The Analysis and Classification of Birth Data

Raul Robu

Department of Automation and Applied Informatics, University Politehnica Timişoara, Bulevardul Vasile Pârvan, Nr. 2, 300223, Timişoara, Romania, raul.robu@aut.upt.ro

Ştefan Holban

Department of Computers, Faculty of Automation and Computers, University Politehnica Timişoara, Bulevardul Vasile Pârvan, Nr. 2, 300223, Timişoara, Romania, stefan.holban@cs.upt.ro

Abstract: The paper presents a study regarding the births that took place at the Bega Obstetrics and Gynecology Clinique, Timişoara, Romania in 2010. The analysis began from a dataset including 2325 births. The article presents a synthesis of the studies that analyze birth data. The Apgar score is the main subject in many studies. On one hand, researchers investigated the relation between the Apgar score and different factors such as the newborns' cry, the level of glucose in the blood from the umbilical cord, the mother's body mass index before the pregnancy, etc. On the other hand, there are studies that demonstrate that the Apgar score is important for the ulterior evolution of babies. The article presents the attributes from the dataset and how they were preprocessed in order to be analyzed with Weka. The values of each attribute were investigated and the results were presented. The past experience regarding births, expressed through the dataset values, was then used to build classification models. With the help of these models, the Apgar score can be estimated based on the known information regarding the mother, the baby and possible medical interventions. The purpose of these estimations is consultative, to help in identifying which values of the input variables will lead to an optimal Apgar score, in certain circumstances. The classification models were built and tested with the help of ten classification algorithms. After the model that produces the best results of classification was determined, a dedicated application was developed, with the aid of the Weka API which classifies the birth data by using LogitBoost algorithm.

Keywords: birth data; classification; data mining; LogitBoost; Weka

1 Introduction

Everyday, databases of enormous size are collected. The analysis of these data may help extract interesting and useful information. This analysis could be done individually for each attribute, following the frequency of certain values, of the medium, maximum or minimum values, etc., but also, treating the attributes together, by using data mining techniques, such as classification, clustering, discovering association rules, etc.

The classification techniques apply successfully in the medical field. Building classification models that allow an estimate, with a certain degree of confidence, of a class attribute based on different values of the input variables, may be particularly useful especially if the class attribute is a characteristic that is difficult to obtain through medical methods that they either endanger the patient's life or are prohibitively expensive. Such a classification model was built in [1], to estimate if a pulmonary nodule is cancerous or not, according to different input variables that are the nodule's characteristics and taking into consideration the patient's characteristics, obtained through noninvasive tests (the SPN diameter, border character, presence of calcification, patient's age, smoking history, results of CT densitometry, etc.). The classification model built in [1] can estimate if the nodule is cancerous or not. The medical procedure that determines whether such a nodule is cancerous or not is invasive and implies tissue sampling and analysis, an operation that is not always recommended due to certain patient's health state.

Another example of classification study in the medical field is the classification of patients that develop post operatory complications following gastric cancer operations. Two methods were used: logistic regression and neuronal networks. The data from patients in Taiwan that suffered a gastric cancer operation were used. The ANN model had a better performance than logistic regression [2].

The classification models can also be built to simulate the value of the output variable based on different values of the input variables in order to choose values for the input variables so that the output variable has the desired value.

In the dedicated literature there are many studies that investigate birth data. A part of these studies analyze and classify the newborns' cry. Another part of the studies focuses on the Apgar score. In different papers the following are investigated: the relation between the Apgar score and the newborn's cry, between the Apgar score and the glucose level from the blood of the umbilical cord, between the Apgar score and the body mass index of the mother in the pre-pregnancy period, etc. There are several studies that investigate the effects of the Apgar score on the subsequent evolution of the baby. A short synthesis of the studies regarding birth related data is presented in chapter two. These studies demonstrate that it is important for the newborns future evolution that each newborn obtains an Apgar score that is as high as possible. That is why, in this paper, we built classification models which allow an estimation of the interval in which the Apgar score will be

situated considering mother's data, baby's data and data regarding the medical interventions that could possibly help the birth. The purpose of this estimation is to determine, based on the experience of previous birth data, experience represented by the classification model and based on the mother's and baby's data, which medical interventions will lead to the best Apgar score. For example, for a given birth, we can simulate, based on experience (the model) and on the mother's and baby's data, the Apgar score if the mother gives birth naturally as opposed to a caesarean delivery. After testing the models, the result was the model performed with an 80% confidence level. These estimations can be used with a consultative role by doctors and they should help them in the decision making process, so that the newborns have an Apgar score that is as high as possible.

The analyzed data contains information regarding births that took place in 2010, at the *Bega Obstetrics and Gynecology Clinique, Timişoara, România* which we will refer to hereinafter as the *Bega Clinique*. The initial data set contains information for 2325 births and 19 attributes for each birth. The analysis was done with *Weka* [3]. Preprocessing the data was necessary in order to introduce and analyze data in *Weka*. Following this operation, data regarding 2086 births remained as well as 15 attributes for each birth. Data were analyzed both statistically and using some classification techniques. The statistical analysis revealed interesting information, such as the fact that approximately 60% of the women that gave birth in 2010 had a caesarean operation; 62 under aged women became mothers; the medium weight of the newborns was 3194 grams, but two babies that weighed over 5000 grams, etc.

