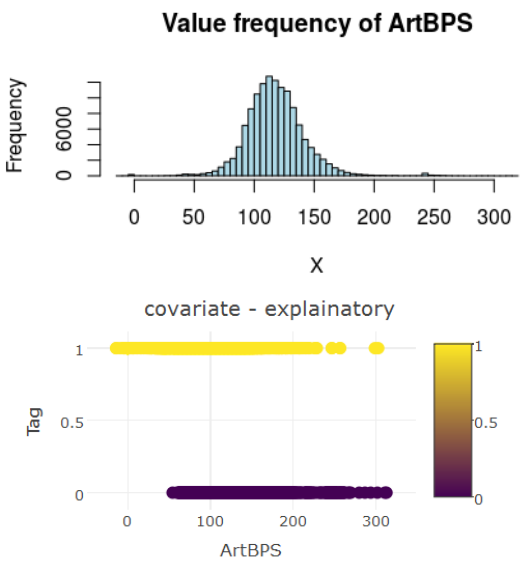
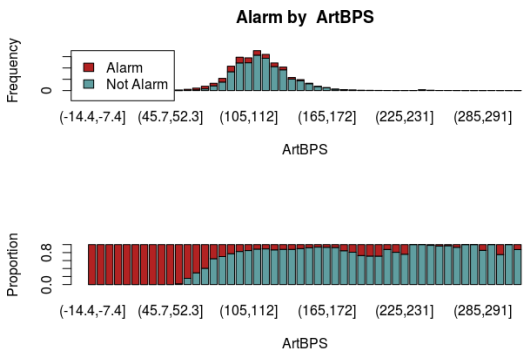
The visualization tool available at this link: https://galhever.shinyapps.io/ml_visualization_tool/. The tool presents essential descriptive statistics of the data in some visualization dimensions that are part of the medical analysis and a prerequisite for the understanding of further statistical evaluations.



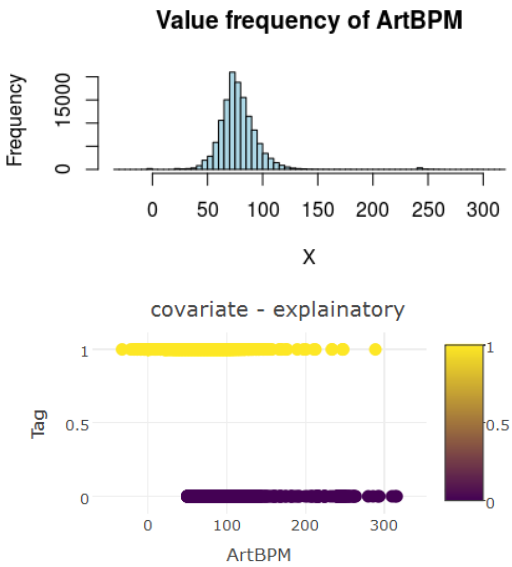
We checked the alarm frequency among the week. Most of the alarms occur between 12-13 and 17-18 and the distribution of alarm amount among the day is normal. Alarm frequency by hour and weekday available at this link: http://rpubs.com/galhev/253302

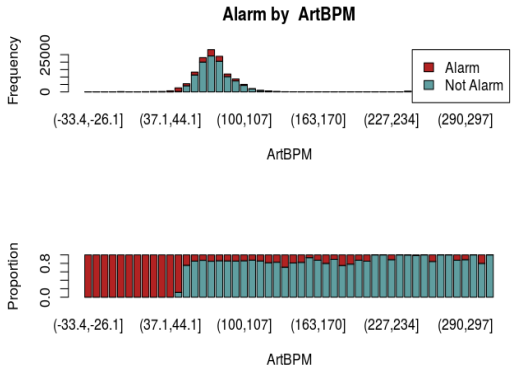


Covariate-Explanatory Histograms and Scatter plots

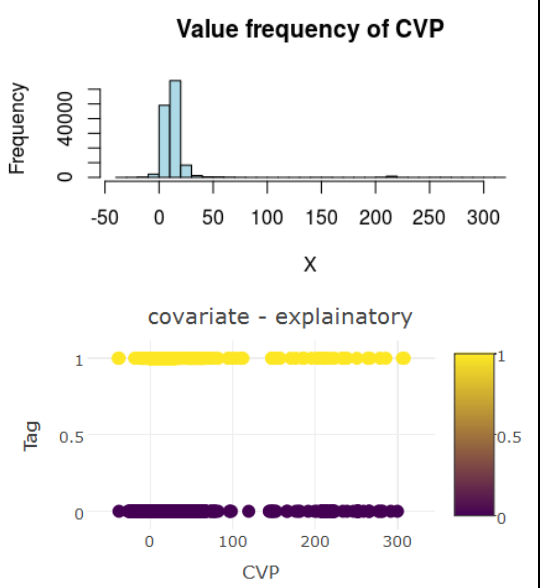
Variable Name		
ArtBPS		

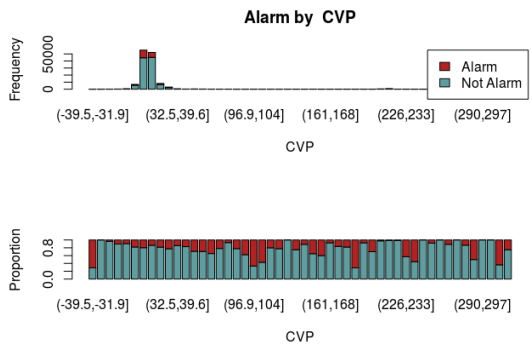
ArtBPM



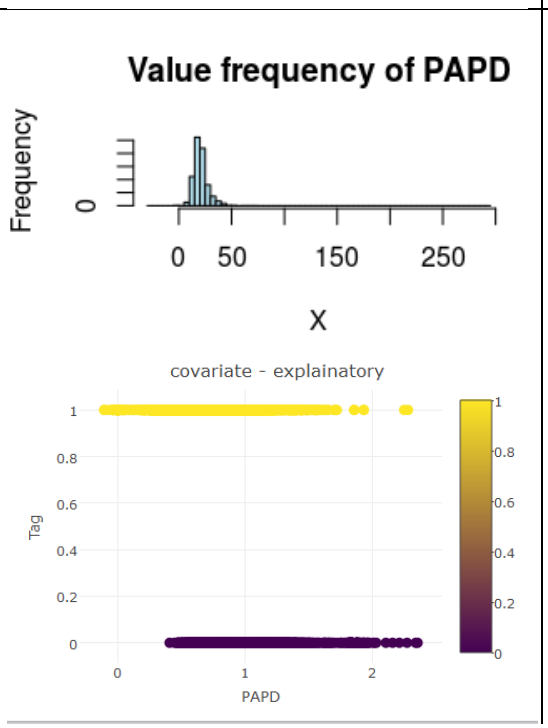


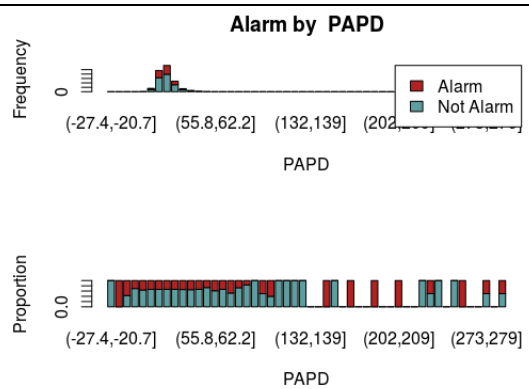
CVP



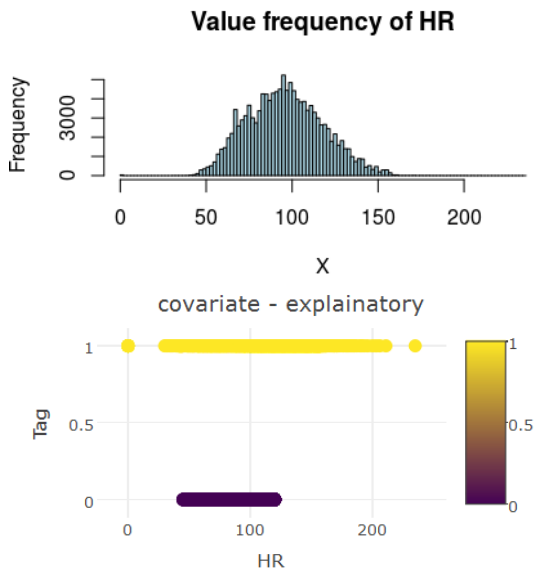


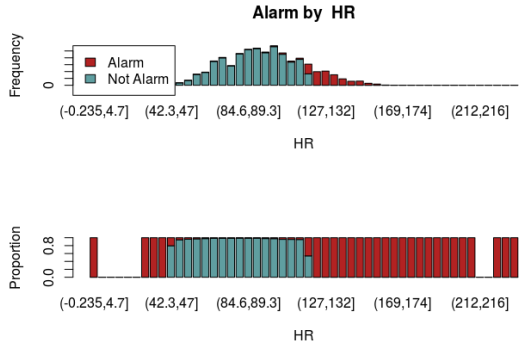
PAPD



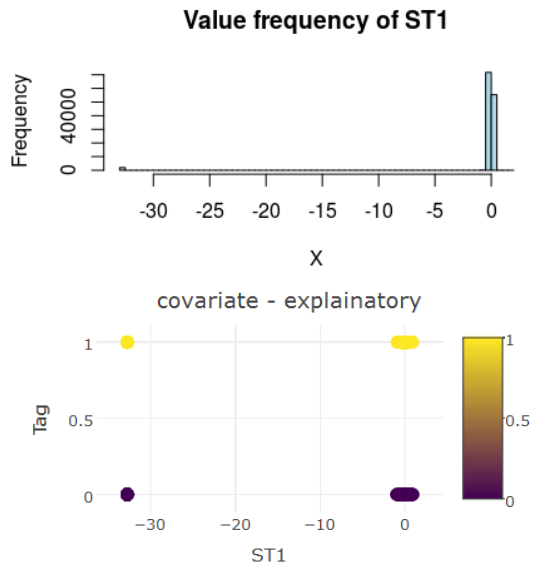


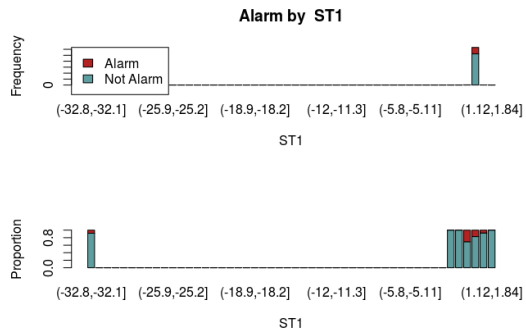
HR


Value frequency of HR
Frequency
0
3000
0
50
100
150
200
X
covariate - explainatory
Tag
1
0.5
0
0
100
200
HR

