Predict hourly patient discharge probability in Intensive Care Units using Data Mining

Filipe Portela[1*], Rui Veloso[1], Sérgio Oliveira[1], Manuel Filipe Santos[1], António Abelha[2], José Machado[2], Álvaro Silva[3] and Fernando Rua[3]

[1]Algoritmi Research Centre, University of Minho, Guimarães, [2]CCTC, University of Minho, Braga, [3]Intensive Care Unit, Centro Hospitalar do Porto, Porto

*Corresponding addresses: Information System Department, University of Minho, 4800-058, Guimarães, Portugal

{cfp, mfs@dsi.uminho.pt, ruiveloso51046@gmail.com, sergiomdcoliveira@gmail.com {abelha, machado}@di.uminho.pt, moreirasilva@gmail.com, fernandorua.sci@hgsa.min-saude.pt

Abstract: The length of stay (LOS) is an important metric to manage hospital units since a correct prevision of the LOS can contribute to reduce costs and optimize resources. This metric become more fundamental in intensive care units (ICU) where controlling patient condition and predict clinical events is very difficult. A set of experiences was made using data mining techniques in order to predict something more ambitious than LOS. Using the data provided by INTCare system it was possible to induce models with a very good sensitivity (95%) in order to predict the probability of a patient be discharged in the next hour. The results achieved also allow for predicting the bed occupancy rate in ICU for the next hour. The work done represents a novelty in this area and contributes to improve the decision making process providing new knowledge in real time.

Keywords: LOS, INTCare, ICU, Data Mining, Occupancy Rate

1 Introduction

The patient Length of Stay (LOS) is one of the most important hospital indicators. This indicator allows to have a better recourses allocation and to understand for example the services occupancy rate. There is a set of studies about this subject, however they cannot be applied to Intensive Care Units (ICUs).

ICUs are recognized as being critical environments where the patients are in risk-life conditions (high risk of having a multiple organ failure). Here it is very difficult to prognostic the LOS because the patient can be admitted to an indeterminate period of time. This happens because normally the patient goes to ICU when he is in critical situation with a reserved prognosis and sometimes the physicians do not know the problem. Adding to this fact, there is a constant change of the environment and the patient condition which makes it difficult to create patterns associated to LOS. For example, for a hospitalization upper than 5 days, the risk of contracting an infection is high and consequently the LOS increase.

This situation leads to the needs of extra resources and expensive cost no standardized at patient admission. In ICU the resources are very limited and adequate to the number of patients in the Unit. Other example is the number of nurses and clinical materials necessary in an unit: it is determinate according to the number of patient hospitalized. ICU is considered the most expensive unit of a hospital. Bearing in mind that one of the Hospital goals is to reduce the cost and optimize the resources, it is very important to assess how much time will be a patient admitted in the Unit. Knowing the hospitalization period, it is possible to predict the service occupancy rate.

In the recent years, several research projects have emerged involving the development and deployment decision support systems for ICU, as well as the use of DM tasks to analyze and predict many parameters that are useful and important to the decision making process in these units. In order to help the decision support at level of patient discharge, some data mining models were induced with the objective to predict the probability of a patient be discharged in the next your. These models can predict with a good sensibility (95.88%) the patient discharge hour. The models uses data related until a maximum of five days (120 hours) of patient admission. This work was based in the study started by INTCare project. INTCare could predict the patient outcome and patient organ failure for the next hour (Portela, Santos, Machado, Abelha, & Silva, 2013). This work aims to use the same idea but predicting the discharge time in the ICU of Centro Hospitalar do Porto (CHP).

This paper is divided into eight sections. After a first introduction of the paper, the main concepts and related work are presented in section two. The following sections present the work developed, namely the data sources used, the data transformation made and the data mining models induced. In the sections five and six, the main results are presented and discussed. Finally, a set of considerations are done and some future work presented.

2 Background

2.1 Length of stay

Planning and provisioning assume an important role in hospital resource management. One way to perform this management is doing the analysis of metrics that evaluate and optimize the resource management. The Length of Stay (LOS) in the hospital has proved to be a very good measure since it measures the consumption of hospital resources to a particular patient, hospital or group of patients (Marshall, Vasilakis, & El-Darzi, 2005).

The LOS is the number of days that a patient is subjected to treatment, since their admission date until the discharge date. This value can be accounted in the hospital or at home, depending on where the patient is receiving health cares. The use of this indicator allows not only knowing the amount of expenses that the patient represents to the hospital, but also evaluating the quality of health care services provided when it is possible.

Since these measurements have to correct in order to maximize the optimization of resources, it becomes crucial to develop models capable to predict the LOS of inpatient, primarily in critical care units like the Intensive Care Units (ICU).

There are several studies using Data Mining (DM) tasks and algorithms as method to predict the LOS, however none of them are specific to ICU (Azari, Janeja, & Mohseni, 2012; Liu et al., 2006; Tanuja, Acharya, & Shailesh, 2011).

2.2 Bed Occupancy Rate

The Bed Occupancy Rate (BOR) is the ratio between the number of patients and the number of beds in a given period, usually daily or yearly. The way to calculate this indicator can be expressed by the following expression:

$$BedOccupancyRate = \frac{Total\ Number\ Patients\ per\ period}{Bed\ Days\ Available\ per\ period} * 100(\%)$$

The analysis of this indicator is very important because it allows get around a problem that sometimes occurs in ICU, where the bed occupancy rate is higher than 100%. In ICU this problem assumes even a greater importance, because more patients can arrive at any time to the unit.

