Analysis and Determination of Risk Factors Leading to Preterm Birth Using Data Mining Techniques in R

Adriana-Georgiana MALEA, Ştefan HOLBAN, Nicolae MELIȚĂ
Politehnica University of Timişoara
Blv. Vasile Pârvan, No. 2,
RO-300223 Timişoara

***Abstract* – This paper aims to present a method of analysis a real medical dataset using Data Mining techniques. The method determines the risk factors leading to preterm birth and also analysis the quality in obstetrics. As language and development environment for statistical computation we used the R system. The purpose of this study is to highlight the influence of certain factors for premature birth by using graphics created in R.**

The method further, applied on a sample of Romanian births, shows that various socio-demographic, anthropometric, behavioral and medical factors act interconnected in a direct manner on the risk of premature births. To prove this statement we wanted a partitioning based on similarities between the attributes that may be risk factors. We also classified attributes that may represent risk factors according to the type of birth: term or preterm. The results will be outlined with graphics obtained by applying data exploration algorithms.

***Index Terms* – Clustering, Data Mining, Heatmap, Preterm birth, R programming language**

I. INTRODUCTION

Studies on the risk of preterm birth are important due to the relationship that exists between shortened gestation (the number of weeks from conception to birth) and infant morbidity, mortality, and subsequent growth and development [1]. Births before term, earlier than 37 weeks, have a great social impact, being with congenital malformations, the most important source of morbidity and neonatal mortality, with late debilitating consequences. Hence, the present study comes as help for the health professionals and parents, showing that factors such as infections, tobacco, alcohol, previous births, and maternal age too low or too advanced may increase the risk for premature births. It is therefore recommended to consider all the factors that may interfere with a normal pregnancy and also analyze them to minimize the risk. While several studies have calculated the relative risks of a large array of extrinsic and intrinsic factors for preterm births, relatively few have simultaneously examined these factors using Data mining algorithms for clustering and classification. This is unfortunate as such analyses allow for the estimation of the relative contribution of medical, anthropometric, and socio-demographic factors associated with premature births.

The study presented in this paper is based on a research made in both computer and medical field, which refers to the mechanisms of recognizing factors that lead to premature births. As a consequence, we proposed a method that shows the influence of certain factors on preterm births by applying Data Mining algorithms and by using the packages provided by the R environment. We used clustering and classification techniques on a dataset of 546 patients from the Banat region, who gave birth on term or before term in a medical unit profile. Each patient is characterized by 48 features. The interesting topic of this research is that the R language enables the use of data structures like matrices, known as *data frames*[6]. They allow multivariate data processing, in order to achieve connections between patients represented as observations and features represented as variables.

Our research has three phases: pre-processing, clustering and classification. In the pre-processing phase we prepared data and analyzed the influence of the 48 available features, showing that some of them may represent risk factors in case of preterm births. Those features are socio-demographic parameters like: age, gender, education, income, occupation, religion, ethnicity, and environment, anthropometric parameters: height, weight, body mass index, behavioral parameters: tobacco, drugs, alcohol, hours of work per week and medical parameters: menstrual cycle, gastrointestinal disease, liver disease, allergies, gestagenics, infections, previous births and others. In the clustering phase, we used cluster analysis to determine classes with similarities between values of observations regarding the type of birth. We represented the dataset variables, in order to partition patients in two sets: term or preterm birth. Finally, the classification phase shows the data obtained after classifying and misclassifying observations regarding the type of birth. Results are highlighted through graphical facilities provided in R, such as: dendograms, histograms, charts and heatmap site. Through these graphics we represented the dataset diversity and so we could make an idea about risk factors that lead to preterm births. Parts of the R code are included.

II. MATERIALS AND METHODS

Data for the study were gathered from 546 medical records at a hospital from Timisoara. Premature infants were defined as those born before the 37th week of gestation. Among those 48 variables that characterize the dataset there is one that shows the type of birth: on term or before term. This parameter represents a key aspect in our research,

because depending on it we will cluster and classify. R packages allowed us to apply different algorithms and to decide which ones fit best our data.

One method of analysis that allows for the assessment of the contribution of numerous determinants of preterm birth and provides insight into the relationship among them contains three steps: data pre-processing, cluster analysis and classification. The strength of this method lies in the fact that it can more accurately reflect connections between risk factors by graphically illustrating them.

