Classification Algorithms of Data Mining Applied for Demographic Processes

Irina Ioniță
Petroleum-Gas University of Ploiești,
Blvd. Bucuresti, No.39,
Ploiesti, 100680, Romania
irinat@upg-ploiesti.ro

Liviu Ioniță
Petroleum-Gas University of Ploiești,
Blvd. Bucuresti, No.39,
Ploiesti, 100680, Romania
iliviu@upg-ploiesti.ro

Abstract

Data mining is a generous field for researchers due to its various approaches on knowledge discovery in enormous volumes of data that are stored in different formats. At present, data are widely used all over the world, covering areas such as: education, industry, medicine, banking, inssurance companies, research laboratories, business, military domain etc. The major gain from applying data mining techniques is the discovery of unknown patterns and relations between data which can further help in the decision-making processes. There are two forms of data analysis used to extract models by describing important classes or to predict future data trends: classification and prediction. In this paper, the authors present a comparative study of classification algorithms (i.e. Decision Tree, Naïve Bayes and Random Forest) that are currently applied to demographic data referring to death statistics using KNIME Analytics Platform. Our study was based on statistical data provided by the National Bureau of Statistics of the Republic of Moldova corresponding to years 2011 and 2012, data related to deaths and various classification attributes, such as causes of death, areas, sex, years and age groups. A detailed proposal on the possibilities to increase the models' accuracy was also provided in the paper. Our findings indicated that the highest accuracy was achieved by the Decision Tree model (over 90%).

Keywords: Data mining, classification algorithms, Decision Tree, Random Forest, Naïve Bayes, demographic processes

1. Introduction

In the last years, data mining has demonstrated to be an activity of interest because it allows the exploration of high volumes of data stored in different formats in order to extract implicit, potentially useful, and previously unknown information (Witten, Frank and Hall, 2011). The results that one can obtain after applying data mining algorithms are quite surprising, as well as the possibility to improve their research activity. Banks with statistic data are a continuous resource for data engineers or other individual interested in knowledge discovery and data mining. Before benefiting from data mining functionalities, the researcher must have followed several steps such as: data selection, data cleaning, data transformation, data mining (models), pattern evaluation and interpretation, knowledge presentation (Witten, Frank and Hall, 2011; Han and Kamber, 2000). Methods used for analyzing and modeling data can be split in two important categories: supervised learning and unsupervised learning. The former category requires input data identified as predictors (independent variables) and a target (dependent variable) whose value is to be estimated. In this case, the model learns how to predict the value of the target variable, by considering the predictors. We mention several examples of supervised learning, such as: decision trees, regression analysis, neural networks, etc. The latter group of methods is represented by: cluster analysis, correlation,

factor analysis, etc. Unsupervised learning treats all variables equally and the main goal is to find patterns and associations between data (Tan, Steinbach and Kumar, 2005).

In this paper the authors developed a comparative study on data mining algorithms (Decision Tree, Naïve Bayes and Random Forest) based on death statistics data in order to classify the population into seven risk classes, considering attributes such as: cause of death, sex, environment or age category. The obtained models were further used in order to estimate risk classes for the new statistical data corresponding to the year 2016. This paper consists of five sections, organized as follows: section 2 reviews the literature on the topic discussed, section 3 provides a brief presentation of the data mining algorithms used in experiments and the methodology used to conduct the prediction analysis. The experimental results and discussions are presented in Section 4, whereas conclusions and future work are suggested in the last section.

