Real-time Predictive Analytics for Sepsis Level and Therapeutic Plans in Intensive Care Medicine

João M. C. Gonçalves *, Filipe Portela *, Manuel F. Santos *, Álvaro Silva **, José Machado *, António Abelha ***, Fernando Rua ****

**Algoritmi Centre, University of Minho, Guimarães, Portugal*
jgoncalves@di.uminho.pt; {cfp, mfs}@dsi.uminho.pt

*** Serviço Cuidados Intensivos, Centro Hospitalar do Porto, Portugal*
moreirasilva@clix.pt, fernandorua.sci@hgsa.min-saude.pt

**** CCTC, University of Minho, Braga, Portugal*
{jmac, abelha}@di.uminho.pt

Abstract

This work aims to support doctor's decision-making on predicting sepsis level and the best treatment for patients with microbiological problems. A set of Data Mining (DM) models was developed using forecasting techniques and classification models which will enable doctors' decisions about the appropriate therapy to apply, as well as the most successful one. The data used in DM models were collected at the Intensive Care Unit (ICU) of the Centro Hospitalar do Porto, in Oporto, Portugal. Classification models where considered to predict sepsis level and therapeutic plan for patients with sepsis in a supervised learning approach. Models were induced making use of the following algorithms: Decision Trees, Support Vector Machines and Naïve Bayes classifier. Confusion Matrix, including associated metrics, and Cross-validation were used for the evaluation. Analysis of the total error rate, sensitivity, specificity and accuracy were the associated metrics used to identify the most relevant measures to predict sepsis level and treatment plan under study. In conclusion, it was possible to predict with great accuracy the sepsis level (2^{nd} and 3^{rd}), but not the therapeutic plan. Although the good results attained for sepsis (accuracy: 100%), therapeutic plan does not present the same level of accuracy (best: 62.8%).

Key Words: Data Mining; Classification; Intensive Care; Sepsis; Predict Therapeutic Plans, INTCare

1. Introduction

Everyday new patients come into Intensive Care Units (ICU) in a critical health condition. One of the main existing problems in Intensive Medicine is related to therapeutics, more specifically when they should be administered to a patient. IT can have a very important role supporting quality and efficiency in health care, providing the right information, at the right

time, to the right person [1, 2].

It is very difficult for the professionals to care of the patients and, simultaneously, document the operations [3, 4]. In order to overcome this limitation a project was developed called INTCare [5-7]. This project has the objective to make available anywhere and anytime [8, 9] pertinent information about the patient. During the project a lot of data (vital signs, laboratory results, fluid balance, ventilation and ICU scores) were converted into electronic form, enabling the automatic acquisition in real-time. This new reality allows for obtaining fundamental knowledge to the patient treatment in the right time.

In order to reduce sepsis mortality, a set of procedures should be followed in an early stage. Survival medium probability decreases of 7.6% for each hour of delay in presence of an effective antibiotherapy. The elaboration of a therapeutic plan for sepsis may result not only in the reduction of mortality, but also in the substantial decrease of costs for institutions, due to the possible improvement in the usage of existing resources [12].

Consequently, the quick interpretation and precise evaluation of physiological data of intensive care patients' state are going to be crucial for a more efficient and effective decision making by the medical staff.

The objective of this project is to predict the patient sepsis level in real-time, determining whether the patient is in the second or third level of the scale [33]. In addition, making use of the same input variables, a set of models has been developed to predict the therapeutic.

This work was developed in a real environment using real data obtained from the ICU of Centro Hospitalar do Porto, in Portugal. This paper is divided in six chapters. After this introduction, the background and related work are presented in the second chapter. The third chapter makes an overview of the data used by the Data Mining Models. Then, the fourth chapter introduces the Data Mining models developed and variables used. The obtained results are presented in the chapter five. Finally, some conclusions about the work are written and the future work presented.

2. Background

2.1 Surviving Sepsis Campaign

A Surviving Sepsis Campaign (SSC) provides the international guidelines for the treatment of sepsis, severe sepsis and septic shock [13]. SSC is a program leaded by the ESICM (European Society of Intensive Care Medicine), ISF (International Sepsis Forum) and SCCM (Society of

Critical Care Medicine) that aims to improve the survival diagnostic and manage patients with sepsis, approaching the challenges associated with it, and having as a mission [14]:

- Increase the awareness, comprehension and knowledge;
- Alter perceptions and behaviors;
- Increase the rhythm of change in care standards;
- Define the care standards for severe sepsis;
- Reduce sepsis associated mortality by 25% in the next 5 years;
- Work with all stakeholders for an improvement in sepsis management, through specific initiatives.

a) SEPSIS

Sepsis is classified as a severe general infection and it is hard to define, diagnose and treat [3, 4]. It is related to a large number of clinical conditions caused by a systemic inflammatory response of the organism to an infection, which develops in severe sepsis, also combined with simple, multiple or total organ dysfunction/failure leading to death [3, 4]. It is one of the major causes of death in ICU. Daily is killing around 1,400 people worldwide [3, 4]. Although not having a clear clinical definition that can be easily adopted and communicated in its entirety, its absence turns sepsis diagnosis and treatment into a clinical challenge [3, 4]. Tables 1, 2 and 3 show the variables associated to sepsis levels. These variables are essential to the development of classification models.