Studies in the area of hospital management have already come out for few years, but only recently investigators are using techniques such as Data Mining. El-Darzi (El-Darzi, Vasilakis, Chaussalet, & Millard, 1998) suggested an approach to simulate the flow of patients in order to evaluate the number of empty and locked beds in a geriatric department of a hospital. This study revealed several problems in the early stages, since the generated models were unreliable. After some modifications the investigators achieved results more reasonable but still far from desired outcomes.

More recently it has been performed studies that pretend to analyze the BOR using DM techniques. Alapont (Alapont et al., 2005) presented an investigation in hospitals resource management, with the intend to improve the utilization of hospital resources, avoid situations of BOR superior to 100% and optimize of surgical blocks schedules. Classification and regression were used as techniques and linear regression, support vector machines, decision trees and artificial neural network as algorithms.

Zhu (Zhu, Hen, & Teow, 2012) have focused in bed management in order to help planning the transfer and allocation of beds in the National University Hospital of Singapore. In this work the authors used several DM techniques like decision trees, artificial neural networks and logistic regression.

2.3 INTCare Related Work

INTCare is a research project developed to Intensive Care Units. The main goal of this project is to change the way of medical staff take decision and create new knowledge at the moment they need to decide. With this project, it was possible to change the way that data is acquired and processed. Now the entire process of extract transforming and loading (ETL) is executed automatically and in real-time. This new reality brings the possibility of having a set of new data available electronically, allowing the automation of clinical tasks the development of decision and prevision models.

After some offline tests (Vilas-Boas, Santos, Portela, Silva, & Rua, 2010), it was possible to induce in real-time data mining models using online-learning. These models are capable to predict the patient organ failure (cardiovascular, coagulation, hepatic, renal and respiratory) and patient outcome, only using a few variables: case mix, SOFA, critical events and some ratios. This situation was only possible due the introduction of intelligent agents to perform automatically the ETL tasks. The results attained (table 1) were very satisfactory and motived future developments.

To evaluate the results, it was developed a quality measure based in three metrics: Total Error $\leq$ 40%, Sensitivity $\geq$ 85% and Accuracy $\geq$ 60%. Sensitivity is the most important measure because in the medicine the physicians want models sensible to one result; they prefer to know that a patient will have a problem and act according to the provided information than do nothing and a problem appear. In table 1 it is possible to observe the models which achieved the quality measure and the respective results.

The work developed in this project gave origin to a pervasive (Varshney, 2007) intelligent decision support system called INTCare (Portela, Aguiar, Santos, Álvaro Silva, & Rua, 2013; Portela, Gago, et al., 2013; Portela, Santos, et al., 2013; Portela, Santos, Machado, et al., 2012).

The variables used to predict the organ failure and patient outcome will be adapted to the current study.

3 Materials and Methods

3.1 Data Sources

The next figure (figure 1) shows the data acquisition architecture and the agents associated to the process (Portela, Gago, et al., 2011; Portela, Santos, & Vilas-Boas, 2012). The data is collected by the agents, are stored into tables and then processed and sent to the database server. The data are provided by five data systems:

a) Laboratory (LR);

b) Pharmacy System (PHS);

c) Electronic Nursing Records (ENR);

d) Electronic Health Records (EHR);

e) Vital Signs Monitor (VS).

The pre-processing agent is responsible to execute the tasks presented in the following section.

3.2 Data pre-processing and transformation

a. **Data pre-processing**

This phase is applied to values from three different data sources: electronic nursing records, bedside monitors and laboratory.

The pre-processing agent is responsible to execute a set of tasks that include validate all the collected data, prepare the data mining input table (DMIT) (creation of the model input structure for each patient), validate the sources input data, process the variables with all-day frequency.

The values collected are validated in accordance with the limits defined by the intensive care units (Portela, Santos, et al., 2011).

b. **Data Transforming**

The INTCare system by means of its agents executes automatically and real-time transformation tasks on the input attributes so there can be developed predictive models. These transformations include the discretization and normalization of the original values of the variables.

The CM and SOFA variables as well as the accumulated critical events (ACE) and ratios are prepared and inserted into the DMIT. All the values are considered according to a range of values (Min and Max). In the case of the values be real numbers, some ranges were introduced according to the normality and importance of the value for the ICU.

The first transformation is an analytic task where the values collected are transformed according to a set of rules.

For the Case Mix attributes is only used one value to represent each case. For the SOFA, the DM models are based on the worst SOFA value received hourly. This process is applied to the variables presented in table 2.

The second phase of the transformation uses Critical Events (CE). The values are categorized as normal (0) or critical (1) respecting some limits (Silva, Cortez, Santos, Gomes, & Neves, 2008). If the values is within the range is classified as normal when is not is classified as critical. Next step is to determinate the typology of the event. There are two typologies, one that corresponds to values that are out of the normal range and another that refers to critical values independently the duration of the event. During the execution of the Data Mining agent there are a set of tasks that are performed namely the identification of value importance and critical event typology, calculation of the ACE and ratios and the elimination of null and wrong values (Portela, Gago, et al., 2013).

A discretization technique was considered in recent experiment. Using an interval (Min and Max) the values were grouped and categorized. By the use of this technique were defined sets considering two characteristics: the respective average and the higher value of the data collected. The defined categories represent the significance of each value to the patient. The ranges created are in conformity with the Clinical Global Impression Severity scale (CGI-S) (W. Guy Modified From: Rush J, 2000). This scale goes from 0 to 7 where 0 defines an inexistence of severity ill for the patient and 7 represents an extreme condition. The criteria used to define percentages group the most part of the patient in values from 0 to 5. The severe cases are assigned to levels 6 and 7. Table 3 represents the rules defined to discretize each one of the continuous values. On the top are the different sets and on the left are identified the variables. The attributes of R1 are determined by the rows (R1 BP Min to R1 TOT Max).

