International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 21 (2016) pp. 10374-10379

Fertility Analysis Method Based on Supervised and Unsupervised Data Mining Techniques

[1] Mendoza-Palechor, Fabio E.; [2] Ariza-Colpas, Paola P.; [3] Sepulveda-Ojeda, Jorge A.; [4] De-la-Hoz-Manotas, Alexis, [5] Piñeres Melo, Marlon

[1,2,3,4] Systems Engineering Department, Universidad de la Costa, Colombia.
[5] Systems Engineering Department, Universidad Autonoma del Caribe, Colombia.

Abstract
The fertility potential analysis represents a research topic of great interest, and could help us to understand all the factors that difficult to have high fertility rates, which its quite relevant to be able to propose solutions and obtain an increase in the fertility levels especially for men. Data mining techniques are widely used for many authors studying this condition, and the metrics to evaluate results using these techniques usually are: accuracy, coverage, positive false and negative false rate. In this study we implemented the following data mining techniques: Decision Trees, Support Vector Machines, Bayesian Networks and K-Nearest Neighbor, and obtained high levels in all metrics, so we can conclude that our method proposal represents a tool to support decision making for patient analysis with fertility problems.

Keywords: Data Mining, Decision Trees, Support Vector Machines, Bayesian Networks, K-Nearest Neighbor.

INTRODUCTION

The World Health Organization (WHO) defines infertility as a disease. It's a reproductive system disease that manifest as the inability of obtain clinical pregnancy after 12 months of having unprotected sexual relationships. Infertility causes several effects in different types of personal health: physical, mental, emotional, psychological, social and even religious, in the couples that suffer from it. It's one of the most important causes of depression, and its social, psychological and cultural consequences have been catalogued in six levels of severity, ranging from guilt, fear, depression to violent death or suicide [1]. This behavior is also confirmed by [2], and clearly infertility can affect negatively generating frustration, and personality weakness.

The fertility potential analysis represents a research topic of great interest, and could help us to understand all the factors that difficult to have high fertility rates, which its quite relevant to be able to propose solutions and obtain an increase in the fertility levels especially for men. Some authors [3], say that during the last three decades, several reports have suggested that the semen quality in regular men has decreased, and [4] talks about a trend in decrease of sperm count and seminal fluid volume in the last fifty years.

In [5] they mention that fertility rates have decreased drastically in the last two decades, especially in men, due to environmental issues and lifestyle that can affect the quality of semen. Several artificial intelligence techniques have become an emergent technology for decision support systems in medicine to patient identification with fertility problems.

In [6] a research was conducted to study semen volume, sperm concentration, progressive motility, vigor and percentage of normal forms and multiple anomalies. The semen volume didn't decrease, but an important decrease in total sperm count was found (443,2 million in 1976 to 300.2 million in 2009), also in motility (64% in 1976 to 49% in 2009) and vigor (88% to 80%).

In [7] the authors consider that semen analysis is standard for routine diagnosis of infertile couples studies through the sperm count, and it´s strongly related to male infertility, and they also express the importance of sperm concentration in male infertility, since it is a relevant factor in diagnosis of this disease.

Based on previous studies in existing literature, we can infer that the proposed model would become an excellent tool to identify patterns and predict behaviors in fertility analysis, and also suggest its implementation in other areas of healthcare.

MATERIALS AND METHODS

Dataset Analysis and Preparation :

For this research the "Fertility Dataset" was used, stored in the Machine Learning Repository UCI [8]. This dataset has 10 attributes and 100 instances. In the next table the attributes of the dataset are presented:

Table 1: Fertility Dataset Attributes

Attribute	Description
Season in which the analysis was performed	• Winter • Spring • Summer • Fall. (-1, -0.33, 0.33, 1)
Age at the time of analysis	• 18-36 (0, 1)
Childish diseases	• yes=0 • no=1
Accident or serious trauma	• yes=0 • no=1

Surgical intervention	• yes=0 • no=1
High fevers in the last year	• less than three months ago =-1 • more than three months ago= 0 • no=1
Frequency of alcohol consumption	• several times a day • every day • several times a week • once a week • hardly ever or never
Smoking habit	• never = -1 • occasional=0 • 3) daily=1
Number of hours spent sitting per day	• 0 • 1
Output: Diagnosis	• normal (N) • altered (O)

Source: *Created by author*

Decision Trees :

Decision trees are considered one of the data mining algorithms most widely used for classification and prediction. Their structure is based on nodes, where each interior node corresponds to one entry variable and is divided in children nodes from the values of each input variable. Each leaf node represents a particular value of an output variable.