Table 1. Definition of Sepsis, adapted from [13, 16, 17]

Variables	**Sepsis is defined as a documented or suspected infection with one or more of the following**
General	Fever (core temperature >38.3°C) Hypothermia (core temperature <36°C) Heart rate >90 min–1 or >2 SD above the normal value for age Tachypnea Altered mental status Significant edema or positive fluid balance (>20 mL/kg over 24 hrs) Hyperglycemia (plasma glucose >120 mg/dL) in the absence of diabetes
Inflammatory	Leukocytosis (WBC count >12,000 μL–1) Leukopenia (WBC count <4000 μL–1) Normal WBC count with >10% immature forms Plasma C-reactive protein >2 SD above the normal value Plasma procalcitonin >2 SD above the normal value

Table 2. Definition of Severe Sepsis, adapted from [13, 16, 17]

Variables	**Severe sepsis is defined as sepsis associated with organ dysfunction, hypoperfusion, or hypotension**

Organ dysfunction	Arterial hypoxemia (PaO2/FIO2 <300) Acute oliguria (urine output <0.5 mL·kg-1·hr-1 or 45 mmol/L for 2 hrs) Creatinine > 2.0 mg/dL Coagulation abnormalities (INR >1.5 or aPTT >60 secs) Thrombocytopenia (platelet count <100,000 ìL–1) Hyperbilirubinemia (plasma total bilirubin > 2.0 mg/dL or 35 mmol/L)
Tissue perfusion	Hyperlactatemia (>2 mmol/L)
Hemodynamic	Arterial hypotension (SBP <90 mm Hg, MAP <65 mm Hg, or SBP decrease >40 mm Hg)

Table 3. Definition of Septic Shock, adapted from [13, 16, 17]

Septic shock is defined as acute circulatory failure unexplained by other causes
Acute circulatory failure is defined as persistent arterial hypotension (SBP <90 mmHg, MAP <60, or a reduction in SBP >40 mm Hg from baseline despite adequate volume resuscitation).

b) Challenges and costs

Intensive care professionals (doctors and nurses) regard sepsis as one of the most challenging and hard to manage conditions, because its course varies from patient to patient, and can develop as a result of numerous circumstances [14]. Sepsis patient management involves a variety of therapeutic interventions, being the effectiveness of the treatment most likely if serious sepsis is avoided through appropriate and preventive care. Once diagnosed, the objective of the therapy is to eliminate the underlying infections with antibiotics [14]. Due to the challenges of diagnosing and treating the complex condition in which they are, about 10% of sepsis patients don't receive antibiotic treatment in adequate timing, raising mortality by 10-15% [15].

Relatively to costs, sepsis imposes a significant burden on health resources, being responsible for 40% of ICU total expenses, which amount to up to $7.6bn USD in Europe, annually, and $16.7bn USD in the US, in the year 2000; being the medium cost per case of approximately $22.000 USD [14].

c) PIRO

According to Levy et al. (2003), PIRO is a classification system proposed by John Marshall for stratifying the patients' state in regard of: (P) Predisposition, (I) Insult infection, (R) Response and (O) Organ dysfunction [16-19]. This concept was proposed by the second international consensus conference about sepsis definitions in 2001, which gathered simultaneously the most important medical societies: SCCM - Society of Critical Care Medicine; ESICM - European Society of Intensive Care Medicine; ACCP - American College of Chest Physicians; ATS - American Thoracic Society; SIS - Surgical Infection Society [16, 17]. The imprecision and heterogeneity of the population defined as sepsis patients led to the

introduction of this new sepsis classification system, known as PIRO [18]. The PIRO concept was proposed with the objective of improving sepsis detection [19]. Besides that, organizing these patients in more homogenous groups can help improve clinical practice, determine prognostics and include these patients in clinical studies [17]. This concept allows, according to [20], for: a more adequate stratification of patients according to seriousness groups; outline clinical studies to evaluate therapeutic strategies in patients with severe sepsis, and use this tool to analyze outcomes. PIRO is used for the daily classification of the patients' dysfunction/organic failure degree [21].

It is important to stress that the PIRO concept is rudimentary and appears as a research proposal and developing concept, needing further testing and perfecting, before being considered for routine application in clinical practice [16, 17].

Its elaboration demands extensive evaluation of sepsis' natural history to define the variables that predict not only an adverse outcome, but also the response potential to the therapy [16, 17].

d) Therapeutics plans

Regarding the therapeutic plan, quick interpretation and precise evaluation of physiological data in ICU patient state monitoring are crucial for a more efficient and effective decision making by the doctors.

2.2 Cross Industry Standard Process for Data Mining

a) CRISP-DM Phases

CRISP-DM (CRoss-Industry Standard Process for Data Mining) methodology was adopted to guide the work. The life cycle of CRISP-DM consists of six phases [32]:

- Business understanding - The initial phase was to understand the problem, with a focus on project objectives and requirements from the standpoint of business, after converting problem objectives into goals DM;
- Data understanding - At this phase, began the collection and subsequent use of data, with a view to understanding, analysis and troubleshooting of quality. Following the identification of relationships among data, or the detection of interesting subset thereof, in order to be subsequently analyzed so as to identify hidden knowledge;
- Data preparation - This phase involves all activities necessary to build the final data set. These data will be used by modeling tools for subsequent analysis by DM

algorithms. The tasks of data preparation include selection of tables, attributes and records, as well as transformation and cleaning of data, with a view to their subsequent analysis by modeling tools;

- Modeling - At this phase, we selected several modeling techniques and their parameters were adjusted to optimize the results. Normally, there are several techniques for the same type of DM goal, and some have specific requirements on how the data is presented. Sometimes we need to return to the stage of data preparation;
- Evaluation - This phase is aimed to evaluate the usefulness of the models. Within this work was adopted Confusion Matrix (CM): Accuracy, Sensitivity and sensibility and Cross-validation (CV) 10 folds.
- Deployment - The creation of models is not the end of the project. Even if the purpose of the models is to increase knowledge about the data, the information obtained must be organized and presented so that the user can use.