The classification models were built using *Naive Bayes [4], J48 [5], k-Nearest Neighbour [6], Random Forest [7], Support Vector Machines [8], AdaBoost [9], LogitBoost [10], JRipp [11], REPTree [12], SimpleCart [13]* algorithms. They were tested through cross validation with 10 folds. The best model, from the prediction accuracy point of view, was obtained using the *LogitBoost* algorithm. It allows an estimation with an 80% accuracy of the interval of the newborn's Apgar score based on data regarding the mother, baby and medical interventions. The classification models built in the beginning of the study had little accuracy. In order to increase their accuracy we had to redo the preprocessing phase, as well as the building and testing phases, several times. After we successfully built the models with a satisfactory accuracy, the algorithm that built the best model was chosen and we developed an application dedicated to classifying data regarding births that uses the *Weka*'s *API*. The application allows building classification models with the aid of *LogitBoost* algorithm and has an interface in which data regarding the mother, newborn and medical interventions can be added, these data are used by the classification model to estimate the interval of the *Apgar score*. The reliability of the prediction resulted, after testing the model through cross validation, is approximately 80%.

2 Studies on Birth Data

In the current literature, there are different studies that investigate birth data. Some of these studies, analyze and classify the newborns cry. Newborn babies use their cries by instinct, to communicate their needs. The different cries of the infant can indicate different requirements. The paper [14] proposes a method to determine the meanings of infant cries according to a baby expert. It applies the novel *Neuro-fuzzy* techniques for the classification and Perceptual *Linear Prediction* for recognition the infant cries. The results showed that the classification performance obtained by using the *Neuro-fuzzy* techniques yielded the most desirable accuracy over other popular methods. In the paper [15] the authors created a classification model with the aid of the *Support Vector Machines* algorithm which can classify, with 76.23% accuracy, the pathologic cry. Two types of pathologies were analyzed: asthma and ischemic encephalopathy.

There are other studies that analyze the birth data that have in the middle the Apgar score. The Apgar score was introduced by *Virginia Apgar* in 1952 as a means to evaluate the health of the newborns immediately after birth [16]. Apgar score is a clinical test performed on a newborn one and five minutes after birth. It is a composite measure of breathing effort, heart rate, muscle tone, reflexes, and skin color. It is an indicator of the newborn's need for medical attention, shortly after the birth [17] [16]. The Apgar score has survived the test of time as it is still used nowadays in maternities [18] [19].

In [20] the cry characteristics of newborn infants were investigated and correlated with the Apgar scores. The cry of premature and mature infants with low and normal Apgar scores was analyzed using *principle component analysis* (*PCA*). The reduced dimension cry signal was investigated to extract features and to correlate them with Apgar scores. The paper [20] proposes the foundation of the design of an automatic algorithm, to replace the manual Apgar scoring system.

In [21] the authors define birth asphyxia based on fetal condition as measured by umbilical artery blood pH, Apgar scores, and neurologic condition of newborns.

A study of the connection between the glucose level from the blood of the umbilical cord and the Apgar score was done in [22]. The study had two major objectives. First, it tried to determine a standard reference level for the glucose in the umbilical cord. The second important objective of the study was to determine if an abnormal level of glucose in the blood of the umbilical cord influences in a negative manner the Apgar score. Following the investigations, no relation between the two factors was determined.

Obesity is a global health problem and maternal obesity may be associated with an increased risk of pregnancy complications and neonatal death. The purpose of the study in [23] was to evaluate the effect of the maternal pre-pregnancy body mass index (*BMI*) on the newborns Apgar score. The study concludes that is

recommended that obese and overweight women should be treated to normalize their *BMI* prior to pregnancy.

There are also a few studies that investigate the effect of the Apgar score to the child evolution in time. In [24] the authors investigate the association of Apgar score at five minutes with long-term neurologic disability and cognitive function. The conclusion was that five-minute Apgar score less than seven has a consistent association with prevalence of neurologic disability and with low cognitive function in early adulthood. In [25] the authors have investigated the relationship between the Apgar Scores at 5 minutes after birth and School Performance at 16 years of age. The study included 877 individuals in the analysis. Newborns with Apgar scores less than 7, at 5 minutes after birth, showed a significant increased risk of not receiving graduation grades, presumably because they went to special schools, due to cognitive impairment or other special educational needs. One out of 44 newborns with an Apgar score of less than 7 at 5 minutes after birth will go to a special school because of the antenatal or perinatal factors that caused the low Apgar score. Nearly all school children who had Apgar scores of less than 7 at 5 minutes after birth showed an increased risk of graduating from compulsory school without graduation grades in a specific subject or receiving the lowest possible grades and were also less likely to receive the highest possible grade.

As the above mentioned studies demonstrate, it is important that each newborn obtains an Apgar score as high as possible.

3 The Dataset

The data made available for analysis came from Bega Clinique. The initial data set contained data for 2325 births that took place in this hospital in 2010. Data regarding the births have been stored in a Microsoft EXCEL spreadsheet file.