PRE-PROCESSING

Data pre-processing consists of two phases: data preparation and data analysis. Data preparation is the initial phase and involves selecting data of interest to discover knowledge. Data analysis is the phase where we apply different techniques of mathematical statistics and artificial intelligence depending on the nature of data to be processed (categorical, nominal, mixed).

The preparation phase is aimed to visualize medical data content and process it so that information becomes homogeneous and uniform. Consequently, we noted that there are many missing values, values that will be filled in later, but we found that there are columns that have the same value for all observations, that means columns with unique factors. Therefore it was decided to eliminate those columns because they adversely affect the analysis. The reason is the impossibility of determining the influence of these attributes on premature births, because they contain only one type of information. For example, for attribute named *Alcohol* we identified only one existing value: *no*. As most values around 90% were *no*, it was assumed that the rest had the same value and therefore we resorted to deleting the attribute. For the other two attributes representing medical data there were approximately 95% missing values which can lead to errors in processing the information. Attributes VDRL and HIV both contain negative values for most fields and for the unknown we assume that it is negative too. Column headed 16-27SG has only values of 1 and the remaining are missing values, so we removed it. Other four columns contained over 95% missing values and we decided to remove those columns because they can not be relevant afterwards. At the end of this phase, 11 attributes with unique factors were eliminated because the nature of their values could not distinguish between a term or preterm birth.

In the data analysis phase, we read the information in a *data frame* and we mark unknown values as *Na* (*Not a Value*), because values containing *Na* can be processed easier with specific functions, i.e. *is.na.* Furthermore, we set the class of each variable through *col.classes* parameter and we name each column with *col.names* because we want to refer each attribute by its name, not by using a numeric index. We identify that there are 14.58% numerical attributes and the rest are nominal.

```
>nasteri<-read.csv(file="test_corectat.csv
",sep=",",header=FALSE,dec=".",na.strings="?"
,
colClasses=c('numeric','factor','factor','fac
tor','factor','factor','factor','numeric','nu
meric','numeric','factor','numeric','numeric'
,'factor','factor','factor','factor','factor'
,'numeric','factor','factor','factor','factor
','factor','factor','factor','factor','factor
','factor','factor','factor','factor','factor
','factor','factor','factor','factor'),col.na
mes=c('Varsta','Localitate','Ocupatie','Origi
ne_etnica','Ciclu_menstrual','Grupa_sanguina'
,'Rhesus','Greutate','Inaltime','Indice_masa_
corporala','Tigari','Cate_tigari','Nr_consult
uri_sarcina','Religie','Etnie','Mediu','Nivel
_instruire','Categorie_asigurat','Ore_de_munc
a_pe_saptamana','Tip_program','Alergii','Cons
um_lichide','Relatia_cu_tatal','Mai_mare_37sa
pt','Mai_putin_de_15SG','Mai_putin_15sapt','T
ipul_ultimei_nasteri','Gravida','Para','Infec
tii','Medicatie','Antibiotice_Ovule','Gestage
ne','Antispastice','Preparate_Fier','Vitamine
','US_Gestation_Weeks'));
>nasteri      <-      nasteri[apply(nasteri,
1,function(x) !sum(is.na(x))>25), ]
```

Code 1. Reading the dataset.

Using the *summary()* function from R, we make a statistical analysis on our knowledgebase and we review those attributes which express better patients' diversity. We want to show that diversity imposes the method chosen, as the only solution that can satisfy and offer good results. Hence, we analyze numerical attributes like: height, weight, body mass index, number of consults per pregnancy and number of cigars per day through histograms and boxplots, showing the type of distribution for each attribute.We also use boxplots for nominal attributes like: occupation, menstrual cycle, ethnicity, sanguine group, education, religion, environment, medication. Moreover, we change the parameters of every function we use from the *graphics package*, so that we obtain a proper representation. We end this pre-processing phase by gathering the most representative histograms and boxplots in order to proceed the next steps.