R2 attributes ($BP, O2, HR\ and\ Total$) follow the same rule, e.g. for the first level (1) they range from 0.00 (0%) to 0.10 (10%).

Finally, all ACE attributes are grouped in accordance to their importance and the number of occurrences. These values were defined by ICU experts but can be modified in the future. Resuming, the first task consists in a data validation by the system procedures. This procedure verifies if the values collected are within ICU ranges and if they have assigned a correct patient id. In the second task, the Data Mining Input table (DMIT) is created. This table is composed by 120 rows (hours) by patient and have one column for each DM variable and other to the target column. The target column (discharge) is filled hourly. An agent verifies if the patient is admitted in the ICU in the current hour. If the result is true the agent insert 0 in the column, otherwise he insert 1 in all rows until last row, corresponding to a total of 120 hours. For example, imagining that a patient is discharged at the seventieth hour of admission, the agent will put 0 in all hours between 1 and 70 and in the others (71 to 120) the agent will put the value 1. This 1 reflects that the patient is not hospitalized in the respective hour.

The third task takes into account the variables used by DM and fill the DMIT table with:

- Case Mix (CM) - Age, Admission type, Admission from;
- SOFA - Cardiovascular, Respiratory, Renal, Liver, Coagulation, neurologic;
- Critical Events Accumulated (ACE) - ACE of Blood Pressure (BP), ACE of Oxygen Saturation (SO2), ACE of Heart Rate (HR) and Total ACE;
- Ratios1 (R1) - ACE of BP/elapsed time of stay, ACE of SO2/elapsed time of stay , ACE of HR/elapsed time of stay, Total of ACE / elapsed time of stay;
- Ratios2 (R2) - ACE of BP / max number of ACE of BP , ACE of SO2/ max number of ACE of SO2 , ACE of HR / max number of ACE of HR, Total ACE/ max number of total ACE.
- Ratios (R) is a union of the two ratios set: R = R1 U R2. Where
- ACE is the number of accumulated Critical Events per patient during their admission by hour and type;
- elapsed time of stay is the number of hours since patient admission until now;
- max number of ACE is the maximum number of ACE verified by a patient per hour.

These values are refreshed anytime a worse value appears.

3.3 Data Mining

During this study, it was experimented hundreds of models using variables suggested by the literature review and other obtained in patient admission. Other models were developed using the worst values of a patient in the first 24 hours. It was induced models combining a set of techniques and using: Patient Age and Sex, Admission Type, Admission from, admission situation (transplant, hepatic problems and others). Although various models have been tested, none of them allowed achieving satisfactory results - best accuracy was around 70%.

Due to the unsatisfactory results obtained a new set of data was explored in order to achieve the target. Using the data initially created to develop the models present in section 2, the results achieved were totally different.

a. **Techniques description**

To this study it was used Oracle Data Mining (ODM) and considered three of DM techniques available: Decision Trees (DT), Support Vector Machine (SVM) and Naive Bayes (NB).

The selection of Data mining techniques was based in three distinct aspects: interpretability of the models, engine efficiency and the suitability for small datasets. SVM achieved the second and third goals. DTs and NBs met the first two goals. In the case of NBs, they also can present good results using a reduced number of data.

i. Decision Trees

A Decision Tree (DT) consists in a form of representing a set of rules that follows a hierarchy of values and classes. They express a simple and conditional logic and they can be easily interpreted by users (Santos & Azevedo, 2005). The DTs are compose by nodes and branches, where the connections between nodes are established by the branches that correspond to the value of the attributes. The nodes on the bottom of the tree are known as leafs and indicate the classes in which a row can be classified. The top node is known as root and contains all the training examples that can be divided in classes. All the nodes except the leafs are known as decision nodes because taken decisions are executed in this nodes. All the decision nodes have a finite number of children nodes and that number is equal to the number of values that a characteristic can assume (Cios, 2007). The models developed based in Decision Trees (DT) produce rules in the form of IF predictive information THEN target (Concepts, 2005). This models uses CART algorithm (Breiman, Friedman, Olshen, & Stone, 1984) and Gini coefficient. CART is a decision tree learning technique that produces DT as a binary recursive partitioning procedure capable of processing continuous and nominal attributes as targets and predictors (Steinberg & Colla, 1997).

ii. Nave Bayes

The Bayes theory gives a structure for statistical methods in the classification of patterns in classes based on probabilities (Cios, 2007) This theory appears by the work realized by Tomas Bayes (Bellhouse, 2004) but effectively was the French mathematician Pierre Simon de Laplace who developed the theorem as it is known and used (Santos & Azevedo, 2005). The Bayes theorem is expressed by the following form:

$$P(H|X) = \frac{P(H|X) * P(H}{P(X)},$$

Where,

$P(H|X)$ represents the a *posteriori* probability,

H is conditioning of X, i.e. P(X—H) represents the confidence in X, the conditional probability represents the occurrence of event X conditioning the occurrence of event H.

$P(H)$ represents the a *priori* probability of H.

The a posterior probability is slightly different from the a priori because is based in more information where $P(H)$ is independent of X (Santos & Azevedo, 2005) .