In the execution of the decision tree, samples in each interior node are divided in subsets based on attributes, and this process is repeated in each subset derived from recursive partition. In every step, during the growth of a decision tree, one of the input variables is selected for sample division. On the chosen variable base, the distribution point is determined through value test on the attribute, and as a result, the most used tests are impurity and entropy [9].

Support Vector Machines (SVM) :

They are decision algorithms that allow to solve problems of classification and prediction efficiently due their automatic learning system. These are based in the statistic learning theory developed by [10], where a mathematical model is proposed for resolution of classification and regression problems [11]. Other authors mention that SVM is a margin classifier trained by a dataset based on feature vectors. SVM tries to find an optimal level between two different classes of feature vectors with a maximum margin (distance from the optimal hyperplane to the nearest vector). To make classification of the non-separable dataset, a linear feature of SVM it's a vector project in a space of high dimension using a kernel function, such as radial base kernel function [12].

The construction of support vector machines (SVM) is based on the idea of transforming or projecting a dataset belonging to a dimension n, to a superior dimension space applying a kernel function – Kernel Trick. From this created space, data will be handled as a linear problem, looking for a solution without considering the data dimensionality [13].

The success of support vector machines relies on three fundamental advantages: first is their solid mathematical foundation. Second is the concept of minimization structural risk [14, 15], which translates in minimizing the probability of error in classification of new examples. This case usually happens when there are few training data. The third advantage is based on the powerful tools and algorithms to find a solution fast and efficiently [16].

Naive Bayes :

Bayesian networks are considered an alternative to the classic expert systems oriented to decision making and prediction under uncertainty in probabilistic terms [17]. In [18, 19], is presented a structure of four levels. The superior level is composed by a set of variables represented by nodes and arrows related in terms of influence. In a lower level would be the levels or states, also known as space state [20, 21] that can take each of the variables of the model. In third place, you have a set of probability conditional functions, one for each state of the variable conditioned to the possible values of the variables that set the value of the variable. Finally, in the lowest level there would be a set of algorithms that would allow the network to recalculate the probabilities assigned to every level when some evidence of the model is known.

K- Nearest Neighbors :

It´s an algorithm used for classification and data regression. The algorithm stores all known cases and classifies or assign a feature to new cases based on similar features [22]. This method must be one of the first options when there is little to none previous knowledge of the data distribution. The method was developed for the need of performing discriminant analysis when reliable parametric estimates of probability density are unknown or difficult to calculate [23]. The main difficult of this method is to find the value of k, because if its value is too high, there is the risk of setting the classification based on the majority of values and if it's too low there can be a lack of precision in the classification caused by the shortage of selected data as comparison instances [24].

One of the advantages more relevant of the k-NN method is that can change radically the classification results without modifying its structure, changing the metric used to find the distance. The metric must be selected according to the problem that needs to be solved. The advantage of being able to change metrics, allows to get different results without changing the algorithm, only the procedure to measure distances.

EXPERIMENTATION

For this research the "Fertility Dataset" was used, which can be found in the Machine Learning Repository UCI [8]. The purpose of this research is to have a comparison of different data mining techniques, which are used usually in task of classification, prediction and segmentation. Next, we describe the methodology used.

First, dataset was downloaded and prepared to start training and evaluation of the method created, second, an analysis of data mining tools to be used was made, taking into account their availability in the Weka tool and their license terms of use. In the process of data preparation, it was necessary to run a balancing process for the data, because the output variable data was out of scale, and this situation has a negative impact on the proposed model, since the learning process must be

made under one situation only. You can see an screenshot of the initial state of the data and the results after the balancing process.

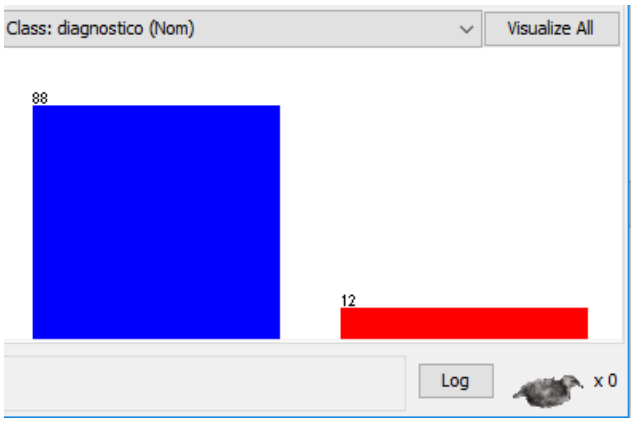


Figure 1: Dataset without balancing process
Source: *Created by author*

Figure 2: Dataset after balancing process
Source: *Created by author*

After data preparation and tool selection, the clustering method to be used was selected, given the need of joining users based on their fertility. For this process, it was used the K-Means Simple method, and the results are presented in the next table:

Table 2: Data Aggregation Results with Clustering Method

Cluster 0	96 (52%)
Cluster 1	88 (48%)

Source: *Created by author*

In table 2, you can see the amount of users in each cluster, where cluster 0 represents the patients with fertility problems and cluster 1, are the patients without this condition.