2. Related work

Identifying the factors that influence death rate, the relation between these factors, as well as their significance, represents the main challenge in data mining. Death prediction is a sensible topic, though approached in the literature, as proven in (Sarvestani, Safavi, Parandeh and Salehi, 2010; Shoiab, Ajit and Hisham, 2017; Chen, Huang, Hong, Cheng and Lin, 2011; Rathore, Tomar and Agarwal, 2014; Ioniţă, and Ioniţă, 2016). Findings of such research studies can be used in order to prevent diseases and poor health condition of the population, to perform various activities such as immunization of children or to inform young women regarding the risk of breast cancer and so forth. In Sarvestani et al., (2010) an analysis of the prediction of the survivability rate of breast cancer patients by using data mining techniques is presented. The authors investigated three data mining techniques, namely Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms. After performing several experiments, the authors concluded that the achieved prediction performances were comparable to the existing techniques and that C4.5 algorithm has a much better performance than the other two techniques. In the approach proposed in Shoiab et al. (2017), death prediction was performed by using various classifiers and the results were analyzed by using errors. The authors used various classification techniques such as: Multilayer Perceptron, Multilayer Regression, SMOreg and Linear Regression. Experimental results proved the efficacy of the proposed approach in terms of higher accuracy when using various statistical error measure (Shoiab et al. 2017). Other studies focused on heart diseases prediction (Chen, Huang, Hong, Cheng and Lin, 2011; Masethe and Masethe, 2014). Data mining algorithms such as J48, Naïve Bayes, REPTREE, CART, and Bayes Net were applied in this research in order to predict heart attacks. The research results revealed a prediction accuracy of 99%. In Chen, Huang, Hong, Cheng and Lin, (2011) the authors developed an artificial neural network algorithm used for classifying heart disease based on several features and a user-friendly heart disease predict system (HDPS). The accuracy of the prediction was near 80%. Paper (Ioniţă, and Ioniţă, 2016) presents a case study on the classification of patients with thyroid dysfunctions into three classes (i.e. 1 – hypothyroidism, 2 – hyperthyroidism, 3- normal) by using CART and TreeNet models and discusses possible methods to improve the accuracy of the considered classification models. In the experiments described in Ioniţă, and Ioniţă, (2016), by comparing the obtained results with those already existing in the literature, for most of the experiments, the accuracy of CART model was over 93% and the accuracy of TreeNet model was 94.97%.

3. Research methodology

For the experiments developed in this paper, we proposed the following data mining algorithms: Decision Tree, Naïve Bayes and Random Forest, well-known supervised learning methods (Caruana and Niculescu-Mizil 2006).

In order to solve a given problem of supervised learning, the following steps are needed:

- determine the type of training examples;
- gather a training set (that needs to be representative);

- determine the input feature representation of the learned function (the accuracy of the learned function depends very much on the feature vector, which should not be too large, but which should also contain enough information so that one may accurately predict the output);
- determine the structure of the learned function and the corresponding learning algorithm (Decision Tree, Naïve Bayes, Neural Networks etc.);
- run the learning algorithm on the gathered training set;
- evaluate the accuracy of the learned function (using a test set).

A decision tree is a very simple representation used for classifying observations and it graphically consists in a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and the tree leaves are the classes or class distributions (Quinlan, 1986)**Eroare! Fără sursă de referinţă.**. Example of decision trees mentioned in the literature are: ID3, C4.5, CART etc. Algorithms used for constructing decision trees usually work top-down and consist in choosing a variable at each step that best splits the set of items. The measures developed for selecting the best split are usually based on the degree of the child nodes (the smaller the degree of impurity, the more skewed the class or class distribution). Example of impurity measures includes: Entropy(t), Gini(t) and Classification error(t), were *t* denotes the *t* node of the decision tree (Tan, Steinbach and Kumar, 2005).

Naïve Bayes is a classification algorithm used for solving binary problems (two class) and multi-class classification problems and it is particularly suited when the dimensionality of the inputs is high. The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and can handle an arbitrary number of independent variables whether continuous or categorical. Parameter estimation for Naive Bayes classifier uses the method of maximum likelihood (Caruana and Niculescu-Mizil 2006).

Random forests or random decision forests (Ho, 1995) are an ensemble learning method used for classification, regression and other tasks, the main tasks being to build a multitude of decision trees at training time and to output the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Some of the random forests features are: they run efficiently on large databases, they can handle thousands of input variables, and generate an internal unbiased estimate of the generalization error during the forest building process (Prinzie and Van den Poel, 2008)**Eroare! Fără sursă de referinţă.**.

In the next section, we shall discuss the experiments performed within our research and the results on predicting death risk classes for population, obtained by considering several predictor variables that are detailed below.