The classification method of Nave Bayes was developer from the Bayes theorem (Tuffry, 2011). Nave Bayes provides an simple approach, with clear semantics in the representation, using and learning of probabilistic knowledge (Witten & Frank, 2005). There are sets of data where the Nave Bayes arent capable of getting good results because the attributes are treated like they are totally independent. The addition of redundant attributes distorts the learning process (Witten & Frank, 2005).

iii. Support Vector Machines

The Support Vector Machines (SVM) are a method used for the classification of linear and non-linear data. This method uses the mapping of non-linear data to transform the training data in higher dimensions. In this new dimensions the method searches for the optimal hyperplane separating it linearly, creating a decision boundary. So that the non-linear mapping is appropriated to a sufficient higher dimension, the dada from two classes can be divided by a hyperplane. The SVM find this hyperplane by support vectors and margins defined by support vectors (Han, Kamber, & Pei, 2006). The output of an SVM binary classification model using (Milenova, Yarmus, & Campos, 2005) is given by:

$$f_i = b + \Sigma_{j=1}^{m} \; a_j \; y_j \; K(x_j, x_i),$$

Where,

f_i(also called margin) is the distance of each point (data record) to the decision hyperplane defined by setting $f_i = 0$;

b is the intercept;

$\alpha$is the Lagrangian multiplier for the j^{th} training data record x_j;

$y_j$is the corresponding target value (± 1).

b. **Configurations**

Table 4 presents the configurations considered for each one of the techniques. For each parameter is indicated whether the used value is a default value (Default) or an user-specified value (Input).

c. **Training and Testing Datasets**

In order to evaluate the 21 models induced, an automatic and real-time test phase was performed using the data transformed. The original dataset was divided into two data sets using the holdout sampling method. 70% of data were considered for training and 30% for testing (stratified by the target). The data used in the models corresponds to admissions / discharges made in ICU of CHP between 2012.02.01 and 2013.07.12:

- Period in analysis: 527 days;
- Number of patients: 249;

- Number of records: 21886;
- Time Frame Considered: the number of inpatient days from 1 to n ($n \leq 120$ only first five days are considered in the maximum). Invalid records were not considered based on the following criteria:
- Exclusion criterion I: Patient with intermittent data, i.e., the collecting system failed at least one hour in a continuous way;
- Exclusion criterion II: Existence of null values.

Observing table 5 is possible to identify that the used sample have fields with quite heterogeneous distributions because no value of the variation coefficient is inferior to 20%, value that is used to determine if a distribution is heterogeneous or homogeneous. This table present the minimum value (MIN), maximum value (Max), average (AVG), standard deviation (STDDEV) and coefficient of variance (VC). Table 6 presents the distribution of the classes (in percentage) for each one of the independent variables. For example the admission type is divided in two classes: programed (21.74%) and urgent (78.26%) the Respiratory Sofa is 0 in 31.86% of the cases and more (1) in 68.14%. Figure 2 represents the target distribution. The discharge patients are represented by the value 1, whereas the patients represented by the 0 value are all the ones that are hospitalized.

d. **Models developed**

In this paper only it is presented the second round of tests, where was induced 21 models using online-learning and combining 7 scenarios (s) x 1 target (z) x 3 techniques (t):

s:

1 = {CASE MIX}

2 = {CASE MIX, ACE, R}

3 = {CASE MIX, ACE, R1}

4 = {CASE MIX, ACE, SOFA}

5 = {CASE MIX, ACE, SOFA, R}

6 = {CASE MIX, ACE, SOFA, R2}

7 = {CASE MIX, ACE, SOFA, R1}

z:

1 = discharges

t:

1 = Support Vector Machine

2 = Decision Trees

3 = Naïve Bayes

Although the classes present in table 3 were created the induced models were attained without the use of these classes, i.e., for those variable it was used the real values.

Each model is induced automatically and in real-time using streamed data. The data mining engine uses the data present in input table. This table contains tuples of the type:

$DMIT = < pid,\ date,\ hour,\ v_{ace\_bp},\ v_{aceBP\_time},\ v_{aceBP\_max}, v_{ace\_hr},\ v_{acehr\_time},\ v_{acehr\_max}, v_{ace\_spo2},\ v_{acespo2\_time},\ v_{acespo2\_max}, v_{total\_ace},\ v_{totalace\_time},\ v_{totalace\_max},\ v_{age},\ v_{adminF},\ v_{adminT},\ v_{sofa\_cardio},\ v_{sofa\_resp},\ v_{sofa\_renal},\ v_{sofa\_coag},\ v_{sofa\_hepa}>$

Where,

pid is the patient identification;

date is the date of the values;

hour is the number of hours elapsed since the patient admission;

$v_{ace\_bp...}$ v_{sofa_hepa} are the values obtained for each patient and date.

The next representation demonstrates an instantiation for the model M4 which corresponds to SVM and discharge target:

$\mathrm{DMIT}_{svmDischargePidDate} = <$ discharge, pid, date, hour, v_{age}, v_{adminF}, v_{adminT}, v_{sofa_cardio}, v_{sofa_resp}, v_{sofa_renal}, v_{sofa_coag}, v_{sofa_hepa}, v_{sofa_neuro}, v_{ace_bp},v_{ace_hr},v_{ace_spo2},v_{aceur_max},v_{total_ace} $>$

After the DMIT is filled and validated the models are induced online by the DM agent. This agent runs whenever a request is sent or when it verifies that the performance of the models is decreasing