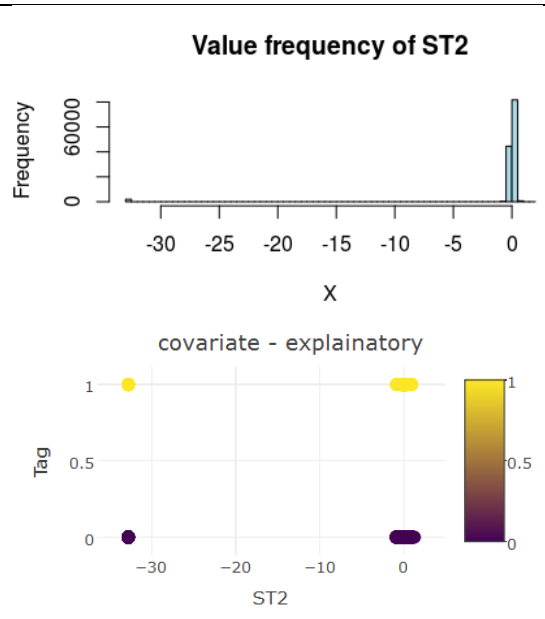

Alarm by HR
Frequency
Alarm
Not Alarm
(-0.235,4.7]
(42.3,47]
(84.6,89.3]
(127,132]
(169,174]
(212,216]
HR
Proportion
0.0
0.8
(-0.235,4.7]
(42.3,47]
(84.6,89.3]
(127,132]
(169,174]
(212,216]
HR

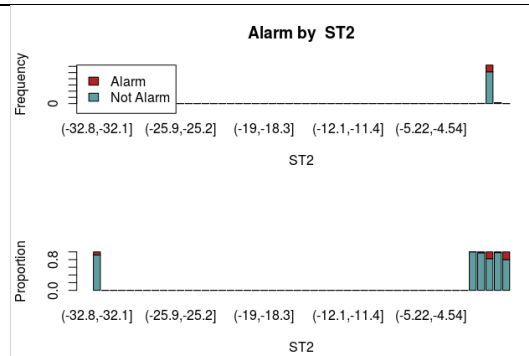
ST1


Value frequency of ST1
Frequency
0
40000
-30
-25
-20
-15
-10
-5
0
X
covariate - explainatory
Tag
1
0.5
0
−30
−20
−10
0
ST1

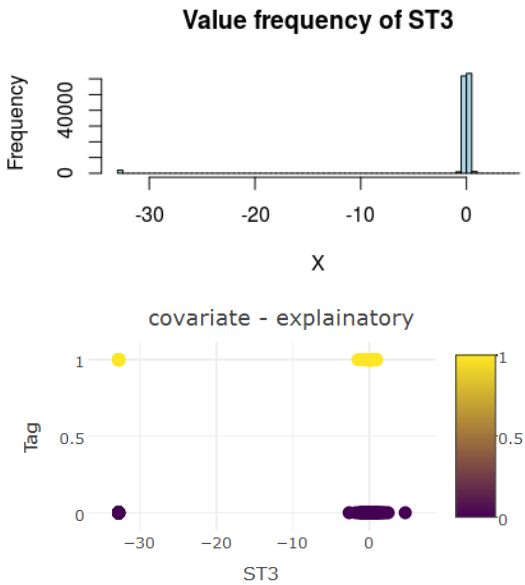

Alarm by ST1
Frequency
Alarm
Not Alarm
(-32.8,-32.1]
(-25.9,-25.2]
(-18.9,-18.2]
(-12,-11.3]
(-5.8,-5.11]
(1.12,1.84]
ST1
Proportion
0.0
0.8
(-32.8,-32.1]
(-25.9,-25.2]
(-18.9,-18.2]
(-12,-11.3]
(-5.8,-5.11]
(1.12,1.84]
ST1

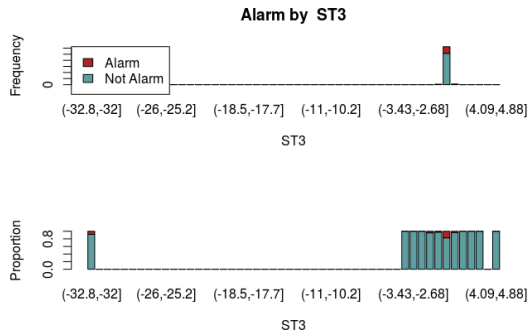
ST2


Value frequency of ST2
Frequency
0
60000
-30
-25
-20
-15
-10
-5
0
X
covariate - explainatory
Tag
1
0.5
0
−30
−20
−10
0
ST2

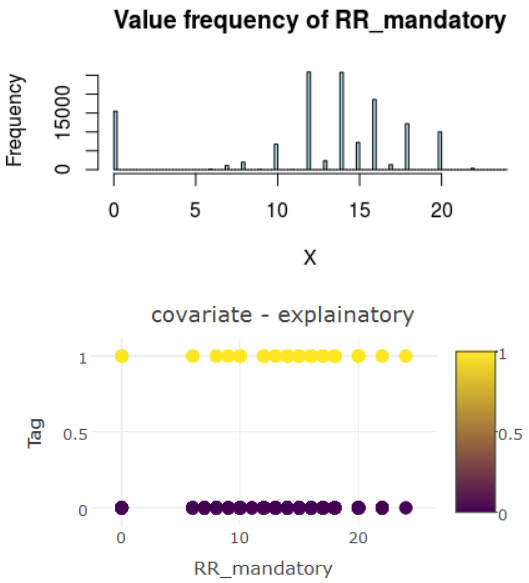

Alarm by ST2
Frequency
Alarm
Not Alarm
(-32.8,-32.1]
(-25.9,-25.2]
(-19,-18.3]
(-12.1,-11.4]
(-5.22,-4.54]
ST2
Proportion
0.0
0.8
(-32.8,-32.1]
(-25.9,-25.2]
(-19,-18.3]
(-12.1,-11.4]
(-5.22,-4.54]
ST2