For each birth, the following data were available:

- ID number
- Month of birth
- Mother's name
- Mother's age
- City of residence
- Location- urban or rural
- Gesta – the number of the pregnancy
- Para – the number of children bore by the mother

- The number of gestation weeks
- Presentation - the position of the fetus when exiting the uterus, more precisely the part of its body that is going to emerge first. The possible values for this field are: cephalic, pelvic, facial and transversal. Cephalic presentation is the normal instance, in which the fetus has the spinal cord parallel to the mother's and the head down with the chin next to the chest. The pelvic presentation means that the fetus emerges with its feet or bottom in front. The facial presentation is when the fetus looks straight ahead and its face will come out first. If the fetus' spine is not parallel (the fetus has an oblique position in the belly) the presentation is transversal.
- The Apgar score - Immediately after birth, even in the first 60 second after expulsion, in the delivery room, an assessment of the newborn's health state is made, evaluating the vital functions and its capacity to adapt to the extra uterine environment. Simultaneously with providing the first nursing measures, the neonatologist will write down the clinical state and behavior of the newborn, quantifying the vital functions with the aid of the Apgar. The Apgar score has values between 0 and 10.
- The baby's gender
- The baby's weight
- If the birth was natural or through a cesarean operation
- Videx – column that indicates if the fetus was extracted using a metal device on its head in order to help the natural birth
- The reason for which a cesarean birth was indicated
- Episiotomy – column that indicates if the perineum was cut or not in order to help the natural birth (0 – it was not, 1 – it was)
- The number of labor hours
- EMP = Manual Extraction of the Placenta, the value 0 indicates that such an intervention was not done and the value 1 indicates that it was done.

4 Data Preprocessing

Preprocessing data was done through a series of actions:

- The columns unimportant for the study were eliminated, such as the id number, the name of the mother (also for privacy reasons), the city of

residence. The column month of birth was kept for statistical analysis, but was eliminated from the number of columns used to build the model

- The errors that could be rectified were corrected. For example:
  - The weight registered for two babies was of 34000g and 34450g. We presumed that a zero was added by mistake at the end of the real weight and we removed this zero.
  - The column hours of labour contained numeric data between 0 and 30, but for two persons 20 minutes had been inserted. The column had to be uniform so it either contains the number of hours of labour or of minutes of labour, so we transformed everything in hours and replaced the two values of 20 min with 0.33 hours
  - For three women that gave birth through caesarean there was no value filled in the episiotomy field. The value 0 (without episiotomy) was filled in because this intervention is specific for the natural birth.
- The instances that contain incorrect values that could not be corrected were eliminated. Those were the following columns:
  - The column number of gestation weeks usually contained the exact number of the gestation week (for example week 40) but for 178 persons, the exact number of the gestation week was not known so the filled in values were 37/38, 38/39, etc. The corresponding instances were deleted.
  - The columns person's age, gesta, para, weeks of gestation have the value zero for ten instances
  - The Apgar score– for other 4 newborns, instead of having a score in this column, the reason for which such a score could not be indicated was filled in –that is because two of them were born at home and the other two were born in the ambulance
- The instances that have missing values in different columns were removed. Columns with missing values were person's age, gesta, para, the Apgar score, sex, weight, type of the birth, videx, reason for caesarean, episiotomy, labour hours. There were 47 instances removed because of missing values
- The attribute recommendation for caesarean was transformed from a String attribute into a nominal attribute. In a first stage, 490 unique values of this attribute were identified with the help of a developed software instrument. Then, we searched for a solution to obtain a smaller number of nominal values. The solution that was found was to group the reasons for the caesarean recommendation. Analyzing data, a number of twenty-six groups of reasons were coded using the letters of the alphabet. Each of the 490 reasons for the caesarean is part of one of the 26 identified groups. For example, the cardio-vascular group contains cerebral aneurism, aneurism, varicose disease,

hypertension, cardiopathy, cardiac insufficiency, maternal cardiac pathology, preeclampsia, mitral valve prolaps, wpw syndrome, thrombophlebitis, bradycardia. In order to fill in the place of each reason, the code corresponding to the group it is part of, a software tool was developed. The identified groups and the letters used for codification are the ones below:

- a- obstetrical antecedents of the mother
- b- cardio-vascular
- c- pelvis conformation
- d- umbilical cord
- e- placenta disattachment
- f- cervix dystocia
- g- endocrine
- h- twin pregnancy
- i- amniotic liquid
- j- broken membranes
- k- neurologic
- l- ophthalmologic
- m- fetus particularities
- n- operatory particularities or accidents
- o- placenta previa
- p- fetus presentation
- q- primipara in age
- r- negative labour trial
- s- renal
- t- uterus rupture
- u- over the term pregnancy
- v- pregnancy with treatment
- w- serological
- x- height of the mother
- y- fetal suffering

- z- unrecommended caesarean (value filled in for the mothers that gave birth naturally)

- The columns were named according to their content. For example, in the file that we received, the letter G was used to name the column that indicates the number of pregnancies (gesta) as well as the column that indicates the weight of the newborn. The newly elected names of the two columns were gesta and weight.

The analysis of the EXCEL file was done using the Filter option. A filer was set for each column and every column was checked individually so that it contained only valid data. After eliminating the instances that contained missing or incorrect data, a number of 2086 valid instances remained (out of 2325). The next phase of preprocessing was to create the *ARFF* file, with the remaining valid data and loading it into *Weka*. The built file *births1.arff* is presented in Figure 1.