Due to the nature of the problem, this encompasses a typical DM objective of classification [5]. The main goal is to predict one target with two classes, i.e., classify whether a patient has or not sepsis.

b) Data Mining techniques

To overcome this problem a set of models were defined using three distinct techniques: Support Vector Machine (SVM), Decision Trees (DT) and Naïve Bayes (NB). The models were induced automatically using Oracle Data Mining [6, 7]. All the data used were obtained and processed automatically.

c) Evaluation

1) Cross-validation (CV) - In order to use all available cases and compare the prediction's precision, the cross-validation (CV) was used. The data set is randomly divided in mutually exclusive subsets of approximately equal *k* sizes (folds). The classification model is trained and tested *k* times (cycles). For each cycle, the model is trained using *k-1* folds and tested using the remaining data. The global estimated precision of a model is calculated by the average of the *k* individual precision measures, as shown in the following equation:

$$PCV = \frac{1}{k}\sum_{i=1}^{k} Ai$$

where PCV is the precision of the cross validation, *k* the number of used folds and *A* is the measure of precision (for example, rate of success, sensibility, specificity) of each fold [22]. The measure of precision used was the sensibility.

2) Confusion Matrix (CM) - Metrics for evaluating sepsis were: the total error rate, acuity, sensitivity. For therapy was considered the specificity. The Confusion Matrix of a classifier indicates the number of correct classifications versus forecasts made for each class on a set of examples [23]. The CM is a technique commonly used in the evaluation classification problems. In the binary case, each example is referred to as being positive or negative [24]. Some rates can be derivate from CM:
   - True Positive Rate (TPR) - corresponds to the number of positive examples correctly classified;
   - True Negative Rate (TNR) - corresponds to the number of negative examples actually classified as negative;
   - False Positive Rate (FPR) - corresponds to the number of positive examples classified as negative;
   - False Negative Rate (FNR) - corresponds to the number of negative examples classified as positive;

- From this matrix many other measures can be derived, such as [24]:
  - **Total error rate** = {FPR+FNR}/{n}x100(%);
  - **Sensibility** = {TPR}/{TPR+FNR}x100(%);
  - **Specificity** = {TNR}/{TNR+FPR}x100(%);
  - **Accuracy of prediction** = {TPR+TNR}/{n}x100(%).

2.3 Related Work

This work is related with the research project INTCare. The main goal of INTCare was the development of an Intelligent Decision Support System to predict organ failure and patient outcome in real-time and using online learning [8, 9]. To attain INTCare goals was necessary to perform a set of changes in the ICU environment [10] and in the way that the data were collected.

INTCare introduced changes in the data sources type and format. Now, the laboratory results are collected in an open format, the patient therapeutic are accessible electronically, the vital

signs are obtained automatically and a new Electronic Nursing Record (ENR) platform for collecting the nursing data in real-time called was developed. ENR is used for monitoring the patient data in real-time. ENR is a web-based touch screen platform available near the patient beds. This platform is used by nurses and by physicians to insert, validate or consulting the patient data. All the data are available anywhere and anytime. The new knowledge obtained by INTCare it is disseminated through ICU platforms / systems. The advances attained in the ICU technology enabled the prediction of the sepsis level and therapeutic plans.

3. Objectives and Methodologies

This research is focused on the prediction of sepsis and on the prediction of the therapeutic plan for patients with microbiological problems, based on sepsis levels.

3.2 Objectives

With the objective of supporting clinical decisions this work promoted the induction of DM models for predicting the best therapy for patients with microbiological problems, setting as a starting point the sepsis level. With that in mind, the following specific goals were considered:

- Understanding to what point is it possible to predict, with a high degree of accuracy, the sepsis level and therapeutic plan of sepsis patients;
- Study and define the variables that influence the therapy;
- Develop and test a set of classification models, based in DM, that will enable the doctor to decide about the best therapy to apply, adequate to the patient's problems;
- Test the classification models with real-world data.

3.1 Research Methodology

For some time Information Systems (IS) researchers were encouraged to consider Action Research (AR) methodology as an adequate research approach from the various adopted methodologies by IS [25]. According to Avison et al. (1999), AR has ideal characteristics for the study of IS with a significant impact in the area [26]. In this article, this methodology was partially adopted because it fits the development of the project and because it uses a systematic cyclical method of planning: action, observation, reflection and evaluation [27, 28]. With the goal of developing a set of decision models that will enable the doctor to decide on the best therapy to apply to the patient, there was the need to explore and survey facts. In an approach

to discover more about the nature, context, relevance and resolution of the problem, the implementation of a series of measures of action was started.

Firstly, data has been studied and prepared. Then, a theoretical model has been proposed and elaborated, allowing for responding to he presented problem. A survey of the necessary indicators was also made, for the model development. After that, the variables that influence the therapy were studied and defined. DM models for pattern prediction were developed. Finally, the models were tested with real-world data, collected in real-time and treated online.

4. Data Overview

4.1 Data description

The Data used to induce DM models were collected at the Intensive Care Unit (ICU) of Centro Hospitalar do Porto from 19-AUG-2011 to 04-JUL-2012, corresponding to 305 days, 394 patients and 12 beds. Different data types were collected from:

- 12 monitors of vital signs;
- 10 mechanical ventilators;
- Pharmaceutical data;
- Laboratory results.

The datasets provided by CHP for this project were generated from a set of real data collected and processed online in real-time, using the following data sources:

- Electronic Health Record (EHR);
- Vital Signs Monitor (VSM);
- Laboratory (LAB);
- Hospital Management Warehouse and Pharmacy (HMWP).