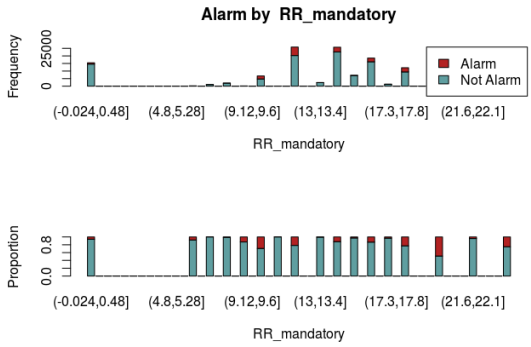
ST3



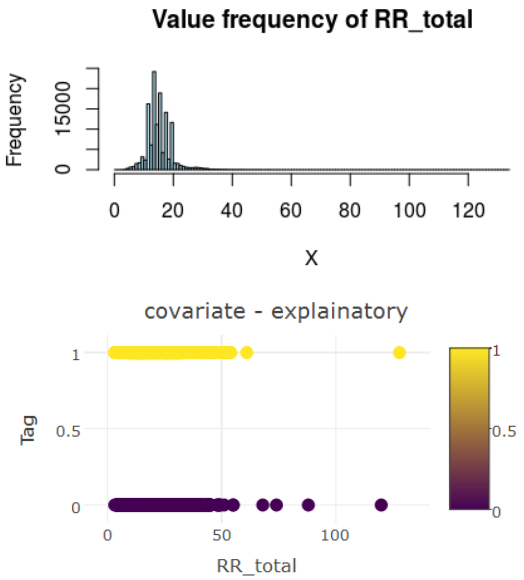


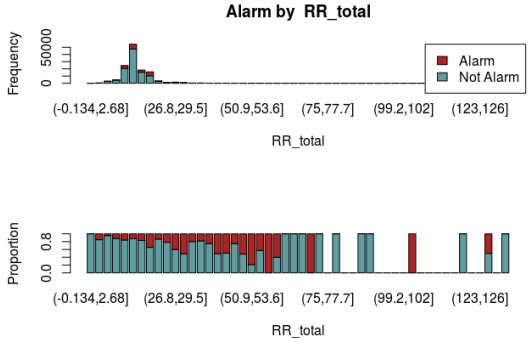
RR Mandatory



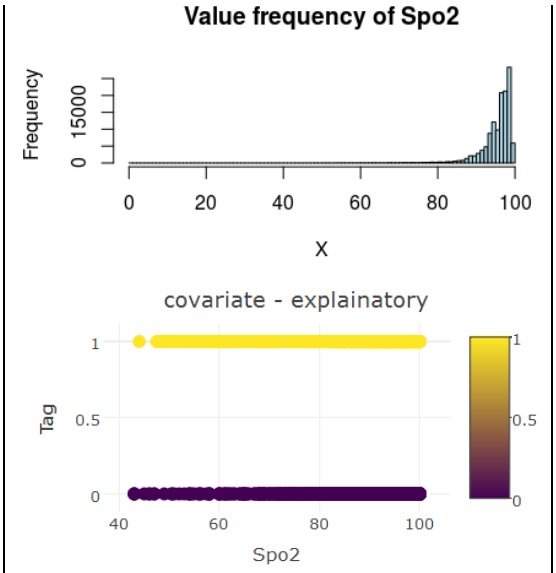


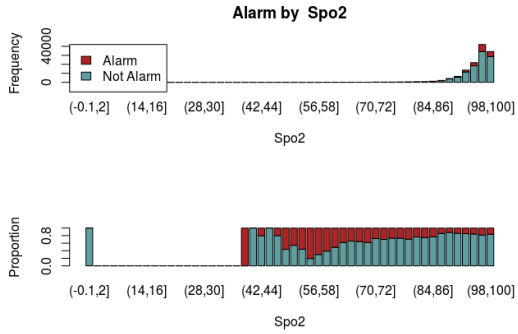
RR Total



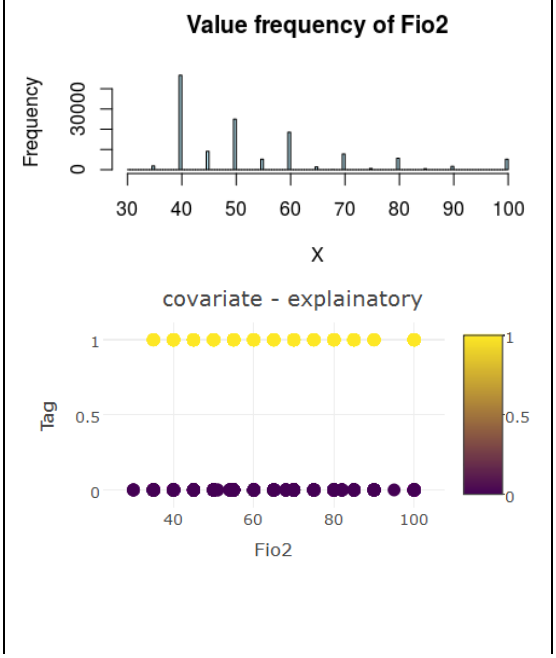


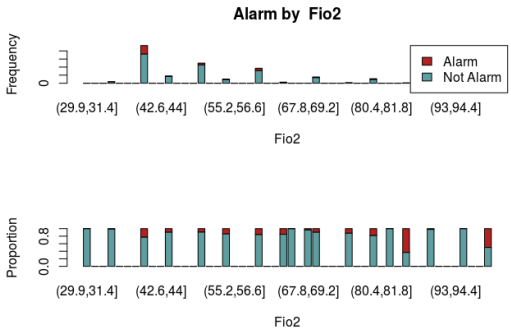
Spo2





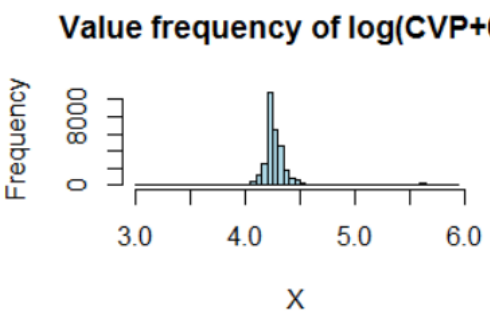
Fio2

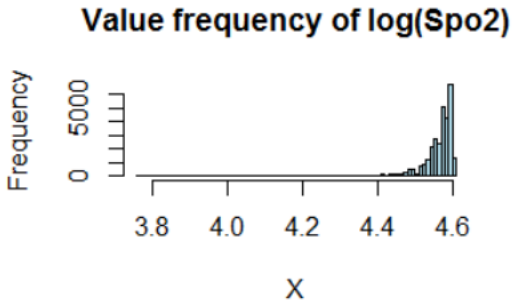


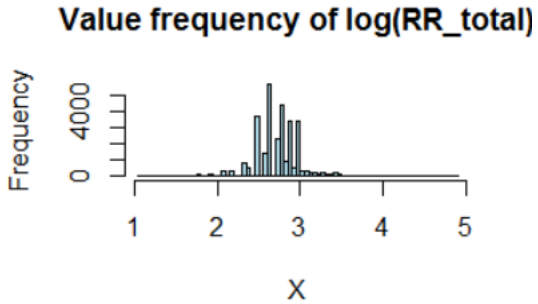


By looking at the histograms, it seems that the ArtBPS, ArtBPM, PAPD and HR are normally distributed and Spo2 has a left-tailed distribution. If we look at the skewness measure for asymmetry (included in appendix 2) we can see that all the ST segments, Sp02 and RR mandatory have a negative values. Negative skew indicates that the tail on the left side of the probability density function is longer or fatter than the right side. Conversely, all the others have positive values, which indicates that the tail on the right side is longer or fatter than the left side. We can understand it also by looking at the distances between the quantiles or by the location measures (median and mean). In case when the mean is bigger than the median it indicates that the tail on the right side and the other case in accordance.

For stabilize the variance errors we will perform variable transformation to close the distribution to normal. The following variables after log transformation:

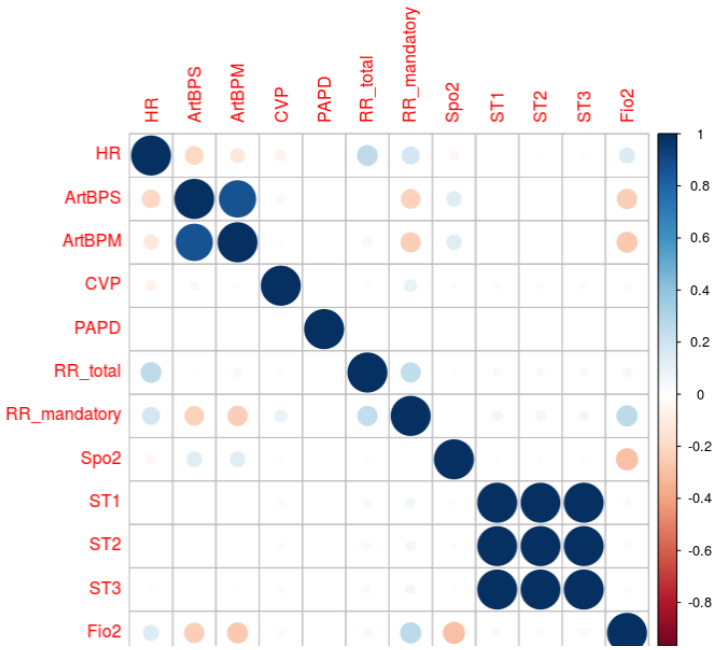






Correlation Matrix

Correlation is one of the most common and most useful statistics. A correlation describes the degree of relationship between two variables. In the correlation matrix below we can see positive correlation between blood pressure-related indicators and heart-beat-related indices. Quite logical since each measure the same signal in a different way.



Modelling

Measurement is a tool for identifying changes. Repeated measures of the same parameter often yield slightly different results, for example, re-measurement of a patient's blood pressure, hence, interpretation of data to detect change is not always a simple matter. This inherent variability is due to factors such as fluctuations in patient's biological processes, differences in service processes, and imperfections in the measurement process itself. Measurement include three steps. First is determining and defining key indicators; second is collecting an appropriate amount of data; and third is analyzing and interpreting these data. This section focuses on the third component the analysis and interpretation of data using machine learning (ML) and statistical process control (SPC) methods. (Benneyan, Lloyd, & Plsek, 2003)

Handling the Unbalanced Data Problem

The data for many classification problems is inherently imbalanced, as machine learning models are often used to sort through huge populations of negative (uninteresting) cases to find the small number of positive (interesting and alarm-worthy) cases. Conventional algorithms' loss functions usually attempt to optimize quantities such as error rate, thus they are often biased towards the majority class as they are not taking the data distribution or the different errors importance into consideration.

While testing different algorithms with our data, we've encountered classifiers in which the minority examples were treated as outliers of the majority class and ignored, so the algorithm simply generated a trivial classifier that classified every example as the majority class, i.e. created 100% misclassification for the "alarm" samples (100% false negative rate). In our case, a false negative error could mean that a patient is coding in his bed and no one of the medical team is notified of it. Thus, the false negative errors are far more severe than the false positive errors, and they must be addressed more carefully.

Learning from imbalanced data has been studied actively in the machine learning community and a vast number of techniques have been suggested. Among approaches are changing the class ratios within the data by oversampling the minority class or undersampling the majority class. At the algorithm level one can adjust the class weights (misclassification costs), adjust the decision threshold and more. We chose to test several methods during the modeling, as will be reviewed in the following section.

Measure selection

Our research focused on false alarm reduction, which consider false positive (type I) errors. On previous chapters we represented the problems of such type of cases. From the other hand we need also to consider in our tests at type II error may be compared with a so-called false negative (where an actual 'hit' was disregarded by the test and seen as a 'miss'). False negatives may provide a falsely reassuring message to patients and physicians that disease is absent, when it is actually present. This sometimes leads to inappropriate or inadequate treatment of both the patient and their disease.

False negatives and false positives are significant issues in medical testing. We don't want to miss any problem in patient's condition and on the other hand we don't want a lot of false alarms in order to prevent alarm fatigue. We want to balance between false positive and false negative and still keeping on high sensitivity that is more important in healthcare domain.

The higher severity of false negative errors (over false positive errors) is not reflected in the naïve accuracy measure normally used. Thus, we explored a few alternative measures which allow weighting of the different error types.

Sensitivity and specificity are terms used to evaluate classification accuracy. The sensitivity and specificity of a quantitative test are dependent on the cut-off value above or below which the test is positive. In general, the higher the sensitivity, the lower the specificity, and vice versa.

$$\boldsymbol{Sensitivity} = \boldsymbol{Recall} = 1 - \alpha = \frac{True\ positives}{True\ positives + False\ Negatives}$$

$$\boldsymbol{Specificity} = 1 - \beta = \frac{True\ negatives}{True\ negatives + False\ Positives}$$

The **sensitivity** of a classification test refers to the ability of the test to correctly identify those patients with a disease, or in our case, the samples that should cause alarm. For example, a test with 80% sensitivity detects 80% of alarms (true positives) but 20% cases which required an alarm go undetected (false negatives). A high sensitivity is clearly important where the test is used to identify a patient in a life-threatening condition.

The **specificity** of a classification test refers to the ability of the test to correctly identify those patients without the illness, or the samples that should be classified as no-alarm. Therefore, a classification test with 80% specificity correctly reports 80% of no-alarm as test negative (true negatives) but 20% of no-alarm are incorrectly identified as test positive (false positives).

Another measure for classification results is the **precision** measure which answers the question - when the classifier predicted an alarm, how often is it correct?

$$\boldsymbol{Precision} = \frac{True\ positives}{Predicted\ negatives}$$

The **F** and **F_β** scores are commonly used for classification evaluation, using the sensitivity and precision measures. The F and F_β score gives beta times more weight to the sensitivity component over the precision component:

$$\boldsymbol{F\ score} = 2 \cdot \frac{Precision \cdot Sensitivity}{Precision + Sensitivity}$$

$$\boldsymbol{F_\beta\ score} = (1 + \beta^2) \cdot \frac{Precision \cdot Sensitivity}{\beta^2 \cdot Precision + Sensitivity}$$

The following measures balance sensitivity and specificity by weights:

$$\boldsymbol{Weighted\ Accuracy} = \beta * Precision + (1 - \beta) Sensitivity$$

$$\boldsymbol{Weighted\ Measure} = \sqrt[W_1+W_2]{Sensitivity^{W_1} \cdot Specificity^{W_2}} = \sqrt[W_1+W_2]{(1-\alpha)^{W_1} \cdot (1-\beta)^{W_2}}$$

A preliminary test was conducted in order to understand the proposed measures better and to choose a measure for our own model comparison. We trained (over a train set including only 7 of the variables) a random forest classifier with 20 trees and 3 variables in each tree, and got the following confusion matrix:

	Reference	
Prediction	No Alarm	Alarm
No Alarm	32,131	8,582
Alarm	1,831	525

Out of 9,107 true alarms, only 525 samples (**6%**) were actually classified as alarms. Also, out of 33,962 no-alarms, only 1,831 samples (5%) were wrongly classified. The following measures were calculated, as we tested several options for weights of the Sensitivity component within the weighted measures:

Weight	Accuracy	Sensitivity	Specificity	Precision	F1_score	Fb_score	Weighted Accuracy	Weighted Measure
1	0.75823	0.05765	0.94609	0.22284	0.09160	0.09160	0.50187	0.23354
5						0.05934	0.20572	0.09190
10						0.05807	0.13842	0.07435
20						0.05776	0.09996	0.06586

As we wanted the Sensitivity to have a magnified weight within the chosen measure, we chose to use the Weighted Measure with 10 times more weight to the Sensitivity component over the Specificity component. By this we are willing to risk a higher rate of false positives, in order to reduce the rate of false negatives.