CLUSTERING

Cluster analysis involves associating with each object a set of G measurements which form the feature vector, X = (X1, . . . ,XG). The feature vector X belongs to a feature space X (e.g., R^G). The task is to identify groups, or clusters, of similar objects on the basis of a set of feature vectors, X1 = x1, . . . ,Xn = xn. Clustering procedures fall into two broad categories: hierarchical methods, either divisive or agglomerative and partitioning methods. [2] In our research we use both categories, because the aim of clustering is to partition the attributes from our data set in two clusters corresponding to term and preterm births, without considering the attribute *US Gestation Weeks*.

Still, if clustering would be made only through this attribute, the two clusters would have 188 patients that gave birth on term and 185 patients that gave birth before term and would look like figure 1. We observe that the two clusters are mutually exclusive and diametrically opposed in

space represented by the set of patients.

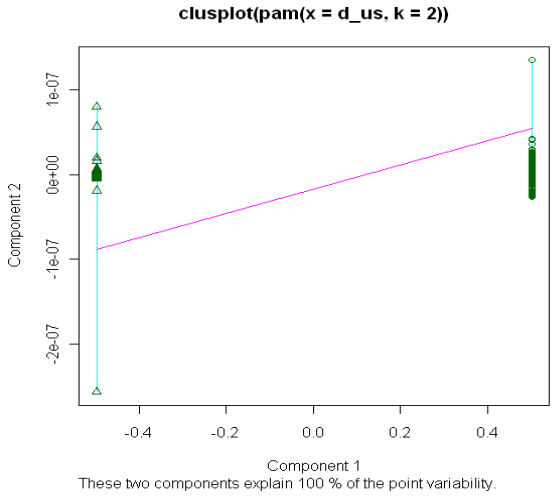

Fig.ure 1. The two clusters: on term and before term births are diametrically opposite.

```
>plot_us2<-
pam(as.numeric(Nasteri$US_Gestation
_Weeks),2)
```

Code 2. PAM Algorithm applied on the dataset.

To apply clustering algorithms it is required to calculate the distance matrix between dataset items. Such data are processed in a matrix of similarities, the matrix D = (dij), for all n objects to be clustered. Once the distance between two objects is chosen, we must define a measure of distance between clusters or groups of observations. We have to determine the distance matrix of each type of data: nominal, binary, numeric, ordinal. To calculate the distance matrices for each of these variables we used a tutorial written by Kardi Teknomo about determining the distance matrices for multivariate data [3].

First, we apply *Hierarchical clustering (hclust) algorithm* and we use dendograms to view nested sequences of clusters. Afterwards, we show that through AGglomerative NESting (agnes) and DIvisive ANAlysis (diana) algorithms we obtain the main structure of the data meaning the shape of the clusters (agnes) and focusing on upper levels of dendogram (diana). Partitioning methods, like Partitioning Around Medoids (pam), K-means (kmeans), Fuzzy C-means clustering (cmeans), Fuzzy Analysis (fanny) and Clustering LARe Applications (clara) partition the data into a number K of mutually exclusive and exhaustive groups, but we show through silhouette and variability of attributes clustered that similarity between attributes is quite low and so groups are not mutually exclusive. Consequently, a classifying phase is required.

CLASSIFYING

Classifying means that classes are predefined and the task is to understand the basis for the classification from a set of labelled objects (training or learning set). This information is then used to classify future observations. [4] Classification algorithms used in this paper are algorithms from MLInterfaces package from the Bioconductor project. The results of the classification algorithms can be observed through the confusion matrix and classifier accuracy. Given a data set we can find if the rules generalize satisfactory the accuracy of predictions. Accuracy is expressed in terms of the difference between predicted scores and actual results of the measurements. On the other hand it may be a measure of error rate, i.e. percentage of misclassified records. The confusion matrix compares the known classification of the testing set with the predicted classification based on the tuned machine learning algorithm. [5] Besides classic algorithms from MLInterfaces package, we use cross-validation to assess the prediction error of supervised machine learning. In order to get an accurate assessment it is important that all steps that can affect the outcome are included in the cross-validation process. [5]

In this phase we apply Naïve Bayes algorithm on a set of 250 patients from the initial dataset, with and without cross-validation. Afterwards, we use Support Vector Machines (SVM) with different kernels and Nearest Neighbor Method (knn) algorithm to obtain maximum accuracy for predection of term and preterm birth.

```
clasif      =      MLearn(class~.,data=TMP2,
naiveBayesI,1:250)
>    clasif3    =    MLearn(class~.,data=TMP2,
naiveBayesI,       xvalSpec("LOG",       10,
balKfold.xvspec(10)))
```

Code 3. NaiveBayes Algorithm without and with cross-validation applied on the test set.