Table 4 presents the associated attributes, their description and respective data source, referring to the initial selection.

Table 4. Attributes, description and data source

Attributes	Description	Data source
BILIRUBIN	Bilirubin	LAB
CREATININE	Creatinine	LAB
HR	Heart Rate	VSM
GLUCOSE	Glucose	LAB
TIME	Time when the clinical examination was collected	LAB/VSM

LEUKOCYTES	Leukocytes	LAB
PID	Patient Identifier	EHR/VSM /LAB
DR	Date recording/collection value	EHR/VSM /LAB
MAP	Mean Arterial Pressure	VSM
SBP	Systolic Blood Pressure	VSM
PLATELETS	Platelets	LAB
TEMP	Temperature	VSM
MEDICAMENT	Name of medicaments	HMWP
GROUPMED_I	Group of medicaments (firs level)	HMWP
GROUPMED_A	Group of medicaments (secound level)	HMWP
SUBGROUPMED_I	Subgroup of medicaments (third level)	HMWP

4.2 Data preparation

In the data exploration, the statistical analysis of clinical tests and vital signs variables was made prior to the transformation. For a pre-validation of the values, as in table 5, the range of vital signs values, collected automatically, was determined by ICU doctors.

Table 5. Range of values of vital signs

Attribute	Min	Max
MAP	0	200
SBP	0	300
TEMP	34	45
HR	0	250

In table 6, the minimum and maximum values of vital signs and clinical analysis are shown, based on sepsis limits (tables 1, 2 and 3).

Table 6. Maximum and minimum values of vital signs and clinical analysis defined for sepsis

Attributes	Min	Max
BILIRUBIN	0	2
CREATININE	0	2
GLUCOSE	0	120
LEUKOCYTES	4000	12000
PLATELETS	100	9999
MAP	65	9999
SBP	90	9999
TEMP	36	38,3
HR	0	90

After the exploration of the data, its quality was checked, as in table 7, which shows the percentage of null values in all the selected attributes.

Table 7. Data quality

Attributes	Nulls
BILIRUBIN	0,0368%
CREATININE	0,0036%
HR	0
GLUCOSE	0,0006%
TIME	0
LEUKOCYTES	0,0034%

PID	0
DR	0
MAP	0
SBP	0
PLATELETS	0,0058%
TEMP	0
MEDICAMENT	0
GROUPMED_I	0
GROUPMED_A	0
SUBGROUPMED_I	0

Table 8 presents the data obtained after the transformation. Some new attributes and records were added. Excepting SEPSIS_FINAL variable, the value 0 (zero) has been assigned to variables whose value is within the range and the value 1 for those who are out of it.

Table 8. Data obtained after transformation

Attributes	Data
SEPSIS_BILIRUBIN	{0;1}
SEPSIS_CREATININE	{0;1}
SEPSIS_HR	{0;1}
SEPSIS_GLUCOSE	{0;1}
SEPSIS_LEUKOCYTES	{0;1}
SEPSIS_MAP	{0;1}
SEPSIS_SBP	{0;1}
SEPSIS_PLATELETS	{0;1}
SEPSIS_TEMP	{0;1}
SEPSIS_FINAL	{0;2;3}

Regarding SEPSIS_FINAL variable, value 0 represents patients without sepsis. Value 2 represents patients with severe sepsis and value 3 patients with septic shock. Using the value 1 of the remaining attributes present on table 8 and according to the definitions of sepsis on the tables 1, 2 and 3, a procedure was developed to calculate the sepsis level for each record. The procedure assigns the values of 0, 2, or 3 through a simple verification as defined by sepsis in the referenced tables (1, 2, and 3). For SEPSIS_FINAL was not considered the patient's condition with sepsis (referring to variables of table 1) as this would have to be rated with a value of 1. For this reason it was not possible to assign the value 1 to the variable SEPSIS_FINAL because, according to ICU doctors, these values are easily confused with patient's condition without sepsis. During the data integration process, in order to merge these data it was created a data view for the entire data, combining the attributes of table 4 and those transformed on table 8.

Finally, it was necessary to convert continuous numeric data into a range of classes. The ranges were created using a 7-point scale adapted from the Clinical Global Impression -

Severity scale (CGI-S) [3, 4]. The CGI-S allows the doctors to evaluate the disease severity [3, 4]. In this sense, Bin Quantile Range grouping technique was used considering seven classes [3, 4]. The classes were created using a quantile distribution of the values received by each variable.

5. Data Mining Models

Since the variables correspond to continuous and discrete values, it was opted to use the classification models, distinguishing the independent variables set (analysis and vital signs) from the dependent variable set (final sepsis and medication). The used modeling techniques, as referred before, are DT, SVM and NB, whose characteristics are described in table 9.

Table 9. Characteristics of modeling techniques

Description	Values
Algorithm Name	Support Vector Machine
Active Learning	yes
Automatic Preparation	on
Complexity Factor	0.068966
Kernel Function	Linear
Tolerance	.001
Algorithm Name	Naive Bayes
Automatic Preparation	on
Pairwise Threshold	0
Singleton Threshold	0
Algorithm Name	Decision Tree
Automatic Preparation	on
Criteria For Splits	20
Criteria For Splits(%)	.1
Maximum tree depth	7
Minimum Child Record Count	10
Tree Impurity Metric	Gini

In order to obtain models from the data to classify sepsis level, it was essential to execute DM tasks to obtain new knowledge. After the data preparation and data processing tasks, a modeling phase has been carried out.

Figure 1 shows the data transformations of the classification models for sepsis and the figure 2 shows the data transformations of therapeutic plan classification models.