Downsampling

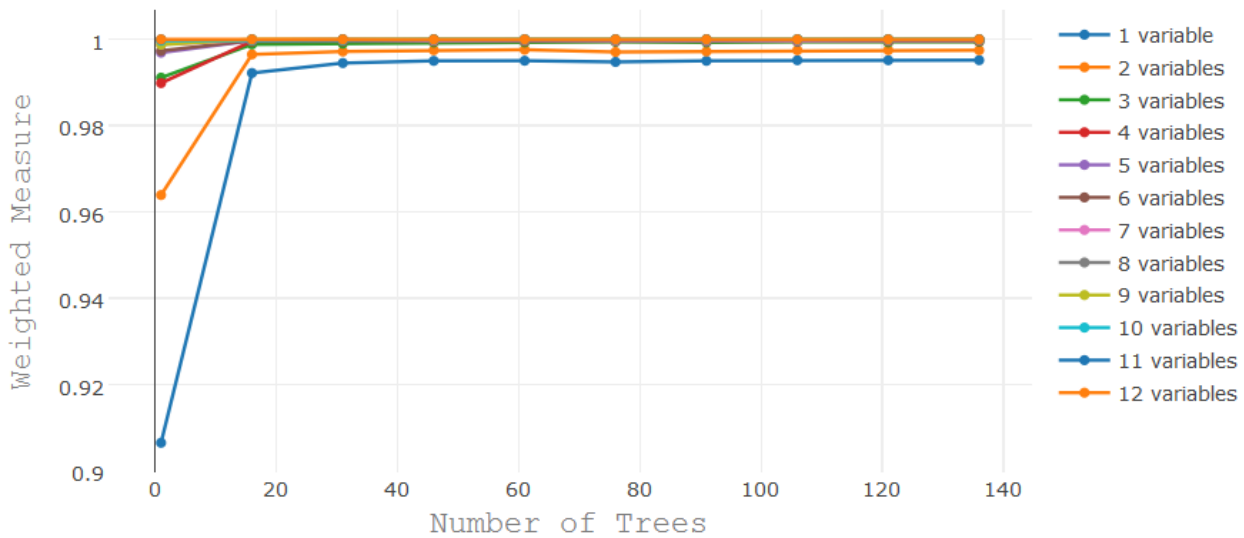
Down-sampling tends to work well empirically but loses some information, as not all of the training data is used for training the classifier. As we wanted to reduce data size for computational time reasons anyway, we've decided to do so while also downsampling the "No alarms" class more aggressively, in order to create a better balance of the classes. The original data had a total of 129,255 samples, of which only 17% of the samples belonged to the alarm class. After reducing the dataset size we got a new set with the following amounts:

	Prior Probabilities	
	No Alarm	Alarm
Amount	20,093	10,154
Probability	66%	34%

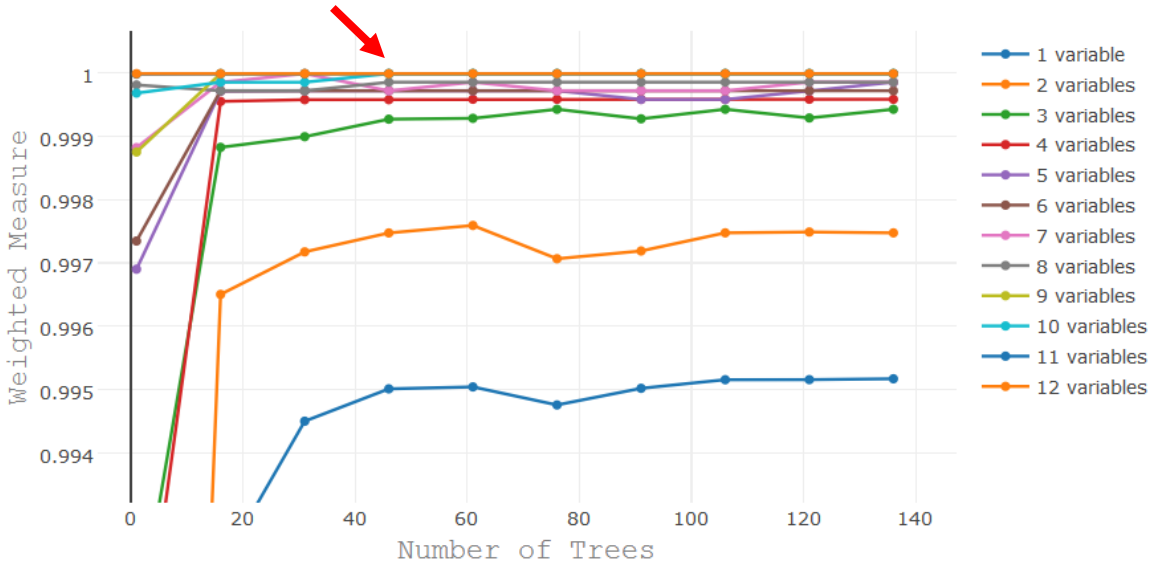
This new dataset was used for the whole modeling section.

Random Forest

The Random Forest algorithm grows an ensemble of decision trees (each is a weak classifier by itself) as their majority vote decision creates a strong classifier of its own. The algorithm for growing a forest requires a few parameters, among them the number of trees in the forest, and the number of variables to be randomly chosen for each tree, thus also determining the trees' depth. In order to determine the parameters for fitting a good model to our data, we've conducted an experiment in which we've tested all possible numbers of variables to be chosen for the trees (1-12), and a number of trees varying from 1 to 150. Each possible combination of parameter was tested using CV-4 over the train set, which included 20,163 samples. The experiment's results can be seen in the following graph.

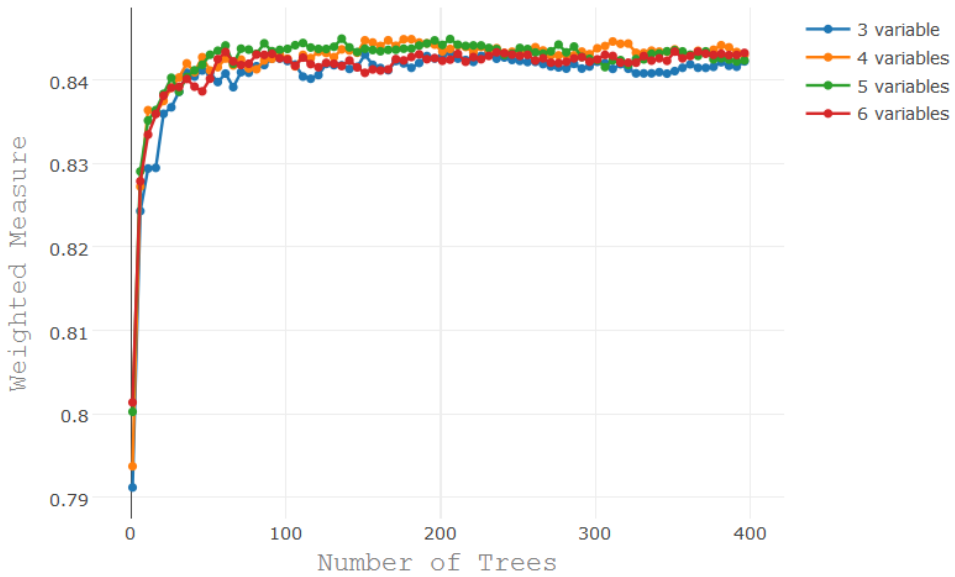


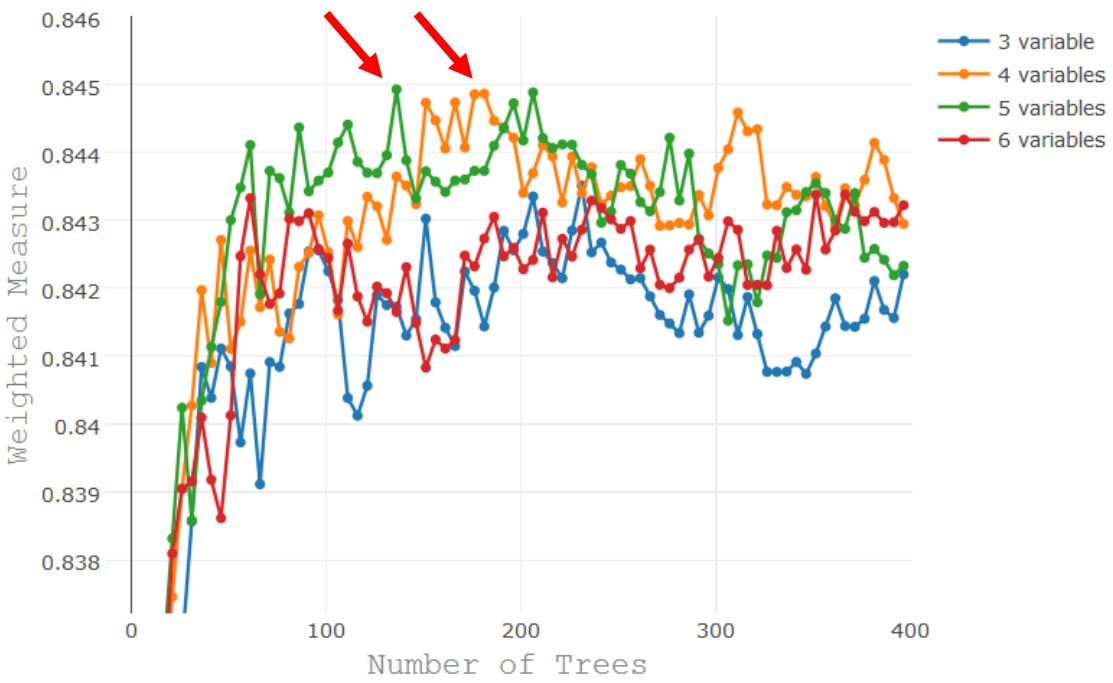
Looking closer at the results we've seen that models with 11 and 12 variables for each tree, produced classifier with an almost perfect measure score of 0.99998 for a vast range of forest sizes.



We've assumed the classifiers' extremely high accuracy was due to the method that was used for tagging the original samples, which used the variables HR, CVP, artBPM, artBPS and PAPD. Thus, we were interested in finding out what amount of usable information lies in the variables which were not used for the original tagging. **Can we classify the alarms correctly using only a small set of variables?**

We've prepared a new dataset containing only 7 variables: ST1, ST2, ST3, SPO2, FIO2, RR mandatory and RR total, in order to estimate how much useful information can be extracted from a subset of the 12 variables. We've conducted an experiment over the training set, in order to find out how many variables should be chosen to grow each tree, and saw the best forests were generated using 3 to 6 variables in each tree. Another experiment was conducted to find a good combination of the parameters: we've tested possible numbers of variables to be chosen for the trees (3-6), and a number of trees varying from 1 to 400.

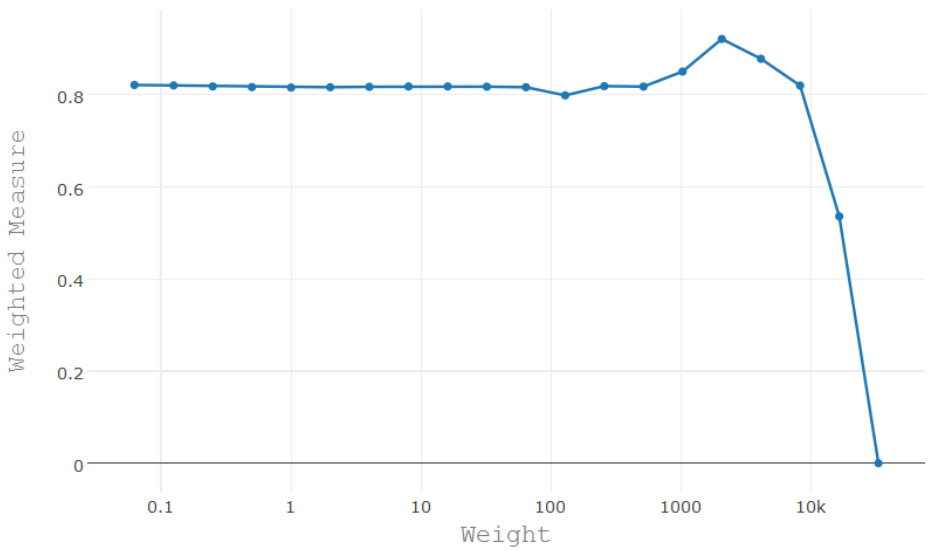




The highest Weighted Measures by CV-4 were of value 0.8449 for 136 trees and 5 variables, and 0.8448 for 176 trees and 4 variables. We were not quite satisfied with the results yet and decided to test another method for balancing the false negative and false positive rates.