III. RESULTS AND DISCUSSION

Following the study, we found that premature birth is largely influenced by primary predictive factors that represent attributes in our medical dataset. They characterize patients that have been grouped and then classified using Data Mining techniques in the programming language R. We started from the premise that the dataset is correct in terms of health, but we also considered the possibility of errors here and there.

In the pre-processing phase we show that 3 attributes are the most relevant for preterm births. The attribute that reflects the *number of cigars per day* shows that 35% of total number of patients smokes on average 10 cigars per day and they present a higher risk of premature births than those who smoke less than the average. This fact is illustrated in figure 2.

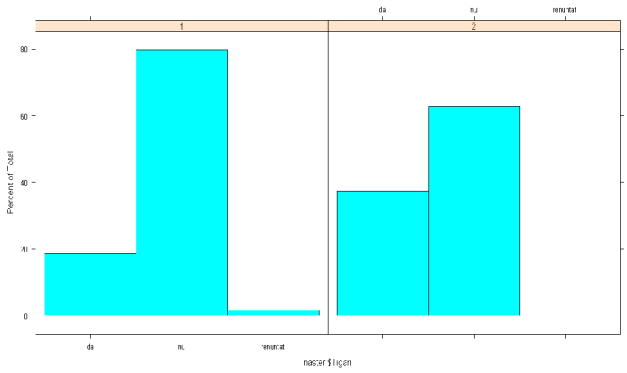

Figure 2. Type of birth in relation with tobacco consume.

In figure 2, the 2 boxes represent patients which gave birth on term, respectively before term and first bar of each

box corresponds to smokers, the second to non-smokers, the third to smokers who quit.

The attribute *Age* shows that patients aged over 35 years old and those between 18 and 25 present a high risk of preterm birth. As there are only few patients under 18 years old, we can not conclude anything. In the figure 3 below, first bar from each box represents patients who gave birth on term and second bar represents patients who gave birth before term.

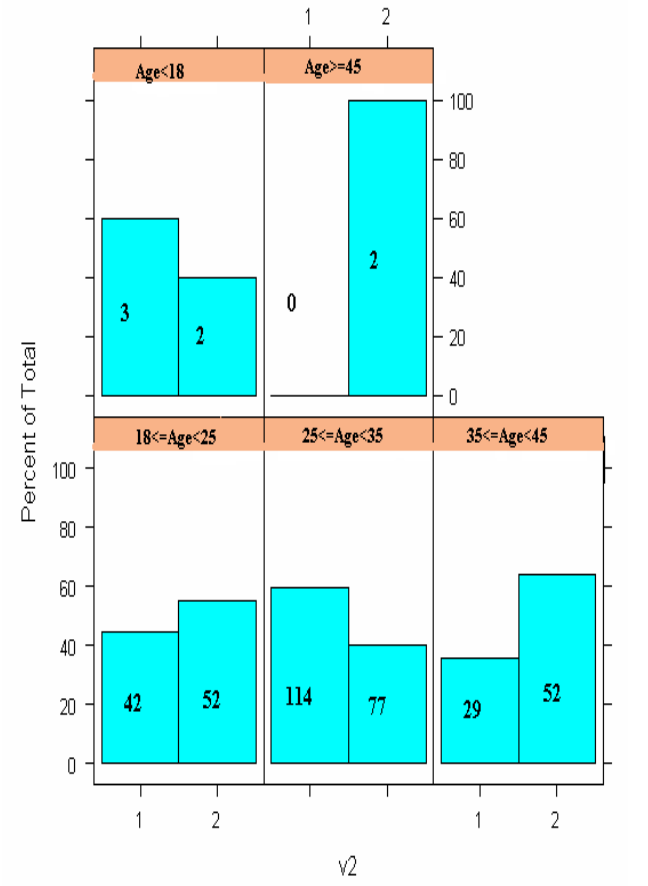

Figure 3. Age influence on preterm births.