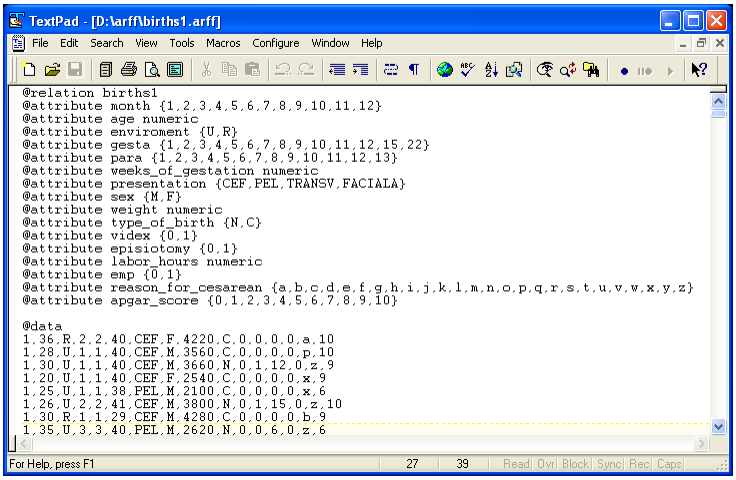

Figure 1
The file *births1.arff*

5 Statistical Analysis

Next, the preprocessed data were loaded into *Weka* and with its help the values of each attribute were analyzed (see Figure 2). The number of instances is 2086.

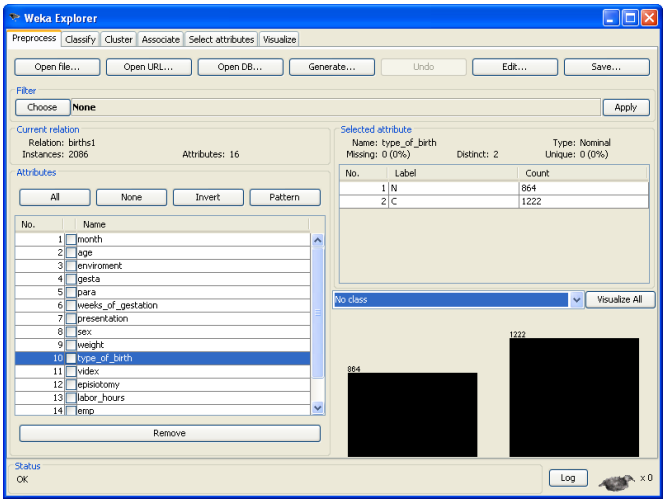

Figure 2

Statistical analysis for each attribute with *Weka*

As it is shown in Figure 3, most births were registered in May, (217 births) which means that September is most frequent month to plan a child.

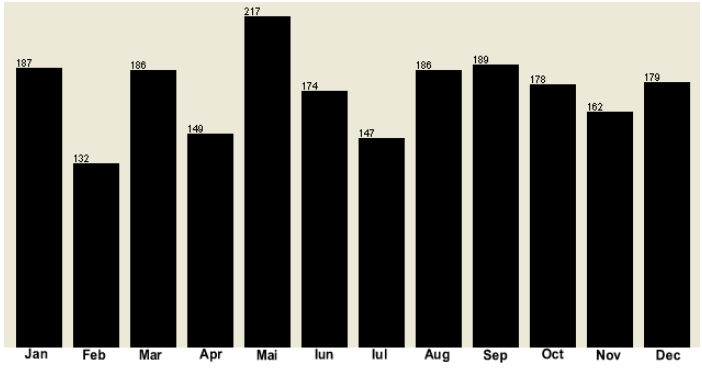

Figure 3

Number of births for each month of the year

The age of the women who gave birth had been between 12 and 45. Most women become pregnant around the age of 28. 62 legally "minor" women became pregnant and 11 of them had the age between 12 and 15. 43 other women who were over 40 years old became pregnant.

Considering their place of residence, 1363 of the women that gave birth in 2010 at Bega Clinique come from an urban location and 723 come from the rural locations.

As far as the indicator *gesta* is concerned (the number of prior pregnancies) it was observed that the majority of women previously had one pregnancy (849 women) or two (596 women), but there are also extreme situations in which women had over 10 pregnancies (23 women).

The indicator *para* states that the great majority of women had their first (1195 women) or second birth (602 women). As usual, we also encounter extreme situations, 7 women that gave birth to over 10 children.

The number of gestation weeks is a numeric value between 19 and 42, and the average value is 38.467 with a standard deviation of 2.409.

The normal presentation, cephalic is the most common (1936 cases), followed by the pelvic one 144 cases, the transversal presentation with 5 cases, and the rarest is the facial presentation (one single case).

As far as the sex of the babies is concerned we see that more boys (1091) were born than girls (995) in Bega Clinique in 2010.

The average weight of the newborns is of 3194 grams, with a standard deviation of 578 grams, but there were 2 babies born with a weight over 5000 grams (see Figure 4).

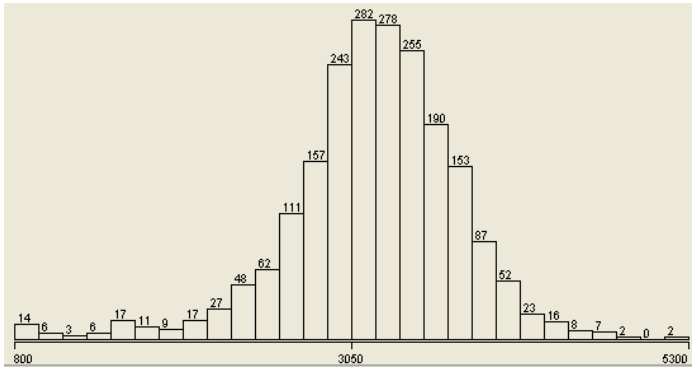

Figure 4
The weight of the newborns

Analyzing the number of women that give birth, we could say that natural birth was less recommended or desired by women than the birth through caesarean because in 2010, 864 women gave birth naturally (approximately 40%) and 1222 women through cesarean (approximately 60%).