4. Results and discussions

Our study used statistical data provided by the National Bureau of Statistics of the Republic of Moldova (2017)[7] corresponding to years 2011 and 2012, that referred to several deaths causes, classified based on areas, sex, years, and age groups. As an analytic platform of data mining, we used KNIME (2017)[8]. Accessed in October 2017, an open solution for data-driven innovation that helps researchers to discover the potential hidden in the enormous volumes of data, mine for unknown insights, or predict new futures. The workflow developed for the current problem is presented in Figure 1 (Decision Tree Model and Naïve Bayes Model) and Figure 2 (Random Forest).

[7] National Bureau of Statistics of the Republic of Moldova. (2017), http://www.statistica.md/. accessed in November 2017.

[8] KNIME. (2017), *https://www.knime.com/knime-analytics-platform.* Accessed in October 2017.

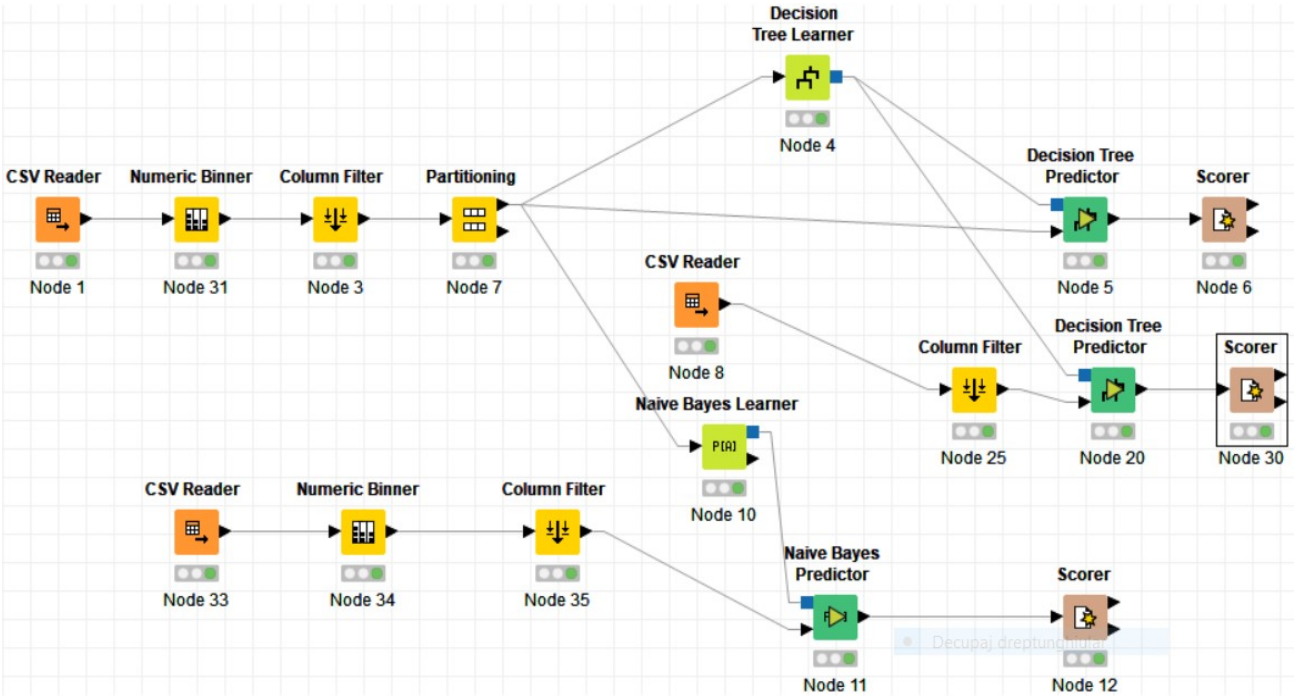

Figure 1. KNIME workflow for Decision Tree model and Naïve Bayes model

The model variables are: *Sex* {male, female}, *AgeCategory* {age0, age1-4, age5-9, age10-14, age15-19, age20-24, age25-29, age30-34, age35-39, age40-44, age45-49, age50-54, age55-59, age60-64, age65_more}, *CausesDeathCategory* {CauseDeathi, i=1 to 22}, *Environment* {Urban, Rural}, *NumberDeaths, DeathRiskClass*{A, B, C, D, E}, the last one being the dependent variable. One variable is numeric (*NumberDeaths*), and the rest are categorical. Some examples of the values of the predictor *CausesDeathCategory* are: *Infections and parasitic diseases, Intestinal infections, Endocrine nutritional and metabolic diseases, Diabetes* (codified as *CauseDeath1, CauseDeath2* etc.). The variable *NumberDeaths* refers to the incidence of death due to a certain disease. The initial data set (2640 observations) was manually preprocessed so that to obtain a training data set suited for KNIME.