4 Results

In order to evaluate the results the metrics: sensitivity, accuracy and specificity were calculated. As mentioned before, sensitivity is the most adequate metric in medicine area. Table 7 and 8 present the results achieved by Data Mining models. Table 7 presents the top 3 models having in account the sensitivity. Table 8 presents the top 3 models for each metric. This approach of presenting results allows to choose the models according what the decision maker want to predict predicted: models equilibrated (accuracy), good to predict discharge (sensitivity) good to predict admission (specificity). After make a correct prevision for all the patient admitted in ICU it is possible determine the Bed occupancy rate for the next 24 hours, having in account the forecast values. Always some new patient arrive or depart from the ICU the Data Mining model is executed and the rate recalculated. This procedure is represented by the following algorithm:

Algorithm *- Bed occupancy rate*

```
Requires: hour, number of beds
Function Bed Occupancy Rate[hour]
 If hour < 24 then
 For all beds do
 Identify Bed occupancy
If bed is empty Then
Insert in Bed N value 0
 Else
Insert in Bed N value 1
 End if
 Next
HourOcupacy[hour] = (total beds occupied/bed total)
 End if
 Return HourOcupacy[hour]
End function
```

5 Discussion and Future Work

The models which present best result was obtained using the decision trees and combining case mix, accumulated critical events and sofa variables as independent variables. This model gives very good confidence when it is necessary understand if a patient will be or not hospitalized in next hours. The model is sensible to predict the patient discharge hour, presenting a result near of 96%. This value represents that in 100 cases the model can hit in 96 cases the last hour of patient hospitalization. In figure 3 it is possible observe the ROC for the best models which present values near the perfect value (1.00).

During this work it was explored a high number of data mining models including regression, classification and clustering techniques. For the first time interesting results to predict LOS in critical units were obtained after an extensive experimental work. The attained results allow conclude that it is possible to predict the patient discharge in a hourly base using the same set of data used to predict organ failure and patient outcome. This makes possible to combine the predicted number of patient hospitalized and the number of unit beds available in order to determine the bed occupancy rate. To the medic community this work represent a novelty and opens new paths in order to support the decision making process at level of the cost and resources optimization. In the future the models will be optimized considering a set of other variables combined with those used in this study. The best model will be included in the INTCare system.

Acknowledgements

This work has been supported by FCT Fundacão para a Ciência e Tecnologia in the scope of the project: PEst-OE/EEI/UI0319/2014. The authors would like to thank FCT (Foundation of Science and Technology, Portugal) for the financial support through the contract PTDC/EEI-SII/1302/2012 (INTCare II).

Referências

Alapont, J., Bella-Sanjun, A., Ferri, C., Hernndez-Orallo, J., Llopis-Llopis, J. D., & Ramrez-Quintana, M. J. (2005). Specialised tools for automating data mining for hospital management [Conference Proceedings]. In (p. 7-19).

Azari, A., Janeja, V. P., & Mohseni, A. (2012). Predicting hospital length of stay (phlos): A multi-tiered data mining approach [Conference Proceedings]. In (p. 17-24). IEEE.

Bellhouse, D. R. (2004). The reverend thomas bayes, frs: A biography to celebrate the tercentenary of his birth [Journal Article]. Statistical Science, 19(1), 3-43.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression [Journal Article]. Wadsworth, Belmont, CA.

Cios, K. J. (2007). Data mining: a knowledge discovery approach [Book]. Springer.

Concepts, O. D. M. (2005). 11g release 1 (11.1) [Journal Article]. Oracle Corp, 2007.

El-Darzi, E., Vasilakis, C., Chaussalet, T., & Millard, P. H. (1998). A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department [Journal Article]. Health Care Management Science, 1(2), 143-149.

Han, J., Kamber, M., & Pei, J. (2006). Data mining: concepts and techniques [Book]. Morgan kaufmann.

Liu, P., Lei, L., Yin, J., Zhang, W., Naijun, W., & El-Darzi, E. (2006). Healthcare data mining: predicting inpatient length of stay [Journal Article].

Marshall, A., Vasilakis, C., & El-Darzi, E. (2005). Length of stay-based patient flow models: recent developments and future directions [Journal Article]. Health Care Management Science, 8(3), 213-220.

Milenova, B. L., Yarmus, J. S., & Campos, M. M. (2005). Svm in oracle database 10g: removing the barriers to widespread adoption of support vector machines [Conference Proceedings]. In (p. 1152-1163). VLDB Endowment.

Portela, F., Aguiar, J., Santos, M. F., Álvaro Silva, & Rua, F. (2013). Pervasive intelligent decision support system - technology acceptance in intensive care units [Book Section]. In Springer (Ed.), Advances in intelligent systems and computing. Springer.

Portela, F., Gago, P., Santos, M. F., Machado, J., Abelha, A., Silva, A., & Rua, F. (2013). Pervasive real-time intelligent system for tracking critical events in intensive care patients [Journal Article].

Portela, F., Gago, P., Santos, M. F., Silva, A., Rua, F., Machado, J., ... Neves, J. (2011). Knowledge discovery for pervasive and real-time intelligent decision support in intensive care medicine [Conference Paper]. Springer - Accepted for publication.

Portela, F., Santos, M. F., Gago, P., Silva, A., Rua, F., Abelha, A., ... Neves, J. (2011). Enabling real-time intelligent decision support in intensive care [Conference Proceedings]. In EUROSIS-ETI (Ed.), 25th european simulation and modelling conference- esm'2011 (Vol. I, p. 446 pages).

Portela, F., Santos, M. F., Machado, J., Abelha, A., & Silva, A. (2013). Pervasive and intelligent decision support in critical health care using ensembles [Book Section]. In Information technology in bio-and medical informatics (p. 1-16). Springer Berlin Heidelberg.