From segmentation or data clustering, the classification methods to be compared were selected, and they are: Decision Trees, Support Vector Machines, Naïve Bayes, and Lazy IBK, and the metrics to be evaluated are: True Positive Rate (TP Rate), False Positive Rate (FP Rate), Precision, Recall, calculated with the equations found in Table 3.

Table 3: Evaluation Metrics of Classification Methods

$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$ **(Equation No 2)**	$TpRate = \frac{True\ Positive}{TruePositive + False\ Negative}$ **(Equation No 4)**
$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$ **(Equation No 3)**	$FpRate = \frac{False\ Positive}{False\ Positive + True\ Negative}$ **(Equation No 5)**

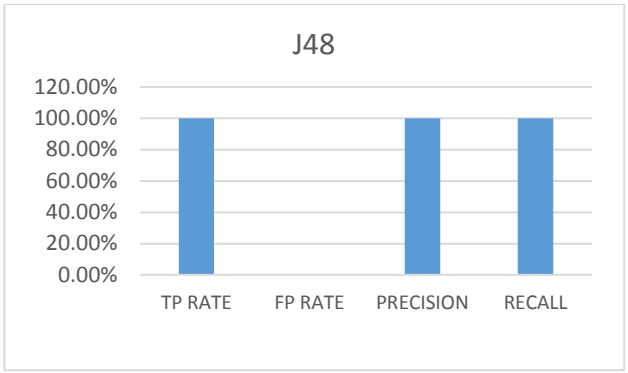
Source: *Created by author*

The data distribution for the training and test process was implemented through crossed validation, the tool selects a data percentage for training and other percentage for testing. After the training phase, each data mining method is applied with the metrics previously mentioned.
Finally, after identifying the best method, a test file was used to check the results for the best classifier obtained. The methodology of the research can be resumed in Figure 3.

Figure 3: Research Methodology Phases
Source: *Created by author*

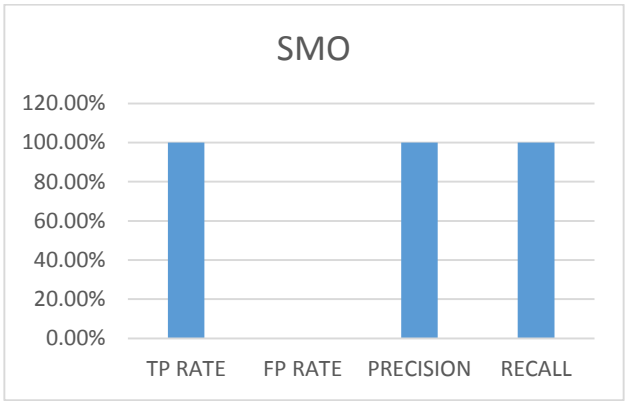
RESULTS

In this research, the starting point was the data preparation, continuing with choosing the best clustering technique, because the data segmentation will affect the results of the classifiers used. The next set of graphs reflect the results obtained in each phase of the methodology applied.



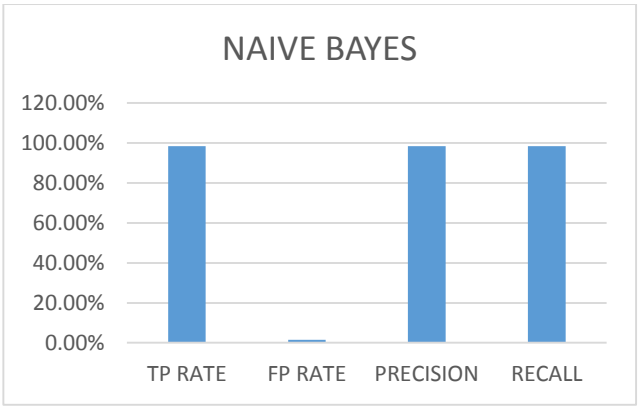
Graph 1: J48 Algorithms results
Source: *Created by author*

In graph 1, you can have the results obtained with the decision trees algorithm J48 which are: 100% (TpRate), 0% (FpRate), 100% (Precision), 100% (Recall).