Fig. 1. DM process to the Sepsis target

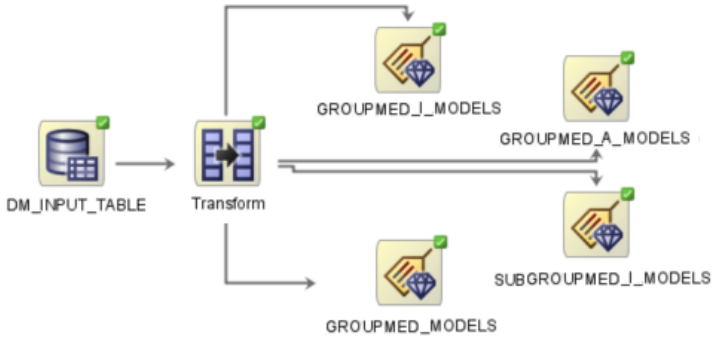


Fig. 2. DM process to the therapeutic plan targets

Once variables correspond to continuous and discrete values, the classification models were used distinguishing the independent variables (lab results and vital signs) and the dependent variables (the score for sepsis and medicaments group). For modeling/evaluation 70% of the data was used for training and the remaining 30% for testing (Holdout sampling). During the modeling task some parameters adjustments were made. All continuous and discrete values with more than 10 classes, excepting the episode number and date, were classified using the Quantile Bin Range.

To generate the models, the process went through two stages:

- The 1st stage is constituted by models M1, M2, M3 and M4, for sepsis level and therapeutic plan, in which attributes M1, M2 and M3 are selected manually and M4 automatically;
- The 2nd stage is constituted by models M5, M6, M7, M8, M9 and M19, for the therapeutic plan, in which all attributes are selected manually.

Each model was developed using three different techniques (SVM, DT and NB) related to sepsis level condition and to the therapeutic plan. The input attributes (independent variables) are:

CaseMix = {PID; Date}

Sepsis = {Bilirubin; Creatinine; HRMax; HRMin; Glucose; Leukocytes; MAPMax; MAPMin; SBPMax; SBPMin; Platelets; TempMax; TempMin}

VarSepsis = {SepsisBilirubin; SepsisCreatinine; SepsisHR; SepsisTemp; SepsisGlucose; SepsisLeukocytes; SepsisMAP; SepsisSBP; SepsisPlatelets}

The target attribute (dependent variables) is:

Target1 = SepsisFinal

Target2 = Medicament / Drug

Each model can be represented as follows:

Model{M1}*{SVM;DT;NB}*{target1;target2} = {CaseMix; Sepsis}

Model{M2}*{SVM;DT;NB}*{target1;target2} = {CaseMix; VarSepsis}

Model{M3}*{SVM;DT;NB}*{target1;target2} = {CaseMix; Sepsis; VarSepsis}

Model{M4}*{SVM;DT;NB}*{target1;target2} = {Automatic}

For example, the independent variables related to M1 (casemix and sepsis), can be represented as follows:

SEPSIS_SVM_M1_SVM = SVM * {target1} * {CaseMix; Sepsis}

SEPSIS_DT_M1 = DT * {target1} * {CaseMix; Sepsis}

SEPSIS_NB_M1 = NB * {target1} * {CaseMix; Sepsis}

MED_SVM_M1 = SVM * {target2} * {CaseMix; Sepsis}

MED_DT_M1 = DT * {target2} * {CaseMix; Sepsis}

MED_NB_M1= NB * {target2} * {CaseMix; Sepsis}

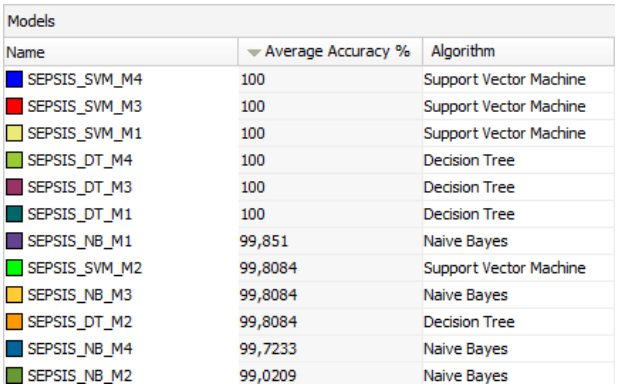
It was generated and tested a total of:

- 12 classification models (1 (target) * 4 (scenarios) * 3 (DM techniques)) to predict sepsis level;
- 64 classification models (3 (target) * 6 (scenarios) * 3 (DM techniques)) to predict therapeutic plan.

In order to use all available cases and compare the prediction's precision, the CV method was used to estimate the synthesis capacity of the classification models. As the ODM implements cross validation in its techniques [29], it was possible to test the model's precision over the used data. Once the results can depend on the random division of the mutually exclusive sub-sets, 10 executions were applied to each 10-fold sub-set, in a total of 10*10 = 100 results for each test configuration. For evaluation, in order to compare the sepsis level classification models, a ROC curve analysis [30] was made. The ROC (Receiver Operating Characteristic) curves are frequently used in the medical field to evaluate decision support computational models, diagnosis and prognosis [31].

6. Results

In order to assess the results attained by the models developed for sepsis a set of measures were applied. Figure 3 shows, ordered by average accuracy, the values of accuracy for each classification model and the technique used. All of the M4 models for each technique use variables defined automatically by the engine. The other models were characterized using a manual selection of the variables.