KNIME node Decision Tree Learner is responsible with training the classification model, based on dataset uploaded by the input node CSV Reader. In a typical data mining project, it is a good practice to evaluate the performance of the model created by applying it on a holdout sample. Therefore, the available dataset needs to be partitioned.

In KNIME, there are dedicated nodes for applying classification models such as: Decision Tree Learner, Decision Tree Predictor, Naïve Bayes Learner, Naïve Bayes Predictor, Random Forest Learner, Random Forest Predictor. After a model *learns* how to classify data (through learner node), we can classify new data by using the predictor node. Scorer node is responsible with accuracy measurements (confusion matrix, accuracy of classification models, Cohen's kappa, Precision, Sensitivity etc.). Column Filter node is used to exclude several attributes from the initial dataset and to maintain only the predictors considered for the classification model. In Figure 2, the KNIME workflow for the Random Forest model is presented. The model *learns* to classify data (Random Forest Learner – node 21) and the model accuracy is analyzed by the node 24 (Scorer). For classifying new data, we used another CSV Reader node and Random Forest Predictor (node 28), the obtained classification accuracy being given by Scorer node (node 29).

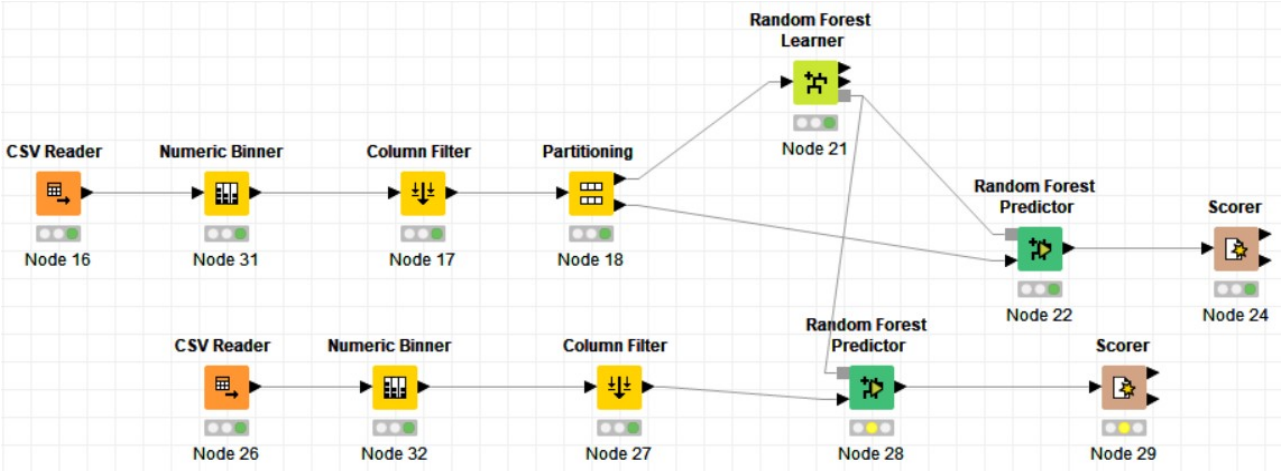

Figure 2. KNIME workflow for Random Forest model

The Partitioning node splits the initial dataset in: training dataset (70%) and validation dataset (30%) (Table 1). In this experiment, we obtained the following accuracy percentages: 95.2% (Decision Tree); 84.4% (Naïve Bayes); 83.4% (Random Forest).

Another experiment consisted in considering the first partition which contained a sample with 1500 records taken from the top of the initial dataset. We obtained the following results for the accuracy rate: 94% - 1410 correct classified data (Decision Tree); 82.01% -935 correct classified data (Naïve Bayes); 86.57% - 987 correct classified data (Random Forest). The second partition was used to test the classification model.