Class Weights

Weighted Random Forest enables to put more "weights" on the minority class, thus penalizing more heavily on misclassifying samples originating in it. The "*randomForest*" package in *R* allows adding such class weights. After experimenting with this option we concluded that only a fairly large relation between the class weights might actually change the classifier generated by the algorithm. We've tried changing that relation in the log scale to find a range in which the classifier will be more accurate. We set the weight for the no-alarm class to 0.001 and the weight for the alarm class to vary between 2^{-4} And 2^{15} and tested the algorithm with 136 trees and 5 variables, using CV-4 over the train set. The following figure portrays the experiment results.



Following the results, we've conducted a final experiment with the alarm class weight varying between 1000 and 4000, for parameter combination of both 4 variables and 176 trees and 5 variables and 136 trees:

The best weighted measure value of 0.919 was for an algorithm with 4 variables and 176 trees and class weights of 0.001 and 1900 for no-alarm and alarm classes, accordingly.

Final Random Forest Tests

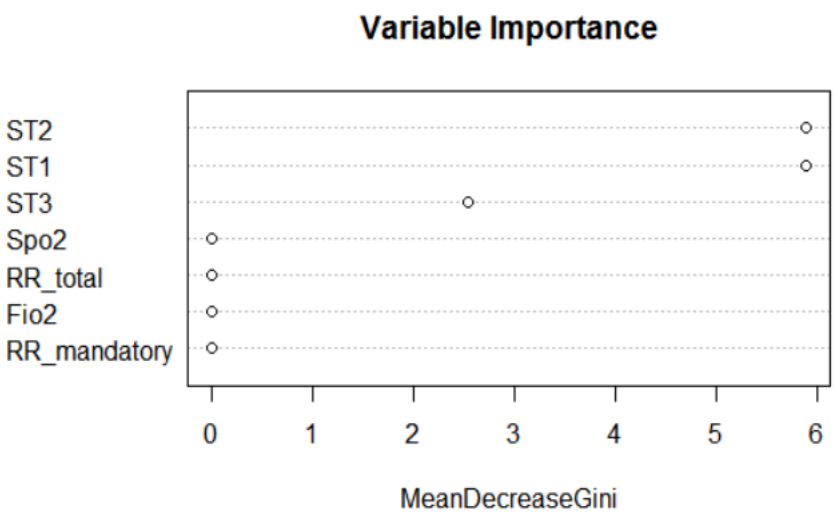
With the chosen setting portrayed above, a classifier was trained over the train set containing the **partial 7 variables set**, and then tested over the test set, producing the following results:

	Reference	
Prediction	No Alarm	Alarm
No Alarm	3,807	189
Alarm	2,872	3,216

Accuracy	Sensitivity	Specificity	Precision	F1_score	Fb_score	Weighted_Accuracy	Weighted_Measure
0.69645	0.944493	0.569996	0.528252	0.677552	0.937182	0.910448	0.902111

A substantial 94.4% of the alarm samples were correctly classified as *alarms*, but this came with a cost: only 56.9% of the *no-alarms* were classified correctly.

We've also extracted the variable importance by the Mean Decrease in Impurity or Gini Index:

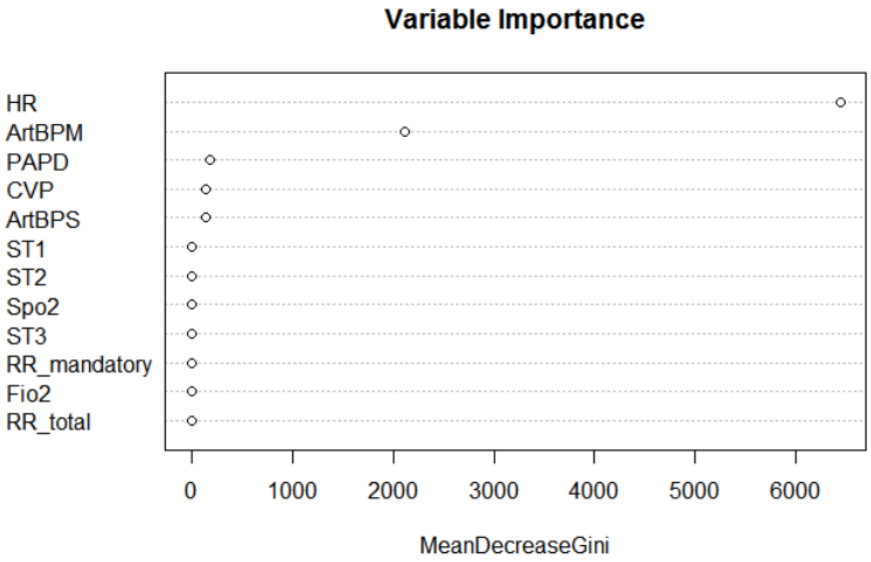


It is apparent that the variables describing measures related to the heart performance (ST1-ST3) have donated the most useful information for the classification.

Another test was conducted using the **full variable data** containing all 12 variables, using the random forest algorithm with 11 variables and 50 trees. The following results were produces:

	Reference	
Prediction	No Alarm	Alarm
No Alarm	6,754	0
Alarm	2	3,328

Accuracy	Sensitivity	Specificity	Precision	F1_score	Fb_score	Weighted_Accuracy	Weighted_Measure
0.999802	1	0.999704	0.999399	0.9997	0.999994	0.999973	0.999973

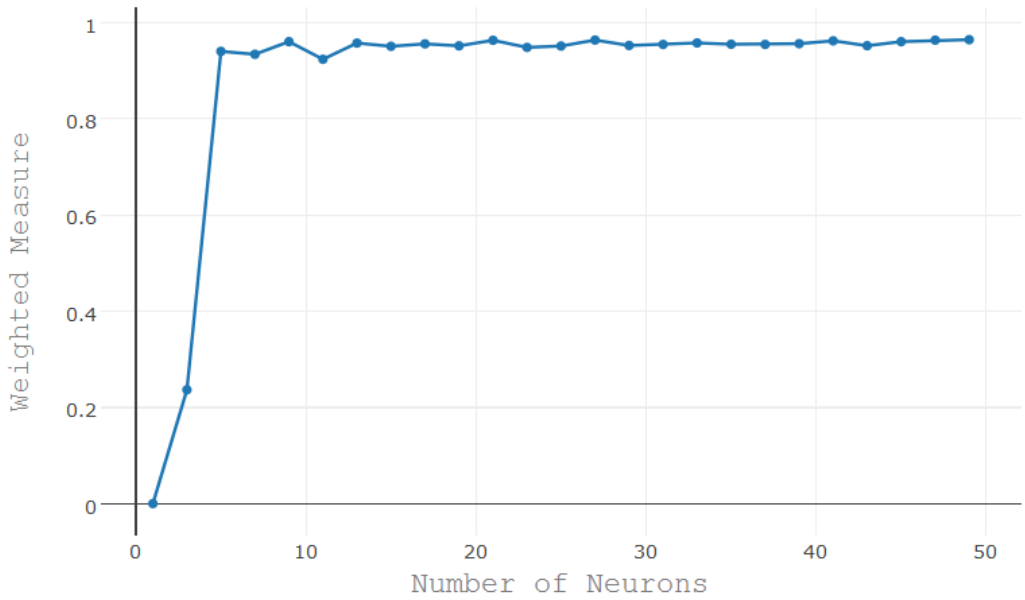


All true alarms were classified as alarms by the fitted model. Corresponding with the previous model's variable importance, the HR variable seems to have donated the most by far to the classification process, hence it is extremely important for recognizing a patient is in critical life-threatening state.

Neural Network

An Artificial Neural Network is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system, which is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. Artificial NNs have a remarkable ability to derive meaning and extract patterns from complicated or imprecise data, making them good models for value prediction or classification.

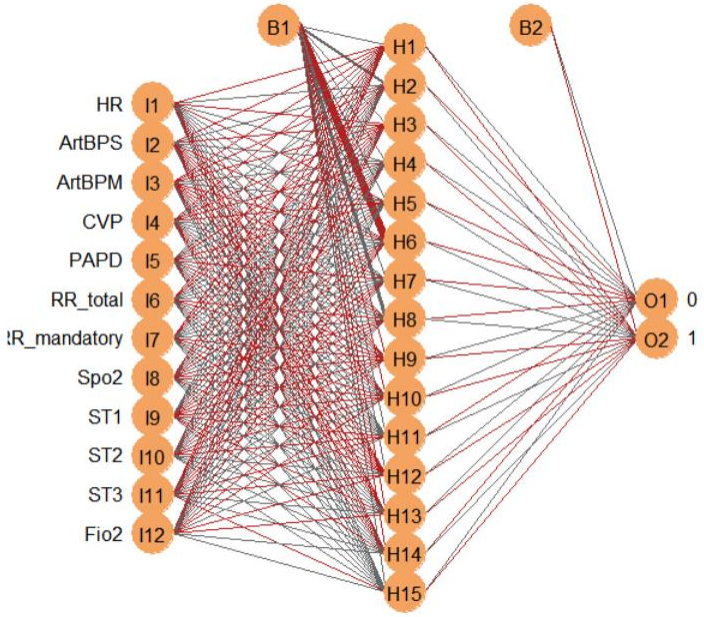
Constructing a NN requires setting the number of neurons for the hidden neuron layer. In order to choose the number of neurons needed we've conducted an experiment assessing the weighted measure of NNs with 1 to 46 neurons, throughout CV-4 over the train set containing the full set of 12 variables.



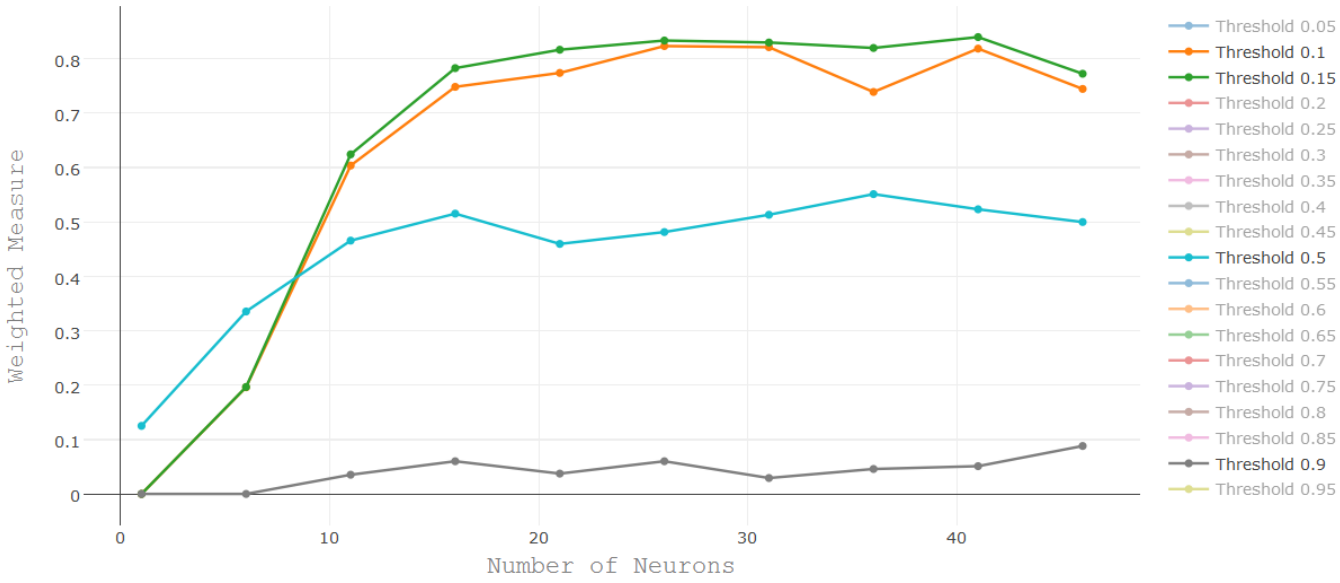
A final model with 15 neurons was chosen, trained over the training set and tested over the test set, producing the following results:

	Reference	
Prediction	No Alarm	Alarm
No Alarm	6,416	100
Alarm	340	3,228

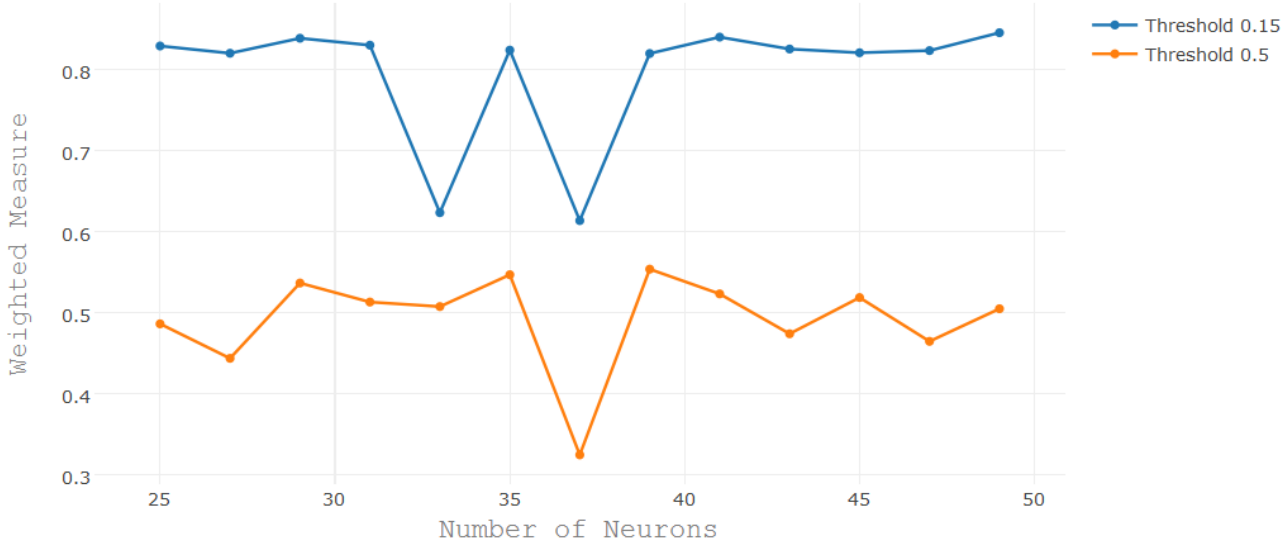
Accuracy	Sensitivity	Specificity	Precision	F1_score	Fb_score	Weighted_Accuracy	Weighted_Measure
0.956367	0.969952	0.949674	0.904709	0.936195	0.96926	0.968109	0.968091



A similar experiment was conducted to determine the number of neurons needed to train a classifier over the **partial dataset** containing only 7 variables. In order to balance between the false negative and the false positive rates, we've changed the threshold for deciding if a sample should be classified to class 0 (no-alarm) or 1 (alarm). The algorithm's results were extracted such that for each sample a score was given to its probability to belong to class 0 or 1. The threshold set was for determining if a sample belongs to the 1 class. For example, if the threshold was set to 0.2, each sample with a probability to the belong to class 1 that's higher than 0.2, was classified as is originated in class 1. We've varied the threshold between 0.05 and 0. 95 and got the following results:



The best results were achieved with threshold 0.15 in the range of 25 to 50 neurons, at which we've tested the algorithm again with more possible number of neurons values:



The best weighted measure of 0.8397 was achieved with 41 neurons and a 0.15 threshold. A final model of these parameters was trained over the training set and tested over the test set, and has generated the following results:

	Reference	
Prediction	No Alarm	Alarm
No Alarm	1,574	207
Alarm	5,105	3,198

Accuracy	Sensitivity	Specificity	Precision	F1_score	Fb_score	Weighted_Accuracy	Weighted_Measure
0.473225	0.939207	0.235664	0.385162	0.546293	0.926018	0.875249	0.828273

This classifier is still inferior to the one we've trained with a random forest model, which produced 94% sensitivity and 56% specificity over the same test set. For the full dataset containing all 12 variables, the random forest classifier also produced better results. This indicates that although useful information can be extracted from the variables which are not used at the decision process of caregivers, a much higher classification quality is achieved from using these variables as well.

K-Means

k-means algorithm takes as input the number of clusters to generate, k, and a set of observation vectors to cluster. It returns a set of centroids, one for each of the k clusters. An observation vector is classified with the cluster number or centroid index of the centroid closest to it.

$$\underset{C}{\operatorname{argmin}} \sum_{i=1}^{K} \sum_{x \in C_i} ||x - \mu_i||^2$$

When μi is the center of each cluster and Ci is a cluster i. In practice, as selected larger K target function value will be smaller (if number of clusters will be similar to number of examples, the target value will be zero). Therefore, in order

to select number of clusters "naturally" without making over adjustment, we will qualify number of clusters by clustering indexes, which also punishes by complexity (number of clusters in our case).

Variables tuning - determining the optimal number of clusters

Davies–Bouldin index

The Davies–Bouldin index (DBI) is a metric for evaluating clustering algorithms. This is an internal evaluation scheme, where the validation of how well the clustering has been done is made using quantities and features inherent to the dataset.

The Davies-Bouldin criterion is based on a ratio of within-cluster and between-cluster distances. The Davies-Bouldin index is defined as:

$$\mathrm{DB} = \frac{1}{K}\sum_{i=1}^{K} D_i$$

where Di,j is the within-to-between cluster distance ratio for the i-th and j-th clusters. In mathematical terms,

$$D_{i,j} = \frac{\bar{d}_i + \bar{d}_j}{d_{ij}}$$

Where d_i is the average distance between each point in the i-th cluster and the centroid of the i-th cluster. d_j is the average distance between each point in the j-th cluster and the centroid of the j-th cluster. di,j is the Euclidean distance between the centroids of the ith and jth clusters.

The maximum value of Di,j represents the worst-case within-to-between cluster ratio for cluster i. The optimal clustering solution has the smallest Davies-Bouldin index value.

Dunn index

The Dunn's index measures compactness (Maximum distance in between data points of clusters) and clusters separation (minimum distance between clusters). This measurement serves as a measure to find the right number of clusters in a data set, where the maximum value of the index represents the right partitioning given the index (partition with the highest separation between clusters and less spread data in between clusters).

First, the maximum distance between each two observations in each cluster will be calculated:

$$\Delta_i = \max_{x,y \in C_i} ||x - y||_2$$

Then the distance between each two observations in different clusters will be maximized:

$$\delta(C_i, C_j) = \max_{x \in C_i, y \in C_j} ||x - y||_2$$

Finally, we will calculate the indicator:

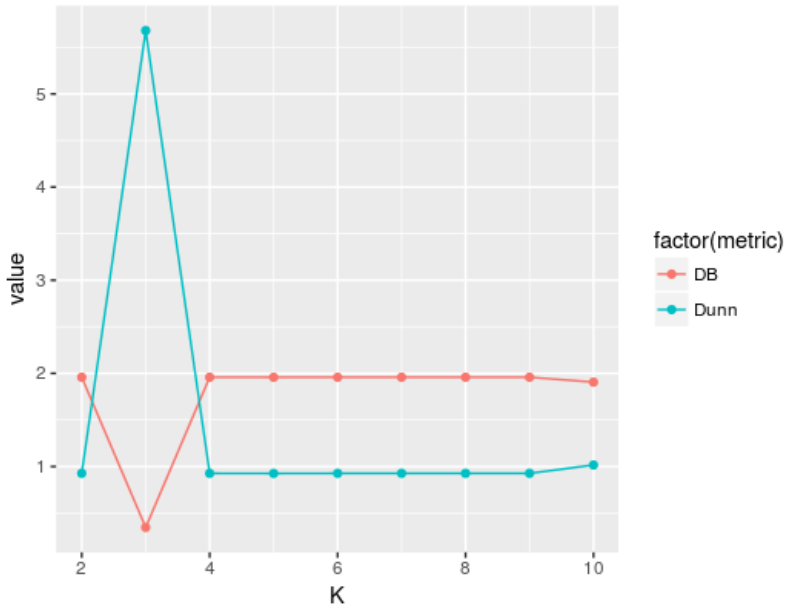
$$\mathrm{DI} = \frac{\min\limits_{1 \le i < j \le K} \delta(C_i, C_j)}{\max\limits_{1 \le l \le K} \Delta_l}$$

The optimal clustering solution has the highest Dunn index value.

We learned clustering schemes for k = {2-10} (number of clusters). For each K value 5 starts points randomly. For each scheme we calculated Dunn and Davies-Bouldin criterions:

Full Model

First test was conducted using the **full variable data** containing all 12 variables.



While K = 3 Dunn criterion has the highest value and Davies-Bouldin has the lowest value, so we checked the model with 3 clusters. We associated each cluster according to the majority "votes" in each section.

K-means - K=3

Predicted\Actual Cluster	1	2	3
No Alarm	16728	3008	357
Alarm	4275	5811	68

	Accuracy	Sensitivity	Specificity	Precision	F1_score	Fb_score	Weighted_Accuracy	Weighted_Measure
1	0.756968	0.572287	0.850296	0.658918	0.612555	0.573033	0.597561	0.593262

K-means - K=2

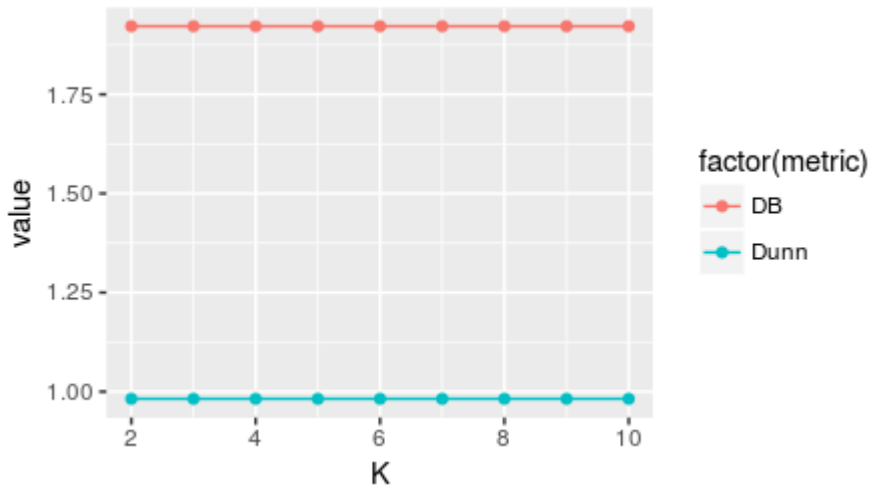
Since in our case we have two clusters (categorially explanatory variable with two options), we examined also K-means with K=2. In this case the majority "votes" for cluster 1 and 2 are both for "No Alarm" so we will choose the bigger percentage for this group and the rest represents the "Alarm" group.