Models

Name	Average Accuracy %	Algorithm
SEPSIS_SVM_M4	100	Support Vector Machine
SEPSIS_SVM_M3	100	Support Vector Machine
SEPSIS_SVM_M1	100	Support Vector Machine
SEPSIS_DT_M4	100	Decision Tree
SEPSIS_DT_M3	100	Decision Tree
SEPSIS_DT_M1	100	Decision Tree
SEPSIS_NB_M1	99,851	Naive Bayes
SEPSIS_SVM_M2	99,8084	Support Vector Machine
SEPSIS_NB_M3	99,8084	Naive Bayes
SEPSIS_DT_M2	99,8084	Decision Tree
SEPSIS_NB_M4	99,7233	Naive Bayes
SEPSIS_NB_M2	99,0209	Naive Bayes

Fig. 3. All the models developed for sepsis

From the twelve scenarios considered relatively to the sepsis level, the top three non-automatic models (one of each technique) were selected for analysis: SEPSIS_SVM_M3, SEPSIS_DT_M3 SEPSIS_NB_M1 (Figure 3). Tables 10, 11 and 12 show the best prediction results for each set of variables selected, considering the technique and the corresponding scenario and the values for total error, sensitivity, specificity and accuracy.

Table 10. Confusion matrix model SEPSIS_SVM_M3

Model SEPSIS_SVM_M3	**Severe sepsis**	**Septic shock**	**Total**	**Corrects**
Severe sepsis	334	0	334	100%
Septic shock	0	1.415	1.415	100%
Total	334	1.415	**1.749**	
Corrects	100%	100%		

Total error	**Sensitivity**	**Specificity**	**Accuracy**
0%	100%	100%	100%

Table 11. Confusion matrix model SEPSIS_DT_M3

Model SEPSIS_DT_M3	**Severe sepsis**	**Septic shock**	**Total**	**Corrects**
Severe sepsis	334	0	334	100%
Septic shock	0	1.415	1.415	100%
Total	334	1.415	**1.749**	
Corrects	100%	100%		

Total error	**Sensitivity**	**Specificity**	**Accuracy**
0%	100%	100%	100%

Table 12. Confusion matrix model SEPSIS_NB_M1

Model SEPSIS_NB_M1	**Severe sepsis**	**Septic shock**	**Total**	**Corrects**
Severe sepsis	109	0	109	100%
Septic shock	1	436	437	99,77%
Total	110	436	**546**	
Corrects	99,09%	100%		

Total error	**Sensitivity**	**Specificity**	**Accuracy**
0,18%	100%	99,09%	99,82%

Confusion matrices permit conclude that the best predictions for sepsis level are provided by SEPSIS_SVM_M3, SEPSIS_DT_M3 and SEPSIS_NB_M1 as presented on tables 10, 11 and 12 respectively.

From the results obtained for sepsis levels the Figure 4 shows the ROC curve obtained from SEPSIS_DT_M3 and SEPSIS_SVM_M3 models, considered the best models in terms of accuracy (100%). In figure 5 it is represented the decision tree obtained from the SEPSIS_DT_M3 model.

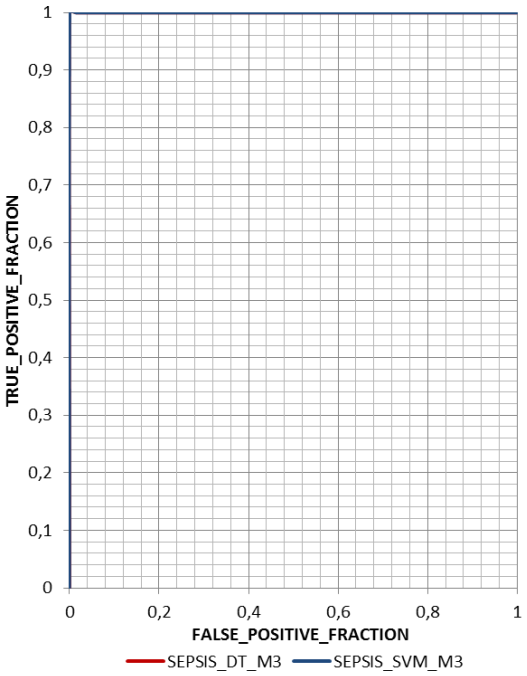


Fig. 4. All the models developed for sepsis

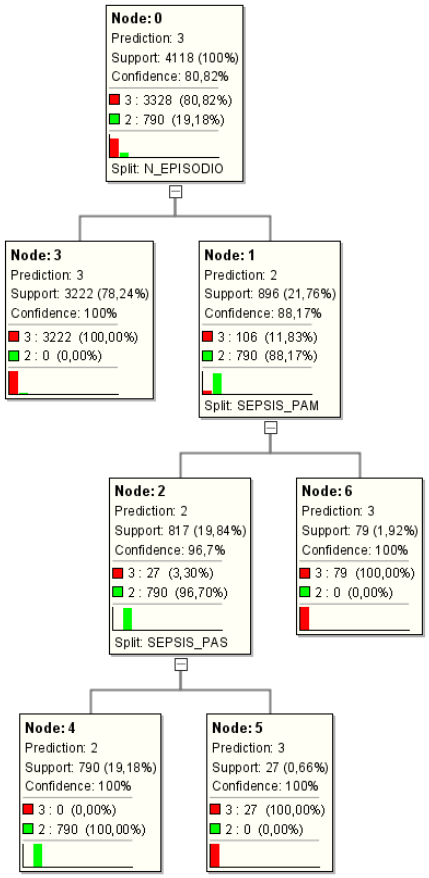


Fig. 5. Decision tree for SEPSIS_DT_M3

In conclusion, it is possible to predict sepsis level with a great accuracy. Regarding sepsis level, the accuracy, sensitivity and specificity results were 100% for SEPSIS_DT_M3 and SEPSIS_SVM_M3 models. SEPSIS_NB_M1 model presented 100% for sensitivity, 99.99% for specificity and 99.82% for accuracy.