Table 1. Accuracy statistics for classification models

Partition=70% Draw randomly		**Accuracy statistics**	**Classification models**		
			Decision Tree	**Naïve Bayes**	**Random Forest**
DeathRiskClass	**A**	TP Rate	1300	588	541
		FP Rate	28	84	114
		TN Rate	497	116	136
		FN Rate	22	5	2
		Recall	0.983	0.992	0.996
	B	TP Rate	119	3	3
		FP Rate	29	7	4
		TN Rate	1663	735	715
		FN Rate	36	48	71
		Recall	0.768	0.059	0.041
	C	TP Rate	207	40	70
		FP Rate	25	21	11
		TN Rate	1594	678	664
		FN Rate	21	54	48
		Recall	0.908	0.426	0.593
	D	TP Rate	126	38	47
		FP Rate	7	12	3
		TN Rate	1704	728	733
		FN Rate	10	15	10
		Recall	0.926	0.717	0.825
	E	TP Rate	6	0	0
		FP Rate	0	0	0
		TN Rate	1841	791	792
		FN Rate	0	2	1
		Recall	1	0	0
Accuracy			*0.952 (95.2%)*	*0.844 (84.4%)*	*0.834 (83.4%)*

In order to validate the model, we used a dataset with records of deaths corresponding to the year 2016, with the same structure as the training dataset. After loading the test data, we obtained similar values for the accuracy: 79.31% (Decision Tree), 79.02% (Naïve Bayes), 80.09% (Random Forest).

In the previous experiment we considered the information gain ratio for the Random Forest model and the *Gini index* for Decision Tree as a split criterion. When the split criterion in the Gini index was changed for the Random Forest model, we obtained 72.28% accuracy (824 correct classified data). As presented in Figure 3, the highest values of the mean accuracy were obtained for the Decision Tree model (over 90%), followed by Random Forest for the sample size 1500.

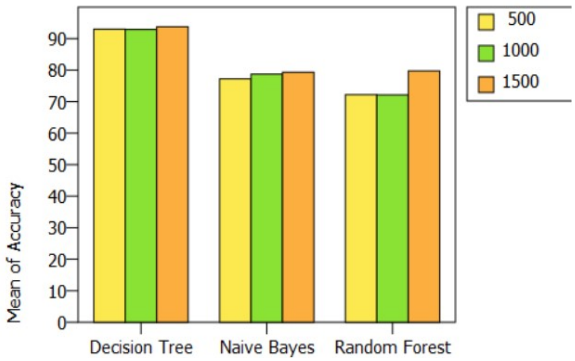

Figure 3. Mean of accuracy for classification models

In Figure 4 we represented the maximum of accuracy obtained in our experiments, considering the two types of partition: the observation taken from the top of the initial dataset (the yellow bars), respectively the partition that was drawn randomly (the green bars).

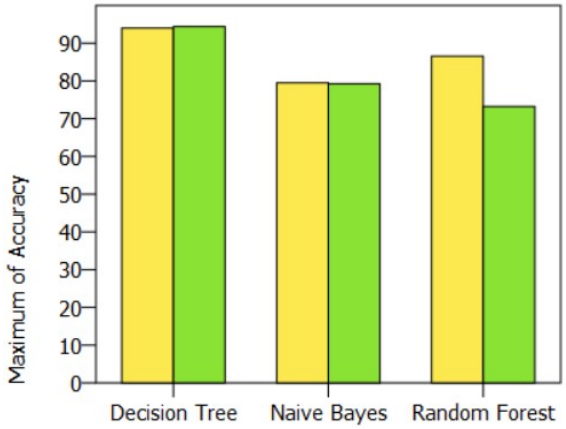

Figure 4. Maximum of accuracy for classification models divided by the type of partition

Decision Tree accuracy proved to have the highest values for both types of partition, followed by Random Forest model (only for the first type of partition).