Portela, F., Santos, M. F., Machado, J., Silva, A., Rua, F., & Abelha, A. (2012). Intelligent data acquisition and scoring system for intensive medicine [Book Section]. In Springer (Ed.), Lecture notes in computer science - information technology in bio- and medical informatics (Vol. 7451/2012, p. 1-15). Viena, Austria. doi: 10.1007/978-3-642-32395-9_1

Portela, F., Santos, M. F., & Vilas-Boas, M. (2012). A pervasive approach to a real-time intelligent decision support system in intensive medicine [Book Section]. In Springer (Ed.), Communications in computer and information science (Vol. 0272, p. 14).

Santos, M. F., & Azevedo, C. S. (2005). Prembulo [a]”data mining: descoberta de conhecimento em bases de dados” [Book]. FCA editores.

Silva, A., Cortez, P., Santos, M. F., Gomes, L., & Neves, J. (2008). Rating organ failure via adverse events using data mining in the intensive care unit [Journal Article]. Artificial Intelligence in Medicine, 43(3), 179-193. Retrieved from http://www.sciencedirect.com/science/article/B6T4K-4SHN0CW-1/2/030ff2c0163795a31082b76e8ba166f9 doi: doi:DOI:10.1016/j.artmed.2008.03.010

Steinberg, D., & Colla, P. (1997). Cart: classification and regression trees [Journal Article]. San Diego, CA.

Tanuja, S., Acharya, D. U., & Shailesh, K. R. (2011). Comparison of different data mining techniques to predict hospital length of stay [Journal Article]. JOURNAL OF PHARMACEUTICAL AND BIOMEDICAL SCIENCES, 7(7).

Tuffry, S. (2011). Data mining and statistics for decision making [Book]. John Wiley and Sons.

Varshney, U. (2007). Pervasive healthcare and wireless health monitoring [Journal Article]. Mobile Networks and Applications, 12(2), 113-127.

Vilas-Boas, M., Santos, M. F., Portela, F., Silva, A., & Rua, F. (2010). Hourly prediction of organ failure and outcome in intensive care based on data mining techniques [Conference Paper].

W. Guy Modified From: Rush J, e. a. (2000). Clinical global impressions (cgi) scale [Journal Article]. Psychiatric Measures, APA.

Witten, I. H., & Frank, E. (2005). Data mining: Practical machine learning tools and techniques [Book]. Morgan Kaufmann.

Zhu, Z., Hen, B. H., & Teow, K. L. (2012). Estimating icu bed capacity using discrete event simulation [Journal Article]. International journal of health care quality assurance, 25(2), 134-144.

Tables and Figures

The results attained in the previous work (table 1) were very satisfactory and motived future developments.

To evaluate the results, it was developed a quality measure based in three metrics: Total Error ≤ 40%, Sensitivity ≥ 85% and Accuracy ≥ 60%. Sensitivity is the most important measure because in the medicine the physicians want models sensible to one result; they prefer to know that a patient will have a problem and act according to the provided information than do nothing and a problem appear.

In table 1 it is possible to observe the models which achieved the quality measure and the respective results.

Tabela 1: Data mining results

Target	**Comply the quality measure**	**Sensitivity**	**Accuracy**	**Specificity**	**Terror**
Cardiovascular	YES	**97,95** ± 0,31	**76,81** ± 2,35	**41,81** ± 5,75	**23,19** ± 2,35
Coagulation	YES	**91,20** ± 3,57	**65,69** ± 3,83	**49,61** ± 6,15	**34,31** ± 3,84
Hepatic	NO	69,24 ± 9,41	**82,89** ± 2,57	**87,34** ± 3,22	**17,10** ± 2,57
Outcome	YES	**99,77** ± 0,33	**63,58** ± 3,11	49,58 ± 4,90	**36,42** ± 3,11
Renal	NO	77,17 ± 12,41	43,08 ± 4,66	43,08 ± 4,66	49,09 ± 5,39
Respiratory	NO	67,11 ± 5,67	**63,86** ± 4,27	**60,39** ± 6,75	**36,14** ± 4,27

The first transformation is an analytic task where the values collected are transformed according to a set of rules.

For the Case Mix attributes is only used one value to represent each case. For the SOFA, the DM models are based on the worst SOFA value received hourly. This process is applied to the variables presented in table 2.

Tabela 2: Transformation table rules

ID		***Variable***	***Min***	***Max***	***Value***
	Age	Age	18	46	1
			47	65	2
			66	75	3
			76	130	4
	Admission Type	Urgent	-	-	u
		Programmed	-	-	p
	Admission From	Chirurgic	-	-	1
		Observation	-	-	2
		Emergency	-	-	3
		Nursing Room	-	-	4
		Other ICU	-	-	5
		Other Hospital	-	-	6
		Other Situation	-	-	7
SOFA	Cardiovascular	BP (mean)	0	70	1
		Dopamine	0.01	-	1
	Renal	Dobutamine	0.01	-	1
		Epi / Norepi	0.01	-	1
	Respiratory	Creitanine	1.2	-	1
		Po2/Fio2	0	400	1
	Hepatic	Bilirubin	1.2	-	1
	Coagulation	Platelets	0	150	1
ACE			0	$+\infty$	SET
R1			0	1	SET
R2			0	1	SET

Table 3 represents the rules defined to discretize each one of the continuous values. On the top are the different sets and on the left are identified the variables. The attributes of R1 are determined by the rows ($R1\ BP\ Min\ to\ R1\ TOT\ Max$).

R2 attributes (BP, O2, HR and Total) follow the same rule, e.g. for the first level (1) they range from 0.00 (0%) to 0.10 (10%).