The attribute *Education* shows that patients who have a lower education level gave birth before term. In figure 4 it can be observed that patients in class 2, who gave birth prematurely, about 78% have completed only secondary cycle.

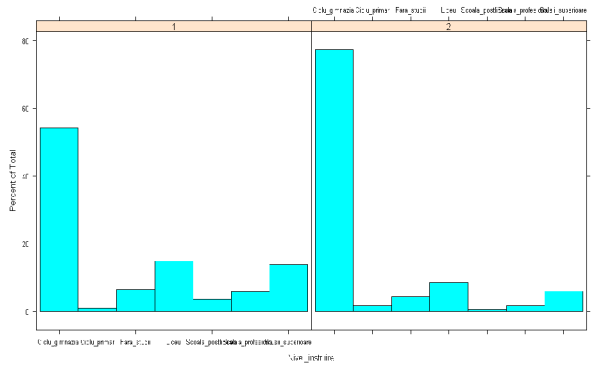

Figure 4. Type of birth in relation with education.

The attribute that refers the number of medical consultations per pregnancy is an important statistical indicator because of two aspects: the first, concerns the fact that 73% of patients have not made any time a medical consult before birth; the second is given by the large number of patients with problems who were in average 6 times at the physician (the maximum number is 15). It can also be seen in figure 5 that urban patients went more often to medical consults during pregnancy than those in rural areas. Patients with problems, who went more than 6 times at the physician, were still those from urban areas.

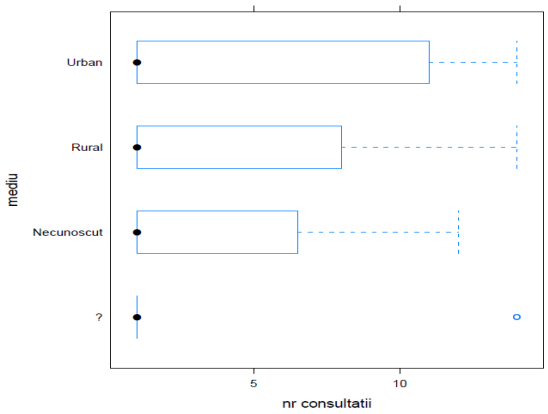

Figure 5.a) The number of medical consults per pregnancy and the living environment (urban or rural) influence preterm births.

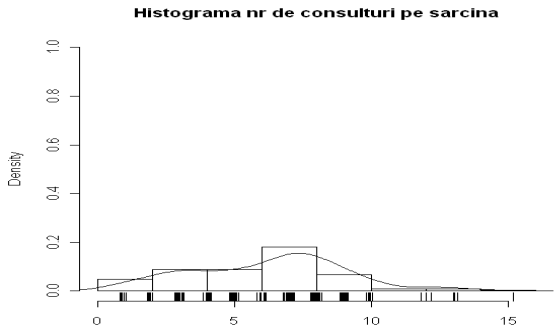

Figure 5.b) Histogram representing the number of medical consults per pregnancy.

A partial conclusion that arises is that data are very different in statistical terms and were collected with errors because there were some extreme values, inconclusive, and so some attributes may influence the risk of premature birth in a small or large extent.

Clustering hierarchical algorithms used have shown that similarity greatly influences the observations grouping[7]. Thus, for a high similarity between two observations, e.g. a short distance, it can be obtained a correct clustering if one takes into account the purpose of the work that is partitioning the medical data set into 2 classes, e.g. a cluster containing the patients who gave birth at term and another who gave birth before term. But if the similarity is low, this means the distance between patient leads to an almost impossible grouping of observations, because they form overlapping clusters within a smaller or greater extent. Specifically, some patients may belong to two or more clusters, which prove the proposed goal of this narrative. In figure 6 we remark an overlap between the 2 clusters for PAM algorithm as well as similarities between the observations, those with yellow code represent a large

distance and those with dark orange code represent a small distance.