Predicted\Actual Cluster	1	2
No Alarm	19736	357
Alarm	10086	68

	Accuracy	Sensitivity	Specificity	Precision	F1_score	Fb_score	Weighted_Accuracy	Weighted_Measure
1	0.654743	0.006697	0.982233	0.16	0.012856	0.006761	0.095382	0.010539

Kmeans - Partial model

We also examined a model which contains only a partial set of variables, containing the variables which help the doctors to diagnose patients' condition, i.e. the following variables: HR, ArtBPS, ArtBPM, CVP, PAPD.



The results of the two indicators shown that it does not matter what K will be selected. For our examining we will use k=2.

Predicted\Actual Cluster	1	2
No Alarm	6970	13123
Alarm	8621	1533

	Accuracy	Sensitivity	Specificity	Precision	F1_score	Fb_score	Weighted_Accuracy	Weighted_Measure
1	0.718881	0.849025	0.653113	0.552947	0.669722	0.844548	0.831215	0.829016

Multivariate Statistical Process Control

Statistical Process Control (SPC) is one of the important tools in quality control (QC) and it's methods have been widely recognized as effective approaches for process monitoring and diagnosis (Srinivasu, Reddy, & Rikkula, 2009). We use SPC for monitoring sequential processes to make sure that the process is stable. SPC is one of the major tools for quality control and management. (Qiu, 2013)

Most monitoring scenarios involve several related variables, and the quality of the measured value is affected by multiple characteristics (Montgomery, 2009). Multivariate statistical process control (MSPC) method relates to this cases which more than one variable are monitored simultaneously (Qiu, 2013). MSPC methodology, based on control charts, that used to monitor the stability of a multivariate process. Stability achieved when the means, variances, and covariance of the process variables remain stable over rational subgroups of the observations (Robert L. Mason & John C. Young, 1996).

Hotelling T^2

Hotelling T^2 multivariate control charts are the multivariate extension of univariate Shewhart control charts was used for analyzing the dataset. The main parameters for Hotelling T^2 calculations are mean ($\bar{x}$) value of each variable and the variance-covariance (S) matrix. These parameters are estimated from preliminary samples (phase I) when the process is assumed to be in control.

$$T^2 = (x - \bar{x})' S^{-1} (x - \bar{x})$$

The control limits for this statistic contain p is the number of quality characteristics observed in each sample. $\bar{x}$ and S is the covariance and the sample mean vector and covariance matrix, respectively.

$$UCL = \chi^2_{\alpha,p}$$

Shewhart and other SPC experts recommend control limits set at 3SD for detecting meaningful changes in process performance while achieving a rational balance between two types of risks, which used also in our research. On phase II we used the parameters (mean and covariance) that were found on phase I on the test set.

Hotteling - Partial Model

First step (phase I) was to identify the in-control training data (which are used to estimate the distribution parameters), and to apply a chart to the training data to see if the training data are really in control. Then, remove all out-of-control data and iterate until all training data are in control. Second step was to apply the control charts established from the in-control training data to test observations.

For this test, we used only independent variables for preventing a singularity. Singularity consider the case when a square matrix does not have a matrix inverse. A matrix is singular if its determinant is 0 and it will happen when the dataset contain dependent variables. The variables that we removed in this test (according to correlation matrix from description statistics chapter) are ST2, ST3, ArtBPS.

The results of the test:

	Accuracy	Sensitivity	Specificity	Precision	F1_score	Fb_score	Weighted_Accuracy	Weighted_Measure
1	0.637398	0.124125	0.901643	0.393827	0.188757	0.124972	0.194808	0.148644

Mass Univariate

Mass univariate approach is a powerful way of dealing with the problem of multiple comparisons and the analysis of a massive number of simultaneously measured dependent variables via the performance of univariate hypothesis tests. This approach conducted a t test or ANOVA on each variable separately, asking whether the activity in each variable differed significantly across conditions, and then performing a Bonferroni correction for multiple comparisons. For an effect to be significant after a Bonferroni correction, the original p value must be less than the alpha value divided by the number of comparisons.

Mas Univariate test – Full model

	Accuracy	Sensitivity	Specificity	Precision	F1_score	Fb_score	Weighted_Accuracy	Weighted_Measure
1	0.637398	0.124125	0.901643	0.393827	0.188757	0.124972	0.194808	0.148644

Logistic-Regression

In statistics, logistic regression model is a regression model where the dependent variable (DV) is categorical. Logistic regression is used in various fields and in most medical fields. In our research, the dependent variable is a binary variable, which represent true or no alarm. Logistic regression is used to predict the odds of being a case based on the values of the independent variables (predictors). The odds are defined as the probability that a particular outcome is a case divided by the probability that it is a not the case.

Full Model – without interactions

Step Forward

```
Call:
glm(formula = train$Tag ~ (HR + ArtBPS + ArtBPM + RR_mandatory +
    ST1 + ST2 + ST3 + Fio2 + cvpLog + RR_total_Log + Spo2_Log),
    family = "binomial", data = train[, 3:14])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4656  -0.4255  -0.1100   0.3019   6.9871

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.149176   1.976551   1.593 0.111100
HR            0.117126   0.001715  68.279  < 2e-16 ***
ArtBPS       -0.046128   0.002043 -22.573  < 2e-16 ***
ArtBPM       -0.010848   0.002738  -3.963 7.42e-05 ***
RR_mandatory  0.028390   0.004804   5.910 3.43e-09 ***
ST1          -0.666451   0.193033  -3.453 0.000555 ***
ST2          -0.091273   0.045470  -2.007 0.044716 *
ST3           0.782798   0.190915   4.100 4.13e-05 ***
Fio2         -0.007614   0.001517  -5.019 5.20e-07 ***
cvpLog        0.825281   0.158039   5.222 1.77e-07 ***
RR_total_Log -0.363826   0.089849  -4.049 5.14e-05 ***
Spo2_Log     -2.905330   0.397746  -7.304 2.78e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28910  on 22684  degrees of freedom
Residual deviance: 12880  on 22673  degrees of freedom
AIC: 12904

Number of Fisher Scoring iterations: 6
```

	Accuracy	Sensitivity	Specificity	Precision	F1_score	Fb_score	Weighted_Accuracy	Weighted_Measure
1	0.897514	0.827626	0.933494	0.864986	0.845894	0.82798	0.83725	0.836732

Step Backward

```
Call:
glm(formula = train$Tag ~ HR + ArtBPS + ArtBPM + RR_mandatory +
    ST1 + ST2 + ST3 + Fio2 + cvpLog + RR_total_Log + Spo2_Log,
    family = "binomial", data = train[, 3:14])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4656  -0.4255  -0.1100   0.3019   6.9871

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.149176   1.976551   1.593 0.111100
HR            0.117126   0.001715  68.279  < 2e-16 ***
ArtBPS       -0.046128   0.002043 -22.573  < 2e-16 ***
ArtBPM       -0.010848   0.002738  -3.963 7.42e-05 ***
RR_mandatory  0.028390   0.004804   5.910 3.43e-09 ***
ST1          -0.666451   0.193033  -3.453 0.000555 ***
ST2          -0.091273   0.045470  -2.007 0.044716 *
ST3           0.782798   0.190915   4.100 4.13e-05 ***
Fio2         -0.007614   0.001517  -5.019 5.20e-07 ***
cvpLog        0.825281   0.158039   5.222 1.77e-07 ***
RR_total_Log -0.363826   0.089849  -4.049 5.14e-05 ***
Spo2_Log     -2.905330   0.397746  -7.304 2.78e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28910  on 22684  degrees of freedom
Residual deviance: 12880  on 22673  degrees of freedom
AIC: 12904
```

	Accuracy	Sensitivity	Specificity	Precision	F1_score	Fb_score	Weighted_Accuracy	Weighted_Measure
1	0.897514	0.827626	0.933494	0.864986	0.845894	0.82798	0.83725	0.836732

Step Both Directions:

```
Call:
glm(formula = train$Tag ~ HR + ArtBPS + ArtBPM + RR_mandatory +
    ST1 + ST2 + ST3 + Fio2 + cvpLog + RR_total_Log + Spo2_Log,
    family = "binomial", data = train[, 3:14])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4656  -0.4255  -0.1100   0.3019   6.9871

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.149176   1.976551   1.593 0.111100
HR            0.117126   0.001715  68.279  < 2e-16 ***
ArtBPS       -0.046128   0.002043 -22.573  < 2e-16 ***
ArtBPM       -0.010848   0.002738  -3.963 7.42e-05 ***
RR_mandatory  0.028390   0.004804   5.910 3.43e-09 ***
ST1          -0.666451   0.193033  -3.453 0.000555 ***
ST2          -0.091273   0.045470  -2.007 0.044716 *
ST3           0.782798   0.190915   4.100 4.13e-05 ***
Fio2         -0.007614   0.001517  -5.019 5.20e-07 ***
cvpLog        0.825281   0.158039   5.222 1.77e-07 ***
RR_total_Log -0.363826   0.089849  -4.049 5.14e-05 ***
Spo2_Log     -2.905330   0.397746  -7.304 2.78e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28910  on 22684  degrees of freedom
Residual deviance: 12880  on 22673  degrees of freedom
AIC: 12904

Number of Fisher Scoring iterations: 6
```

	Accuracy	Sensitivity	Specificity	Precision	F1_score	Fb_score	Weighted_Accuracy	Weighted_Measure
1	0.897514	0.827626	0.933494	0.864986	0.845894	0.82798	0.83725	0.836732

Full Model – second order interactions

We examined also regression with second order interactions between some covariates that a specialized doctor recommended us, as following: HR and ArtBPM, ArtBPM and Spo2_Log and also CVP_Log with Fio2.