The metal cap (videx) was used for only 18 of the 864 babies that were born naturally.

Episiotomy was necessary for 620 cases out of 864 (approximately 70%).

The number of labor hours is between zero and thirty. Usually zero hours of labour had been registered for births through caesarean.

The manual extraction of the placenta was done for only five of the women that gave birth naturally.

The most frequent reasons for caesarean recommendation are:

- The obstetric antecedents of the mother (scared uterus, double uterus, agglutinate cervix, uterus fibroma, etc.) 187 cases
- Negative labor trial, 154 cases
- Cardio-vascular reasons (preeclampsia, hypertension, cardiopathy, etc.) 132 cases

The grade given at birth is usually 10 (908 cases) or 9 (746 cases), but as it can be seen in Figure 5, the whole range of grades (from 0 to 10) was given to babies, including the very low ones.

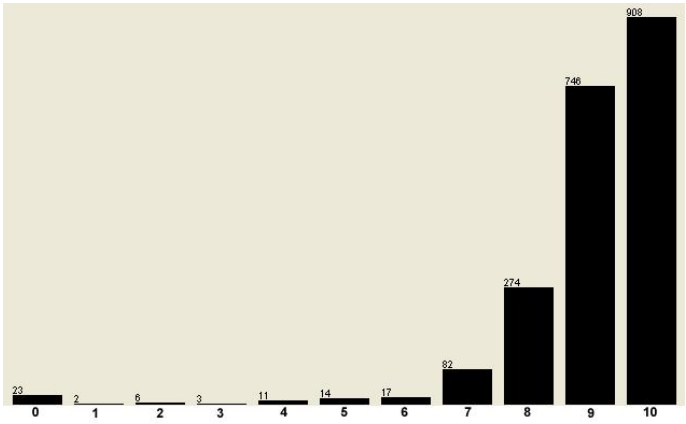

Figure 5
The Apgar score

6 Data Classification

In the next phase, we wanted to build with *Weka*'s help, classification models on birth related data with an accuracy rate as high as possible. The first set of data for which the classification models were built is presented in Figure 1, with the amendment that the attribute month was eliminated among the input attributes, as it was considered irrelevant for the classification. The data set contains 2086 instances and numeric and nominal attributes. The classification models were built

using the *Naive Bayes*, *J48*, *k-Nearest-Neighbour*, *Random forest*, *Support Vector Machines*, *AdaBoost*, *LogitBoost*, *JRipp*, *REPTree and SimpleCart* algorithms. These models were tested through cross validation, using 10 folds. Testing with this technique implies dividing the dataset in 10 subsets. Each algorithm is run 10 times. For each run, a different subset is chosen for testing and 9 other subsets are chosen for training. One by one, each of the 10 subsets will be used for testing. The prediction accuracy of the algorithm is in fact the average of the accuracy of predictions obtained by testing the algorithm on each of the 10 subsets. Testing a subset is performed by estimating the Apgar score for all instances from that subset and then comparing the real score which is known, to the estimated one. The accuracy of the prediction represents the percent of the estimations that were correct. Because the classification models initially built had a very poor accuracy, we returned to the preprocessing phase several times, worked on the dataset and rebuilt and tested the classification models. Finally, the best results were obtained for the dataset in Figure 6.

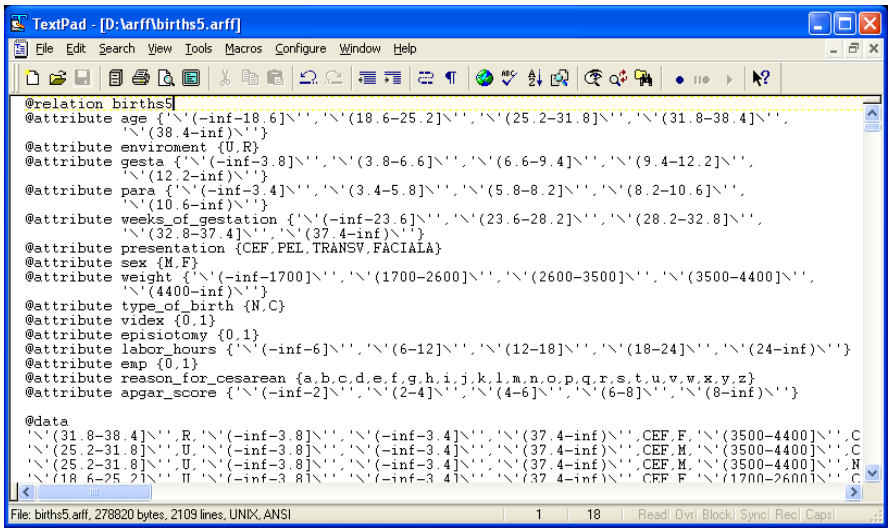

Figure 6
The file *births5.arff*

As it can be observed the numeric attributes *age*, *weight*, *weeks_of_gestation* and *labour_hours* were discretized by distributing the values on 5 equal intervals, and the nominal attributes *gesta*, *para*, *weeks_of_gestation*, *Apgar_score* were transformed in numeric attributes and discretized by dividing the values into 5 equal intervals. The new values of these attributes can be viewed in Figure 6.