Tabela 3: Discretization sets of Data Mining Input

SET		***0***	***1***	***2***	***3***	***4***	***5***	***6***	***7***
R1 BP	Min	-0.1	0.000	0.010	0.021	0.041	0.062	0.082	0.123
	Max	0.000	0.010	0.021	0.041	0.062	0.082	0.123	2.000
R1 O2	Min	-0.1	0.000	0.018	0.036	0.072	0.108	0.144	0.216
	Max	0.000	0.018	0.036	0.072	0.108	0.144	0.216	2.000
R1 HR	Min	-0.1	0.000	0.004	0.008	0.015	0.023	0.030	0.045
	Max	0.000	0.004	0.008	0.015	0.023	0.030	0.045	2.000
R1 TOT	Min	-0.1	0.000	0.020	0.041	0.081	0.122	0.162	0.243
	Max	0.000	0.020	0.041	0.081	0.122	0.162	0.243	2.000
R2	Min	-0.1	0.000	0.100	0.250	0.500	0.750	0.900	1.000
	Max	0	0.100	0.250	0.500	0.750	0.900	1.000	2.000
ACE	Min	-0.1	0	3	5	8	10	12	15
	Max	0	3	5	8	10	12	15	50

Table 4 presents the configurations considered for each one of the techniques. For each parameter is indicated whether the used value is a default value (Default) or an user-specified value (Input).

Tabela 4: Models configurations

Technique	***Setting Name***	***Setting Value***	***Setting Type***
DT	Minrec Node	10	Input
	Max Depth	7	Input
	Minpct Split	0.1	Input
	Imyuritp Metric	Gini	Input
	Minrec Split	20	Input
	Minoct Npde	0.05	Input
	Prep Auto	On	Input
NB	Pairwise Threshold	0	Input
	Singleton Threshold	0	Input
SVM	Conv tolerance	0.001	Input
	Active learning	Enable	Input
	Kernel function	Linear	Default
	Complexity factor	0.142831	Default
	Prep auto	On	Input

Observing table 5 is possible to identify that the used sample have fields with quite heterogeneous distributions because no value of the variation coefficient is inferior to 20%, value that is used to determine if a distribution is heterogeneous or homogeneous. This table present the minimum value (MIN), maximum value (Max), average (AVG), standard deviation (STDDEV) and coefficient of variance (VC).

Tabela 5: Data analysis

	MIN	MAX	AVG	STD DEV	VC
Hours	2	120	55.32	34.48	62.33%
Respiratory	0	1	0.68	0.47	69.12%
Coagulation	0	1	0.42	0.49	116.67%
Renal	0	1	0.18	0.39	216.67%
Hepatic	0	1	0.15	0.36	240.00%
Cardiovascular	0	1	0.64	0.48	75.00%
Total_aCE	0	9	3.12	6.31	201.60%
Total_Max	0.030	3.22	0.34	1.68	494.12%
Total_Hour	0.008	1	0.06	0.10	166.67%
BP	0	9	0.62	1.99	320.97%
BP_Max	0.002	9.67	1.08	5.10	472.22%
BP_Hour	0.008	0.50	0.01	0.03	300.00%
HR	0	9	0.80	1.95	243.75%
HR_Max	0.007	3.20	0.21	1.13	619.05%
HR_Hour	0.001	0.50	0.01	0.03	300.00%
O2	0	9	1.71	4.41	257.89%
O2_Max	0.004	5.45	0.48	2.68	558.33%
O2_Hour	0.003	1	0.03	0.08	266.67%

Table 6 presents the distribution of the classes (in percentage) for each one of the independent variables. For example the admission type is divided in two classes: programed (21.74%) and urgent (78.26%) the Respiratory Sofa is 0 in 31.86% of the cases and more (1) in 68.14%.

Tabela 6: Classes distribution

Variable	Values	Percentage
Admission Type	p	21.74%
	u	78.26%
Age	1	18.07%
	2	39.15%
	3	17.14%
	4	25.63%
Admission From	1	50.36%
	2	0.32%
	3	18.46%
	4	13.52%
	5	1.78%
	6	0.75%
	7	14.80%
Respiratory	0	31.86%
	1	68.14%
Coagluation	0	57.57%
	1	42.43%
Renal	0	81.58%
	1	18.42%
Hepatic	0	84.79%
	1	15.21%
Cardiovascular	0	37.60%
	1	62.40%

In order to evaluate the results the metrics: sensitivity, accuracy and specificity were calculated. As mentioned before, sensitivity is the most adequate metric in medicine area.

Table 7 and 8 present the results achieved by Data Mining models. Table 7 presents the top 3 models having in account the sensitivity.

Tabela 7: Three best models

Model	Technique	Accuracy	Sensitivity	Specificity
M3	DT	0.834941	0.958797	0.814446
M6	DT	0.819020	0.944321	0.798145
M3	SVM	0.805025	0.935412	0.783302

Table 8 presents the top 3 models for each metric.

Tabela 8: Three best models for each metric

Accuracy	Sensitivity	Specificity
M3 – DT	M3 – DT	M5 – DT
0.834941	0.958797	0.814446
M5 – DT	M6 – DT	M4 – DT
0.831107	0.944321	0.798145
M4 – DT	M3 – SVM	M3 – DT
0.829676	0.935412	0.783302

The next figure (figure 1) shows the data acquisition architecture and the agents associated to the process [14, 15]. The data is collected by the agents, are stored into tables and then processed and sent to the database server. The data are provided by five data systems:

a) Laboratory (LR);

b) Pharmacy System (PHS);

c) Electronic Nursing Records (ENR);

d) Electronic Health Records (EHR);

e) Vital Signs Monitor (VS).