Step Forward

```
Call:
glm(formula = train$Tag ~ (HR:ArtBPM + ArtBPS + ArtBPM:Spo2_Log +
    RR_mandatory + ST1 + ST2 + ST3 + Fio2:cvpLog + cvpLog + RR_total_Log +
    Spo2_Log), family = "binomial", data = train[, 3:14])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3714  -0.3014  -0.0421   0.2537   8.4904

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -8.468e-01  2.144e+00  -0.395 0.692803
ArtBPS          -4.600e-02  2.192e-03 -20.991  < 2e-16 ***
RR_mandatory     1.901e-02  5.265e-03   3.610 0.000307 ***
ST1             -8.511e-01  2.189e-01  -3.888 0.000101 ***
ST2             -1.235e-01  4.870e-02  -2.536 0.011213 *
ST3              9.963e-01  2.173e-01   4.584 4.55e-06 ***
cvpLog           7.771e-01  1.812e-01   4.289 1.79e-05 ***
RR_total_Log    -6.127e-01  9.870e-02  -6.208 5.38e-10 ***
Spo2_Log         1.202e+00  4.326e-01   2.779 0.005460 **
HR:ArtBPM        2.126e-03  3.193e-05  66.586  < 2e-16 ***
ArtBPM:Spo2_Log -5.756e-02  1.094e-03 -52.615  < 2e-16 ***
Fio2:cvpLog     -1.165e-03  3.853e-04  -3.023 0.002503 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28910  on 22684  degrees of freedom
Residual deviance: 10429  on 22673  degrees of freedom
AIC: 10453
```

	Accuracy	Sensitivity	Specificity	Precision	F1_score	Fb_score	Weighted_Accuracy	Weighted_Measure
1	0.919862	0.87393	0.94351	0.888449	0.88113	0.874071	0.880255	0.880038

Step Backward

```
Call:
glm(formula = train$Tag ~ (HR:ArtBPM + ArtBPS + ArtBPM:Spo2_Log +
    RR_mandatory + ST1 + ST2 + ST3 + Fio2:cvpLog + cvpLog + RR_total_Log +
    Spo2_Log), family = "binomial", data = train[, 3:14])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3714  -0.3014  -0.0421   0.2537   8.4904

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -8.468e-01  2.144e+00  -0.395 0.692803
ArtBPS          -4.600e-02  2.192e-03 -20.991  < 2e-16 ***
RR_mandatory     1.901e-02  5.265e-03   3.610 0.000307 ***
ST1             -8.511e-01  2.189e-01  -3.888 0.000101 ***
ST2             -1.235e-01  4.870e-02  -2.536 0.011213 *
ST3              9.963e-01  2.173e-01   4.584 4.55e-06 ***
cvpLog           7.771e-01  1.812e-01   4.289 1.79e-05 ***
RR_total_Log    -6.127e-01  9.870e-02  -6.208 5.38e-10 ***
Spo2_Log         1.202e+00  4.326e-01   2.779 0.005460 **
HR:ArtBPM        2.126e-03  3.193e-05  66.586  < 2e-16 ***
ArtBPM:Spo2_Log -5.756e-02  1.094e-03 -52.615  < 2e-16 ***
Fio2:cvpLog     -1.165e-03  3.853e-04  -3.023 0.002503 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28910  on 22684  degrees of freedom
Residual deviance: 10429  on 22673  degrees of freedom
                 AIC: 10453

                 Number of Fisher Scoring iterations: 8
```

	Accuracy	Sensitivity	Specificity	Precision	F1_score	Fb_score	Weighted_Accuracy	Weighted_Measure
1	0.919862	0.87393	0.94351	0.888449	0.88113	0.874071	0.880255	0.880038

Step Both Directions

```
Call:
glm(formula = train$Tag ~ (HR:ArtBPM + ArtBPS + ArtBPM:Spo2_Log +
    RR_mandatory + ST1 + ST2 + ST3 + Fio2:cvpLog + cvpLog + RR_total_Log +
    Spo2_Log), family = "binomial", data = train[, 3:14])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3714  -0.3014  -0.0421   0.2537   8.4904

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -8.468e-01  2.144e+00  -0.395 0.692803
ArtBPS          -4.600e-02  2.192e-03 -20.991  < 2e-16 ***
RR_mandatory     1.901e-02  5.265e-03   3.610 0.000307 ***
ST1             -8.511e-01  2.189e-01  -3.888 0.000101 ***
ST2             -1.235e-01  4.870e-02  -2.536 0.011213 *
ST3              9.963e-01  2.173e-01   4.584 4.55e-06 ***
cvpLog           7.771e-01  1.812e-01   4.289 1.79e-05 ***
RR_total_Log    -6.127e-01  9.870e-02  -6.208 5.38e-10 ***
Spo2_Log         1.202e+00  4.326e-01   2.779 0.005460 **
HR:ArtBPM        2.126e-03  3.193e-05  66.586  < 2e-16 ***
ArtBPM:Spo2_Log -5.756e-02  1.094e-03 -52.615  < 2e-16 ***
Fio2:cvpLog     -1.165e-03  3.853e-04  -3.023 0.002503 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

        (Dispersion parameter for binomial family taken to be 1)

            Null deviance: 28910  on 22684  degrees of freedom
        Residual deviance: 10429  on 22673  degrees of freedom
        AIC: 10453

        Number of Fisher Scoring iterations: 8
```

	Accuracy	Sensitivity	Specificity	Precision	F1_score	Fb_score	Weighted_Accuracy	Weighted_Measure
1	0.919862	0.87393	0.94351	0.888449	0.88113	0.874071	0.880255	0.880038

We can see no differences between stepwise, backward and forward regression results. We will estimate the regression models results by Akaike information criterion (AIC). AIC is a measure of the relative quality of statistical models for a given set of data. L is the maximized value of the likelihood function for the model, k is the number of estimated parameters in the model. AIC value of the model is the following:

$$AIC = 2k - 2\ln(L)$$

We want to minimize this indicator so we will choose the model with the lowest AIC. The lowest AIC was obtained for the model with the interactions with second order, which make sense given that not only the main affect influence the exploratory variable but also the interactions between the independent variables as the specialist mentioned.

Evaluation

In previous chapters, we explored measures which allow weighting of the different error types. We chose to use the **weighted measure** with 10 times more weight to the sensitivity component over the specificity component and the models were tuned so that this indicator will be maximized.

The results of all the models, using full and partial variable sets, are shown in the following tables.

Partial variable sets models:

Model Name	Weighted Measure
NN – partial (only variables which are **not** used for decision making) 7 variables	82.82%
RF – partial (only variables which are **not** used for decision making) 7 variables	90.20%
K-means –partial (only variables used for decision making) 7 variables K=2	82.90%

Full variable set models:

Model Name	Weighted Measure
NN – full 12 variables	96.80%
RF – full 12 variables	99.99%
K-means –full 12 variables K=3	59.32%
K-means –full 12 variables K=2	1.04%
Logistic regression	83.67%
Logistic regression - Interactions	88%
Mass Univariate	14.86%
T^2 **Hotteling**	15.41%

Discussion and conclusions

In this work, we focused on achieving our goal of false alarm reduction in ICU monitors using machine learning and statistics methods. Effective implementation of ML and MSPC algorithms on the sampled medical data required a preliminary stage. First, for understanding the abnormal behaviors of the data and analyzing the patterns in the dataset we used description statistics so we examine the data distribution by visualizations. Then, we used different algorithms for explanatory variable predicting.

The data in this study was divided into train and test sets. The models were trained and tuned using CV-4 (cross-validation) over the train set, and the final models were evaluated over the "unseen" test set.

The evaluation metric we've used for comparison between the algorithms' results was a weighted measure combining the sensitivity and specificity measures. The RF full (12 variables) model test result including 11 variables for each tree and 50 trees has shown the highest sensitivity of 100%, specificity of 99.97% and weighted measure of 99.99%. This model was the best predictor for explaining the dependent variable (true alarm or no alarm).

We saw that using variables that help the decision process of caregivers while they diagnose patients' condition (heart rate, ArtBPS, ArtBPM, CVP, PAPD), improve the results of all the models. Athough useful information can be extracted from variables not used for caregivers decisions (Spo2, ST1, ST2, ST3, Fio2, RR total, RR mandatory), using only those variables produced models of lower quality results, as we've seen in the RF and NN modeling. Thus, for achieving classification models of high quality, we recommend using the variables also used by carefivers at their decision making.

For future work, it is possible to test more complex models and a combination of different algorithms to improve the results of the existing models

Appendix

Appendix 1

Alarm Type	Amount
Bradycardia 1	72
Bradycardia 1, Bradycardia-Hypotension	6
Hypovolemia 1 or 2, Hypovolemia 3	159
Hypovolemia 1 or 2, Hypovolemia 3, Bradycardia 1	3
Hypovolemia 1 or 2, Hypovolemia 3, Bradycardia 1, Bradycardia-Hypotension	4
Hypovolemia 1 or 2, Hypovolemia 3, Tachycardia	1
Hypovolemia 1 or 2, Hypovolemia 3, Tachycardia-Hypotension, Tachycardia	6
Hypovolemia 3	1224
Hypovolemia 3, Bradycardia 1	2
Hypovolemia 3, Bradycardia 1, Bradycardia-Hypotension	9
Hypovolemia 3, LV shock 2	425
Hypovolemia 3, LV shock 2, Bradycardia 1, Bradycardia-Hypotension	2
Hypovolemia 3, LV shock 2, Tachycardia-Hypotension, Tachycardia	43
Hypovolemia 3, Tachycardia	690
Hypovolemia 3, Tachycardia-Hypotension, Tachycardia	278

LV shock 2	337
LV shock 2, Bradycardia 1, Bradycardia-Hypotension	1
LV shock 2, Tachycardia-Hypotension, Tachycardia	118
Tachycardia	6261
Tachycardia-Hypotension, Tachycardia	511

Appendix 2

	Fio2	ST3	ST2	ST1	Spo2	RR_mandatory	RR_total	PAPD	CVP	ArtBPM	ArtBPS	HR
Mean	54.4	-0.46	-0.44	-0.45	95.41	13.14	15.96	20.43	13.37	75.04	112.67	101.88
Median	50	0	0.02	-0.01	97	14	15	19.04	11.03	74.18	113.27	101
SD	17.58	3.86	3.87	3.84	5.54	5.31	4.61	8.92	20.18	21.64	27.69	24.77
Skewness	1.33	-8.24	-8.22	-8.28	-3.89	-1.26	2.72	9.09	8.92	2.18	0.18	0.04
quantile 25%	40	-0.03	-0.02	-0.05	94	12	14	16.038	8.02	64.16	97.24	84
quantile 75%	60	0.03	0.06	0.04	98.5	16	18	23.055	14.03	84.2	128.31	121
Number of observations	30247	30247	30247	30247	30247	30247	30247	30247	30247	30247	30247	30247
Number of missing values	0	0	0	0	0	0	0	0	0	0	0	0

Bibliography

Benneyan, J., Lloyd, R., & Plsek, P. (2003). Statistical process control as a tool for research and healthcare improvement. *Quality and Safety in Health Care*, *12*(6), 458–464. http://doi.org/10.1136/qhc.12.6.458

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. University of California, Berkeley, 110.

Kristensen, M. S., Edworthy, J., & Denham, S. (2015). Alarm fatigue in the perception of medical soundscapes. In *European Congress and Exposition on Noise Control Engineering* (pp. 745–750). Maastricht (The Netherlands).

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

MacGregor, J. F., & Kourti, T. (1995). Statistical process control of multivariate processes. *Control Engineering Practice*, *3*(3), 403–414. http://doi.org/10.1016/0967-0661(95)00014-L

Montgomery, D. (2009). *Introduction to statistical quality control*. *John Wiley & Sons Inc.* http://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C

Qiu, P. (2013). *Introduction to statistical process control. Healthcare facilities management series*. http://doi.org/10.1137/1.9780898719765.ch21

Robert L. Mason & John C. Young. (1996). Multivariate Statistical Process Control.

Sendelbach, S., & Funk, M. (2013). Alarm Fatigue. *AACN Advanced Critical Care*, *24*(4), 378–386. http://doi.org/10.1097/NCI.0b013e3182a903f9

Srinivasu, R., Reddy, G., & Rikkula, S. (2009). Utility of Quality Control Tools and Statistical Process Control To Improve the Productivity and Quality in an Industry. *International Journal of Reviews in Computing*.

Woodall, W. H., & Tech, V. (2006). The Use of Control Charts in Surveillance. *Journal of Quality Technology*, *38*(2), 89–104. http://doi.org/10.1111/j.1600-0420.2005.00625.x