The purpose for the discretization was to reduce the number of attribute values from the dataset, so that classification algorithms could build models with a higher accuracy.

The discretizations were realized using the *Filter* option from the *Preprocess* tab in *Weka* and the filter *weka.filters.unsupervised.attribute.Discretize*.

The results obtained by the classification algorithms on these data are presented in Table 1.

Table 1

The accuracy of the prediction on the *births5.arff* data set

Algorithm	**Accuracy %**
Naive Bayes	78.57
J48	79.09
k-Nearest Neighbour	74.92
Random Forest	76.22
Support Vector Machines	79.62
AdaBoost	79.91
LogitBoost	80.24
JRipp	79.67
REPTree	79.62
SimpleCart	80.00

We can see that the best classification model obtained has a 80.24% accuracy and was built using the *LogitBoost* algorithm. Next, it will be used to make predictions.

We also investigated which is the attribute with the greatest influence on the Apgar score and we determined that it is the weight of the newborn with a correlation coefficient of 38%.

7 Predictions

With the aid of the built classification model we can estimate the interval of the newborn's Apgar score based on the input variables. The input variables are data regarding the mother (age, location, gesta, para, number of weeks of gestation, number of hours of labour, recommendation for caesarean), data regarding the baby (sex, weight) which can be determined through an ecography and data regarding possible medical interventions in order to help the birth (natural or through a cesarean operation, presentation, videx, episiotomy, manual extraction of the placenta). Practically, based on the historical data consisting of 2086 recorded births a classification model was built and it can be used to make different simulations of the interval in which the Apgar score of a newborn will be, for different values of the input variables. We can simulate in which interval the Apgar score will be for different medical interventions on the mother. For

example, we can simulate which will be the Apgar score for a mother if she gives birth naturally or through caesarean.

In order to make predictions we can use *Weka*, although the mechanism used to make predictions with *Weka* is not intuitive. It implies creating a new *ARFF* file in which the instance or instances to be predicted will be filled in. This *ARFF* file is loaded in *Weka* using *Supplied test set* command. Then the *Output predictions* option is checked. Finally, we make a right click on the model built with the *J48* algorithm and choose the *Reevaluate model on current test set* option. For each instance from the test set, the class that the user filled in the *ARFF* file will be displayed as well as the class predicted by the model [26].

The second way to make predictions is to use the *Weka* version that has a dynamic interface [27] with which we can easily make predictions (see Figure 7).

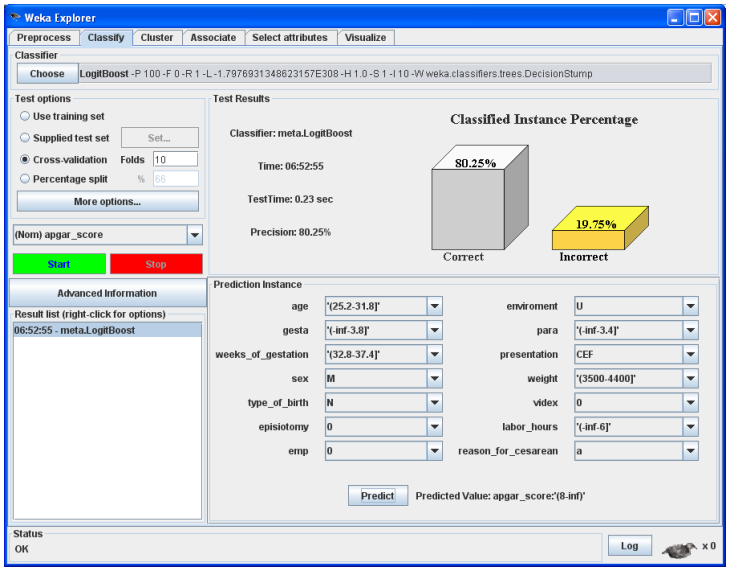

Figure 7

Making predictions with extended Weka

In this *Weka* version, after the classification model is built, the data for the instance that will be predicted can be filled in a dynamic interface, such as the one in the figure and then press the *Predict* button. The result of the prediction is displayed near the button, and the accuracy of the prediction is graphically displayed. The instance that will be predicted is introduced using *JTextField* components for the numeric attributes or *JComboBox* for nominal attributes.

Still, using the classical or extended versions of *Weka* to build classification models in order to make predictions is a little bit much considering that *Weka* offers a lot of facilities that are not used during this process. That is why it is

preferable to develop an application dedicated to classifying data regarding births, that builds the same classification model as *Weka*, using *LogitBoost* algorithm and the data regarding births from the *births5.arff* file. Such an application was developed using *Weka*'s *API* and *Java* and is presented below. The application allows using the created model to make predictions.

8 Developing an Application Dedicated to Classifying Data regarding Births

In order to build the classification model using the *LogitBoost* algorithm based on the data set that contains 2086 instances regarding births from 2010 at the Bega Clinique and to make predictions with this model, an application using *Weka*'s *API* was developed. The graphic interface was realized using *Swing* library. In Figure 8 we can see the tab *Model* that allows building the classification model using the *LogitBoost* algorithm.