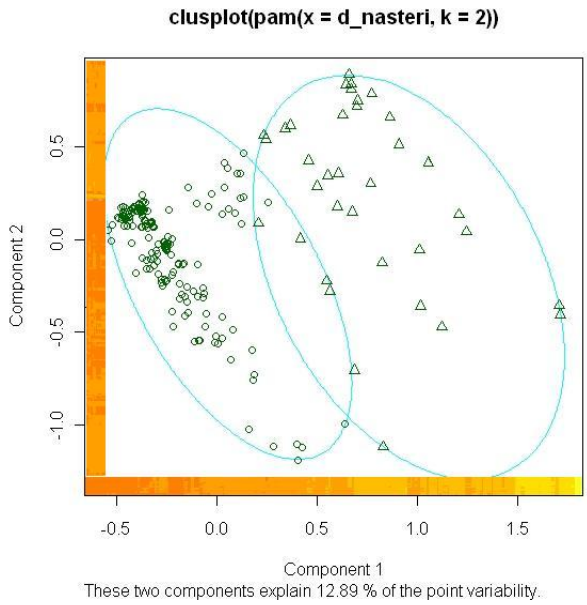

Figure 6. Heatmap representing PAM algorithm results.

Hold out method used for classifying, that method of using only part of the data for the learning phase, depends on the data set configuration. So in case the learning set is expanded to 300, the accuracy in some cases decreases. It was registered an accuracy of 0.09 SVM algorithm with the polynomial function. So it is assumed that the data set is more exposed to errors. However, the best accuracy was obtained by the NaiveBayes Algorithm, where 88% of the observations provided for learning were correctly classified. No other classification algorithm has reached that threshold. However, values exceeding 50% accuracy, were met only for cross-validation, due to the initial partitioning of the data set in learning and test sets.

```
clasif1 = MLearn(class~.,data=TMP2, svmI,
kernel="polynomial", 1:250)
```

Code 4. Support Vector Machine Algorithm with polynomial function applied on the test set.

IV. CONCLUSIONS

The set of patients is characterized by diversity which is pointed out by the values of patient's attributes. This has created difficulties in the way of completing the attributes of a patient, e.g. a similar situation is described in various ways, depending on who completed the patient medical sheet. Some data is statistically correct. This is shown by the fact that the values for different attributes are close to the normal distribution. Because of the medical context there are many fields that are not completed. Medical context refers to the fact that such patients who were registered at the time of birth came for the first time in that hospital.

Cluster analysis techniques were used to partition the dataset in two mutually exclusive clusters, but because there is a low similarity between patients attributes, we have proved that a classifying phase is required. In this phase we extended the learning set from 250 patients to 300 patients and on one hand, we showed that the accuracy has decreased, but on the other hand when we applied cross-validation on the same set, and the accuracy has increased. We concluded that the last observations contain errors and so they negatively influenced the whole dataset.

As a future direction we propose that during the learning phase it would be useful to apply a selection algorithm on the attributes in order to obtain a reduction of the size of the data and to retain only conclusive features.

REFERENCES

[1] Brandt, 1986; Gould, 1986; Kramer, 1987; Livshits et al., 1988; Livshits, 1990; Berkowitz and Papiernik, 1993, " Path analysis of risk factors leading to premature birth".

[2] Sandrine Dudoit and Robert Gentleman, "Cluster Analysis in DNA Microarray Experiments", Bioconductor Short Course Winter 2002.

[3] Kardi Teknomo, "Distance Matrix of Multivariate Data", published in 2006 on the personal website: http://people.revoledu.com/kardi/tutorial/Similarity/MutivariateDistance.html.

[4] Sandrine Dudoit, Robert Gentleman, "Classification in Microarray Experiments", Bioconductor Short Course, summer 2003.

[5] Robert Gentleman, Wolfgang Huber, Vince Carey, Raphael Irizarry, "Machine Learning, Part 1".

[6] Venables, D. M. Smith and the R Development Core Team, "An Introduction to R" manual", Version 2.10.1, 14.12.2009.

[7] N. T. Melita, S. Holban, "A Genetic Algorithm Approach to DNA Microarrays Analysis of Pancreatic Cancer", 9th International Conference on DEVELOPMENT AND APPLICATION SYSTEMS, Suceava, Romania, May 22-24, 2008.

[8] Adriana-Georgiana MALEA, "Analysis and Determination of Risk Factors Leading to Preterm Birth Using Data Mining Techniques in R", Bachelor paper, Universitatea Politehnica Timisoara, July 2009.