In terms of sepsis' therapeutic plan, an analysis looking for the best model per target was carried out. Table 13 summarizes the results obtained, considering only the top eight models from sixty four scenarios tested. In the superior division are presented four models which were developed through manual selection of attributes. In the inferior division are presented four models based on automatic selection of attributes.

Table 13. Best model obtained for target related to therapy

Targets	Classes	Models	Acuity
GROUPMED_I (1st level)	13	MED_DT_M8	48,90%
GROUPMED_A (2nd level)	49	MED_NB_M8	45,30%
SUBGROUPMED_I (3th level)	37	MED_DT_M6	45,19%
MEDICAMENT	75	MED_DT_M8	34,11%
GROUPMED_I (1st level)	13	MED_DT_M4	62,84%
GROUPMED_A (2nd level)	49	MED_SVM_M4	44,00%
SUBGROUPMED_I (3th level)	37	MED_SVM_M4	55,23%
MEDICAMENT	75	MED_SVM_M4	36,24%

7. Discussion

The obtained results are the culmination of an entire process of acquisition and transformation of data, as well as model induction. In the data analysis, a total of 7 202 272 records were treated, from which there were several with null or out of range values. After the data selection, a treatment and transformation process was executed. A final set of data containing a total of 193 122 records was obtained to feed DM models. In regard of the sepsis levels prediction, very good results were attained with the classification models, which represent 100% of accuracy. The same cannot be said regarding the therapeutic plan prediction. The goal was achieved only partially. Results demonstrate that the correlation between sepsis level and therapeutic plan is weak. All data used in this work is from real world, acquired in real-time. This means that the resulted models can be integrated in a decision support system to aid doctors in their decision making processes.

8. Conclusions

Results attained in this work prove that it is possible to predict the Sepsis level in real-time using Data Mining.

This paper presented the classification models induced by using data collected in real time from the ICU of CHP, Porto, Portugal. It was considered a large initial data set which resulted, after processing, in a total of 193 112 records, relating to 394 episodes that occurred during 305 days.

DM techniques were used to extract knowledge. SVM, DTs and NB algorithms were applied in order to search patterns (SEPSIS Levels) and subsequent discovery of useful information.

Results obtained have high acuity. In regard to therapeutics, the acuity results weren't satisfactory but, in spite of that, good levels of assertiveness were verified in some drugs/medication groups.

The assessment of sepsis level is a crucial task for intensive care environments. So as soon as

the risk is identified, more quickly it is applied the best and more accurate treatment. The development of classification models for sepsis is associated to some benefits, such as mortality decreasing and substantial costs reduction for institutions. Besides diagnosing the correct sepsis level, predicting the correct treatment also avoids medication testing costs.

The development of classification models for sepsis can be seen as a major contribute for developing a decision support system. Some experiences were done in order to understand some therapeutics tendencies, however the results weren't satisfactory. The results suggest that sepsis level and patient therapeutic plan aren't related.

9. Future Work

Further work includes:

- To determine new variables that may be used to predict therapeutic;
- To build new models about therapeutic due to the great correlation between the input variables (sepsis level) and target (therapeutic);
- To implement a decision support system based on the models of sepsis developed in this project.

10. Acknowledgments

This work is supported by FEDER through Operational Program for Competitiveness Factors – COMPETE and by national funds though FCT – Fundação para a Ciência e Tecnologia in the scope of the project: FCOMP-01-0124-FEDER-022674. The authors would like to thank FCT (Foundation of Science and Technology, Portugal) for the financial support through the contract PTDC/EIA/72819/ 2006 (INTCare) and PTDC/EEI-SII/1302/2012 (INTCare II). The work of Filipe Portela was supported by the grant SFRH/BD/70156/2010 from FCT.

References

[1] Handel, D. A. and Hackman, J. L. (2010). Implementing electronic health records in the emergency department. *The Journal of Emergency Medicine*, 38(2):257–263.

[2] Nowinski, C. J., Becker, S. M., Reynolds, K. S., Beaumont, J. L., Caprini, C. A., Hahn, E. A., Peres, A., and Arnold, B. (2007). The impact of converting to an electronic health record on organizational culture and quality improvement. *International Journal of Medical Informatics*, 76:S174–S183.

[3] R. L. Mador and N. T. Shaw, "The impact of a Critical Care Information System (CCIS) on time spent charting and in direct patient care by staff in the

ICU: a review of the literature," *International Journal of Medical Informatics*, vol. 78, pp. 435-445, 2009.

[4] K. Häyrinen, K. Saranto, and P. Nykänen, "Definition, structure, content, use and impacts of electronic health records: A review of the research literature," *International Journal of Medical Informatics*, vol. 77, pp. 291-304, 2008.

[5] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression," *Wadsworth, Belmont, CA*, 1984.

[6] O. D. M. Concepts, "11g Release 1 (11.1)," *Oracle Corp*, vol. 2007, 2005.

[7] P. Tamayo, C. Berger, M. Campos, J. Yarmus, B. Milenova, A. Mozes, *et al.*, "Oracle Data Mining," *Data Mining and Knowledge Discovery Handbook*, pp. 1315-1329, 2005.

[8] F. P. Filipe Portela, Manuel Filipe Santos, "Data Mining Predictive Models For Pervasive Intelligent Decision Support In Intensive Care Medicine," presented at the KMIS 2012 - nternational Conference on Knowledge Management and Information Sharing, Barcelona, 2012.