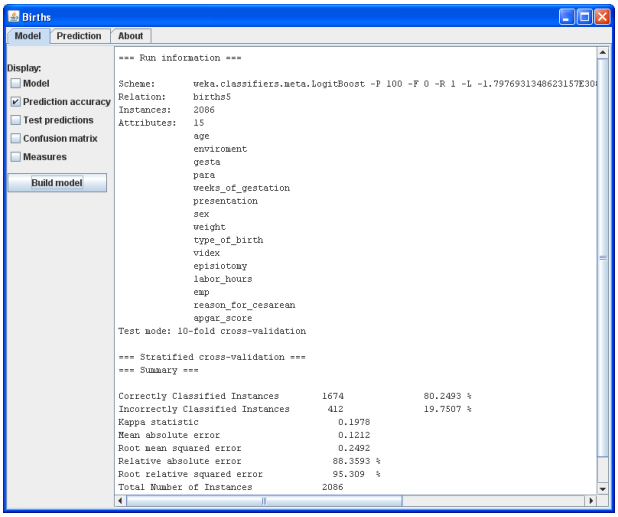

Figure 8
Births tool - model user interface

The display of created model is a user option. Checking the *Test prediction* determines the display of the actual and predicted class for each of the instances tested through *Cross validation*. This option actually corresponds to the option *Output predictions* from *Weka*. We can also display the confusion matrix and *TP Rate, FP Rate, Precision, Recall, F-Measure, ROC Area, Class* measurements by checking *Confusion Matrix* and *Measures*. Practically this tab *Model* corresponds to the *Classify* tab from *Weka* and displays the same information.

In Figure 9 the Prediction tab can be visualized. All variables from the *births5.arff* file are nominal, that is why, in the interface, there are *JComboBox* type components that permit choosing one of the nominal values for each attribute. We introduce the instance to be predicted in the interface and press the button *Predict* and the prediction made by the classification model is displayed near that *Predict* button. On the right side, there is a graph that shows the accuracy of the Prediction obtained by testing the model through Cross validation.

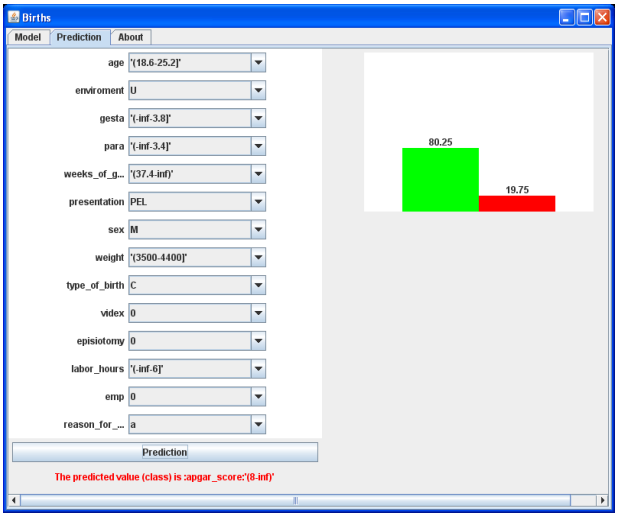

Figure 9
Births tool – prediction user interface

The most important classes from *Weka*'s *API* which have been used while developing this tool are: *Instance*, *Instances*, *Classifier*, *LogitBoost* and *Evaluation*.

The classification model was build using the *buildClassifier()* method that gets as a input parameter the instances that represent the training data set.

```
Classifier classifier =new LogitBoost ();
classifier.buildClassifier(inst);
```

Classifying an instance was done with the help of the *classifyInstance()* method from the *Classifier* class, but it can also be performed with the help of the *distributionForInstance ()* method from the same class. The instance to be classified is transmitted as an input parameter to one of the two methods. The *distributionForInstance()* method returns a *double* type vector with a number of elements equal to the number of values of the class attribute. Each element from the vector represents the probability with which the instance belongs to the corresponding class. The instance is classified as belonging to the class that has the highest probability.

The evaluation of the classifier was done with the aid of the *evaluateModelOnce AndRecordPrediction(classifier, test_instance)* method from the *Evaluation* class.

Conclusions

The paper presents an analysis on the data regarding the births that took place in the Bega Clinique. The initial data set contained 2325 records and after preprocessing, 2086 instances remained. Preprocessing consisted of eliminating the attributes considered irrelevant for the study, eliminating instances that contained missing or incorrect values, rectifying, if possible, the data that were incorrectly filled in, etc. Next, a statistical analysis for each attribute was made and it revealed some interesting information, such as the fact that approximately 60% of the women that gave birth in this hospital underwent a cesarean operation. The following step was building the model, with the help of ten classification algorithms from *Weka* classifiers that allow an estimation of the interval for the value of the Apgar score based on data from the mother, newborn and medical interventions. Finally, a dedicated tool was developed in *Java*, using *Weka API* which builds a classification model with the help of the *LogitBoost* algorithm for the data set regarding births. The application performs predictions with the aid of the built classification model. This tool helps make different simulations of the Apgar score for different values of the input variables with the purpose to choose the input variables so that the Apgar score is optimal.

Acknowledgment

This work was partially supported by the strategic grant POSDRU/159/1.5/S/137070 (2014) of the Ministry of National Education, Romania, co-financed by the European Social Fund – Investing in People, within the Sectoral Operational Programme Human Resources Development 2007-2013.