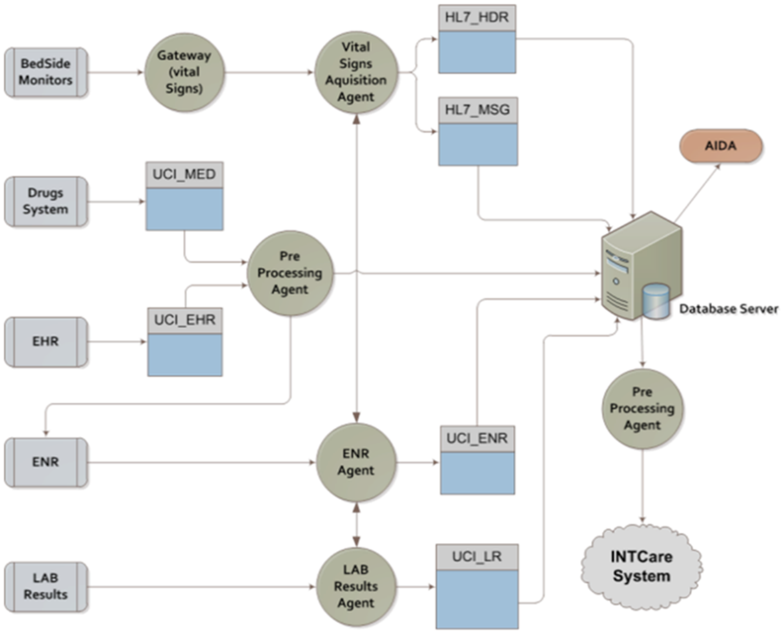

Figura 1: Data acquisition architecture.

Figure 2 represents the target distribution. The discharge patients are represented by the value 1, whereas the patients represented by the 0 value are all the ones that are hospitalized.

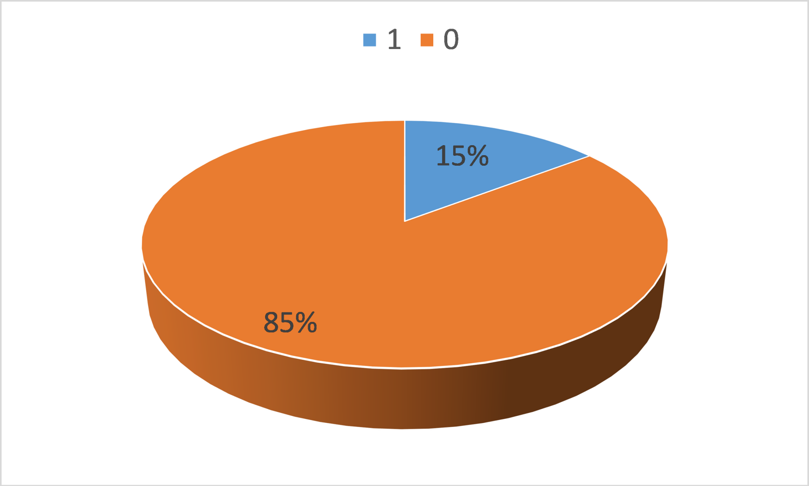

Figura 2: Target distribution.

In figure 3 it is possible observe the ROC for the best models which present values near the perfect value (1.00).

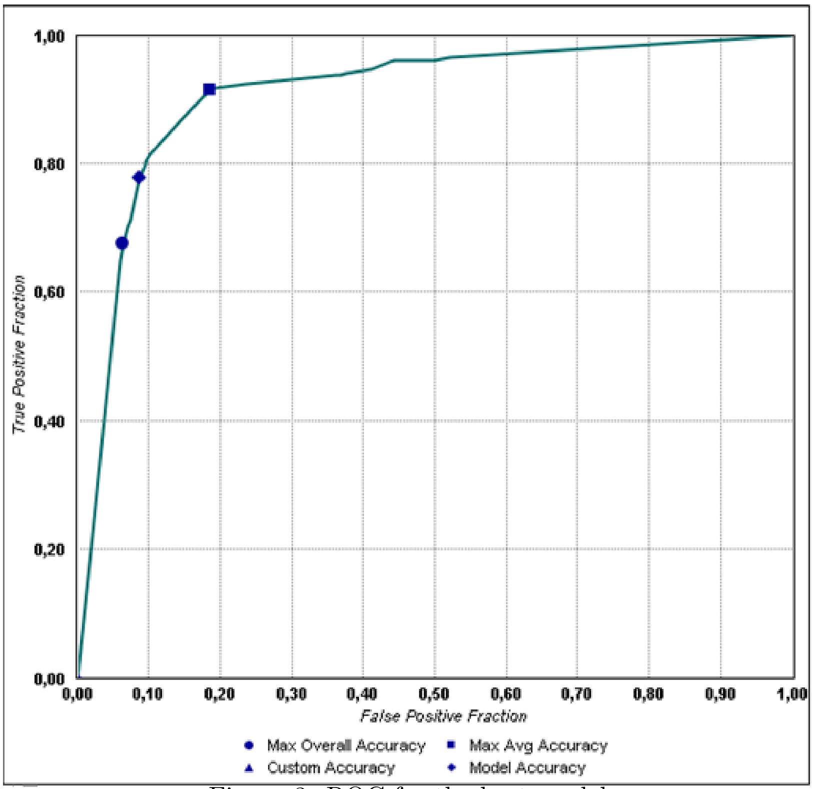

Figura 3: ROC for the best model.