[9] F. Portela, P. Gago, M. F. Santos, A. Silva, F. Rua, J. Machado, *et al.*, "Knowledge Discovery for Pervasive and Real-Time Intelligent Decision Support in Intensive Care Medicine," presented at the *KMIS* 2011- International Conference on Knowledge Management and Information Sharing., Paris , France, 2011.

[10] F. Portela, M. F. Santos, Á. Silva, J. Machado, and A. Abelha, "Enabling a Pervasive approach for Intelligent Decision Support in Critical Health Care," presented at the *HCist 2011* – International Workshop on Health and Social Care Information Systems and Technologies, Algarve, Portugal, 2011.

[11] Kumar, A., Roberts, D., Wood, K. E., Light, B., Parrillo, J. E., Sharma, S., Suppes, R., Feinstein, D., Zanotti, S., Taiberg, L., et al. (2006). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical Care Medicine*, 34(6):1589.

[12] Shorr, A. F., Micek, S. T., Jackson Jr, W. L., and Kollef, M. H. (2007). Economic implications of an evidence-based sepsis protocol: Can we improve outcomes and lower costs?*. *Critical Care Medicine*, 35(5):1257.

[13] Dellinger, R. P., Levy, M. M., Carlet, J. M., Bion, J., Parker, M. M., Jaeschke, R., Reinhart, K., Angus, D. C., Brun-Buisson, C., Beale, R., et al. (2008). Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2008. *Intensive Care Medicine*, 34(1):17–60.

[14] SSC (2010). Surviving sepsis campaign. http://www.survivingsepsis.org/Introduction/Pages/default.aspx. [Website accessed on November 24, 2011].

[15] Lyseng-Williamson, K. A. and Perry, C. M. (2002). Drotrecogin alfa (activated). *Drugs*, 62(4):617–630.

[16] Levy, M., Fink, M., Marshall, J., Abraham, E., Angus, D., Cook, D., Cohen, J., Opal, S., Vincent, J., and Ramsay, G. (2003a). 2001 sccm/esicm/accp/ats/sis international sepsis definitions conference. *Critical Care Medicine*, 31(4):1250–1256. [Special Articles].

[17] Levy, M., Fink, M., Marshall, J., Abraham, E., Angus, D., Cook, D., Cohen, J., Opal, S., Vincent, J., and Ramsay, G. (2003b). 2001 sccm/esicm/accp/ats/sis international sepsis definitions conference. *Intensive Care Medicine*, 29(4):530–538.

[18] Opal, S. (2005). Concept of piro as a new conceptual framework to understand

sepsis. *Pediatric Critical Care Medicine*, 6(3):S55.

[19] Rabello, L., Rosolem, M., Leal, J., Soares, M., Lisboa, T., and Salluh, J. (2009). Entendendo o conceito piro: da teoria à prática clínica - parte 1. *Revista Brasileira Terapia Intensiva*, 21(4):425–431.

[20] Rosolem, M., Rabello, L., Leal, J., Soares, M., Lisboa, T., and Salluh, J. (2010). Entendendo o conceito piro: da teoria à prática clínica: parte 2. *Revista Brasileira Terapia Intensiva*, 22(1):64–8.

[21] Vincent, J., Moreno, R., Takala, J., Willatts, S., De Mendonca, A., Bruining, H., Reinhart, C., Suter, P., and Thijs, L. (1996). The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7):707–710.

[22] Turban, E., Sharda, R., Delen, D., and King, D. (2011). *Business Intelligence: a Managerial Approach*. Pearson Prentice Hall. 2nd edition.

[23] Kohavi, R. and Provost, F. (1998). *Glossary of terms. Machine Learning*, 30(June):271–274.

[24] Santos, M. F. and Azevedo, C. S. (2005). *Data Mining: Descoberta de Conhecimento em Bases de Dados*. FCA-Editora de Informática.

[25] West, D., Stowell, F., and Stansfield, M. (1985). Action research and information systems research. In Ellis, K., Gregory, A., Mears-Young, B., and Ragsdell, G., editors, *Critical Issues in Systems Theory and Practice*.

[26] Avison, D., Lau, F., Myers, M., and Nielsen, P. (1999). Action research. *Communications of the ACM*, 42(1):94–97.

[27] O'brien, R. (1998). An overview of the methodological approach of action research. *Unpublished paper to Professor Joan Cherry, Course LIS3005Y*, Faculty of Information Studies, University of Toronto. April, 17.

[28] McNiff, J. (2002). Action research for professional development. *Concise Advice for New Action Researchers*.

[29] Oracle (2012). Data mining with oracle database 11g release 2 competing on in-database analytics. http://www.oracle.com/us/products/database/options/ advancedanalytics/ 039550.pdf?ssSourceSiteId=ocomkr. [Website accessed on September 20, 2012].

[30] Bi, J. and Bennett, K. P. (2003). Regression error characteristic curves. In *ICML*, pages 43–50.

[31] Lasko, T., Bhagwat, J., Zou, K., Ohno-Machado, L., et al. (2005). The use of receiver operating characteristic curves in biomedical informatics. *Journal of biomedical informatics*, 38(5):404–415.

[32] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). Crisp-dm 1.0. *CRISP-DM Consortium*.

[33] Gonçalves, J. M., Portela, F., Santos, M. F., Silva, Á., Machado, J., & Abelha, A. (2013). Predict Sepsis Level in Intensive Medicine–Data Mining Approach. In *Advances in Information Systems and Technologies* (pp. 201-211). Springer Berlin Heidelberg.

*Corresponding author: Filipe Portela, MsC

Algoritmi Centre, Department of Information System,

University of Minho,

Campus Azurém, Guimarães, Portugal

E-mail: cfp@dsi.uminho.pt