5. Conclusions

Classification models such as Decision Tree, Naïve Bayes and Random Forest are useful tools that can be used to predict or estimate categorical class labels, in our case, the variable *DeathRiskClass*. A dataset on death statistics was analyzed in order to discover potential information about the relation between several personal details (age, sex, environment, the cause of death, the number of deaths) and the death risk class. The study focused on finding a classifier that can "guess" the class label correctly, one proof of its precision being the accuracy of the proposed model. A comparison of the above-mentioned classification models was performed and after the evaluation and the interpretation phases, we concluded that the highest value of accuracy was achieved by using the Decision Tree model (over 90%), for all the

experiments. Future research will consist of studying the behavior of other classification models (i.e. neural networks) and the challenges posed by clustering.

References

Caruana, R. and Niculescu-Mizil, A. (2006), An empirical comparison of supervised learning algorithms. In: *Proceedings of the 23rd international conference on Machine learning,* ACM, pp. 161-168.

Chen, A.H., Huang, S.Y., Hong, P.S., Cheng, C.H. and Lin, E.J. (2011), HDPS: Heart disease prediction system. In: *Computing in Cardiology*, IEEE, pp. 193–198.

Han, J. and Kamber, M. (2000), Data Mining: Concepts and Techniques. Morgan Kaufmann

Ho, T.K. (1995), Random Decision Forests (PDF). *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995*. pp. 278–282.

Ioniţă, I. and Ioniţă, L. (2016), Applying Data Mining Techniques in Healthcare. Studies in Informatics and Control, ISSN 1220-1766, vol. 25(3). pp. 385-394.

Masethe, H.D., and Masethe, M.A. (2014), Prediction of Heart Disease using Classification Algorithms. *Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, 22-24 October, San Francisco, USA*.

Prinzie, A. and Van den Poel, D. (2008), Random Forests for multiclass classification: *Random MultiNomial Logit. Expert Systems with Applications.* 34 (3), pp. 1721–1732.

Quinlan, J. R. (1986), *Induction of Decision Trees. Machine Learning 1*, Kluwer Academic Publishers. pp. 81-106.

Rathore, N., Tomar, D. and Agarwal, S. (2014), Predicting the survivability of breast cancer patients using ensemble approach. *International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, DOI: 10.1109/ICICICT.2014.6781326.

Sarvestani, A. S, Safavi, A.A., Parandeh, N.M. and Salehi, M. (2010), *Predicting breast cancer survivability using data mining techniques.* 2. DOI: V2-227. 10.1109/ICSTE.2010.5608818.

Shoiab, A., Ajit, D. and Hisham, A. (2017), *Death Prediction and Analysis using Web Mining Techniques*.

Tan, P.N., Steinbach M., and Kumar, V. (2005), *Introduction to Data Mining, Chapter: Classification: Basic Concepts, Decision Tree, and Model Evaluation*. Addison-Wesley. ISBN: 0321321367, https://www-users.cs.umn.edu/~kumar001/dmbook/ch4.pdf.

Witten, I., Frank, E. and Hall, M. 2011. *Data mining: practical machine learning tools and techniques.* Morgan Kaufmann.

Irina IONIȚĂ (b. June 30, 1979) received her BSc in Automation and Industrial Informatics (2002), PhD in Automatic Control (2011) from Petroleum-Gas University of Ploiești. Now she is a lecturer in Department of Computer Science, Information Technology, Mathematics and Physics, Faculty of Letter and Sciences, Petroleum-Gas University of Ploiești, Romania. Her current research interests include different aspects of Data Mining applied in various domains (banking, education, industry, health etc.). She has (co-) authored 3 books and more than 40 papers, more than 10 conferences participation, and is a member in more than 5 research projects teams.

Liviu IONIȚĂ (b. April 27, 1973) received his BCs in Mathematics-Informatics (1997), MSc in Information Science (2010) and Advanced Control and Programmable Structures (2010), PhD in Automatic Control (2014) from Politehnica University of Bucharest. Now he is a lecturer in Department of Computer Science, Information Technology, Mathematics and Physics, Faculty of Letter and Sciences, Petroleum-Gas University of Ploiești, Romania. His current research interests include: artificial intelligence, information systems, multi-agent systems, multimedia technologies, operating systems. He has (co-) authored 13 books and more than 30 papers, more than 10 conferences participation and workshops.