References

[1] A. Kusiak, J. Kern, K. Kernstine and B. Tseng, Autonomous Decision-Making: A Data Mining Approach, *IEEE Transactions on Information Technology in Biomedicine*, pp. 274-284, 2000

[2] L. Yi-Chih, L. Yang-Chu and L. Tian-Shyug, Mining the Complication Pattern of Gastric Cancer Patients by Using Artificial Neural Networks and Logistic Regression, *The Journal of Human Resource and Adult Learning*, pp. 151-155, November 2006

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann & I. H. Witten, The WEKA Data Mining Software: An Update, *ACM SIGKDD Explorations Newsletter* (2009), Volume 11, Issue 1, pp. 10-18

[4] G. H. John and P. Langley, Estimating Continuous Distributions in Bayesian Classifiers, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 338-345, 1995

[5] J. R. Quinlan, C4.5: Programs for Machine Learning, *Morgan Kaufmann Publishers*, Inc., pp. 235-240, 1993

[6] D. Aha and D. Kibler, Instance-based Learning Algorithms, *Machine Learning*, Volume 6, pp. 37-66, 1991

[7] L. Breiman, Random Forests, *Machine Learning* 45(1), pp. 5-32, 2001

[8] C. C. Chang and C. J. Lin, LIBSVM: a Library for Support Vector Machines, *ACM Transactions on Intelligent Systems and Technology*, 2, No 3, 2011

[9] Y. Freund and R. E. Schapire, Experiments with a New Boosting Algorithm, *In: Thirteenth International Conference on Machine Learning*, San Francisco, pp. 148-156, 1996

[10] J. Friedman, T. Hastie and R. Tibshirani, Additive Logistic Regression: a Statistical View of Boosting (with discussion and a rejoinder by the authors), *The annals of statistics*, 28, No. 2, pp. 337-407, 2000

[11] W. W. Cohen, Fast Effective Rule Induction, *In: Twelfth International Conference on Machine Learning*, pp. 115-123, 1995

[12] T. Elomaa and M. Kaariainen, An Analysis of Reduced Error Pruning, *Journal of Artificial Intelligence Research*, Volume 15, pp. 163-187, 2001

[13] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, Classification and Regression Trees, *Wadsworth International Group*, 1984

[14] K. Srijiranon and N. Eiamkanitchat, Application of Neuro-Fuzzy Approaches to Recognition and Classification of Infant Cry, In TENCON 2014-2014 IEEE Region 10 Conference, pp. 1-6, 2014

[15] A. Chittora and H. A. Patil, Classification of Pathological Infant Cries using Modulation Spectrogram Features, *In Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*, pp. 541-545, 2014

[16] V. Apgar, A Proposal for a New Method of Evaluation of the Newborn, *Curr Res Anaesth*, 32, pp. 260-267, 1953

[17] V. Naeser, N. Kahr, L. G. Stensballe, K. O. Kyvik, A. Skytthe, V. Backer, C. G. Carson and S. F. Thomsen, Apgar Score Is Related to Development of Atopic Dermatitis: Cotwin Control Study, *Journal of Allergy*, pp. 1-6, 2013

[18] M. Finster and M. Wood, The Apgar Score has Survived the Test of Time, *Anesthesiology*, 102, No. 4, pp. 855-857, 2005

[19] B. M. Casey, D. D. McIntire and K. J. Leveno, The Continuing Value of the Apgar Score for the Assessment of Newborn Infants, *New England Journal of Medicine*, 344, No. 7, pp. 467-471, 2001

[20] R. Sahak, W. Mansor, L.Y. Khuan, A. Zabidi and F. Yasmin, An Investigation into Infant Cry and Apgar Score using Principle Component Analysis, *Signal Processing & Its Applications, CSPA 2009. 5th International Colloquium on*, pp. 209-214, 2009

[21] L. C. Gilstrap, K. J. Leveno, J. Burris, M. L. Williams and B. B. Little, Diagnosis of Birth Asphyxia on the Basis of Fetal pH, Apgar Score, and Newborn Cerebral Dysfunction, *American journal of obstetrics and gynecology*, 161, No. 3, pp. 825-830, 1989

[22] K. Khan and A. R. Saha, A Study on the Correlation between Cord Blood Glucose Level and the Apgar Score, *Journal of clinical and diagnostic research*, Volume 7, Issue 2, pp. 308-11, 2013

[23] L. Sekhavat and R. Fallah, Could Maternal Pre-Pregnancy Body Mass Index Affect Apgar Score?, *Archives of gynecology and obstetrics*, 287, No. 1, pp. 15-18, 2013

[24] V. Ehrenstein, L. Pedersen, M. Grijota, G. L. Nielsen, K. J. Rothman, and H. T. Sørensen, Association of Apgar Score at Five Minutes with Long-Term Neurologic Disability and Cognitive Function in a Prevalence Study of Danish Conscripts, *BMC Pregnancy and childbirth*, 9, No. 1, 2009

[25] A. Stuart, P. O. Olausson and K. Källen, Apgar Scores at 5 Minutes after Birth in Relation to School Performance at 16 Years of Age, *Obstetrics & Gynecology* 118, No. 2, Part 1, pp. 201-208, 2011

[26] D. Rodríguez, Making Predictions on New Data using Weka, Available at: http://www.cc.uah.es/drg/courses/datamining/ClassifyingNew DataWeka.pdf, Accessed: 2014.09.15

[27] R. Robu and C. Hora, Medical Data Mining with Extended Weka, *Proceedings of the 16th IEEE International Conference on Intelligent Engineering Systems (INES)*, pp. 347-350, 2012