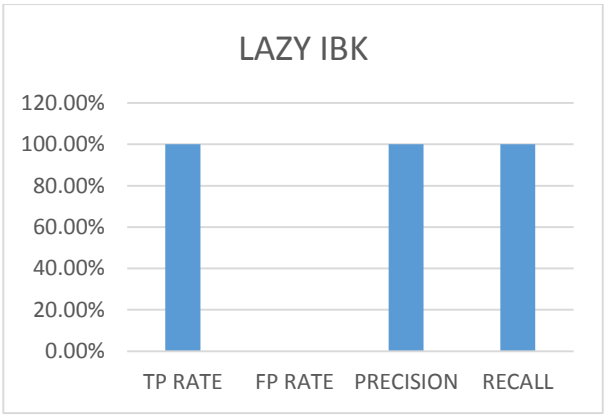
Graph 2: SMO Algorithm Results
Source: *Created by author*

In graph 2, you can have the results obtained with the support vector machine algorithm which are: 100% (TpRate), 0% (FpRate), 100% (Precision), 100% (Recall).



Graph 3: Naïve Bayes Algorithm Results
Source: *Created by author*

In graph 3, you can have the results obtained with the Naïve Bayes algorithm which are: 98.4% (TpRate), 1.5% (FpRate), 98.4% (Precision), 98.4% (Recall).



Graph 4: Lazy IBK Algorithm Results
Source: *Created by author*

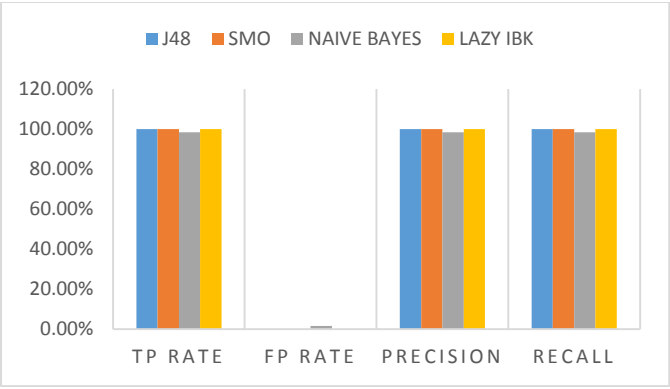
In graph 4, you can have the results obtained with the Lazy IBK algorithm which are: 100% (TpRate), 0% (FpRate), 100% (Precision), 100% (Recall).

The results found with all classifier methods are resumed in table 4.

Table 4: Classifier Algorithm Results

METHOD	TP RATE	FP RATE	PRECISION	RECALL
J48	100,00%	0,00%	100,00%	100,00%
SMO	100,00%	0,00%	100,00%	100,00%
NAIVE BAYES	98,40%	1,50%	98,40%	98,40%
LAZY IBK	100,00%	0,00%	100,00%	100,00%

Source: *Created by author*



Graph 5: Classifier Algorithm Results
Source: *Created by author*

In graph 5, you can find the comparison of all techniques used, and it's easy to notice the best methods achieve the best percentages in all metrics evaluated, which led us to conclude that they can be used in classification processes to assess the fertility conditions with the data valuated. The proposed method is based in the structure shown in Figure 4.

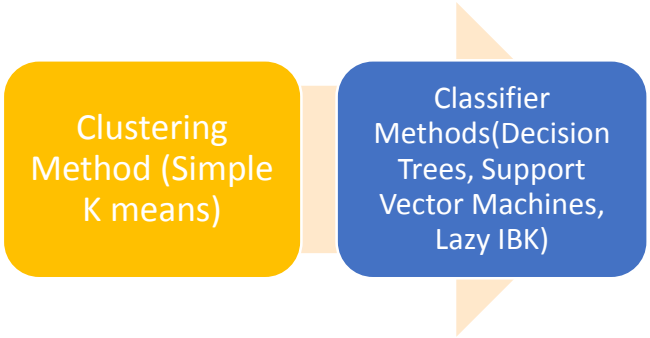


Figure 4: Methods Summary
Source: *Created by author*

CONCLUSION

The goal of this research is to propose a method to build classification processes to identify fertility conditions in men, based on the dataset Fertility Data. To accomplish this, it was necessary to make a segmentation process of the patients with or without fertility problems through the simple k-means method, and then the data mining techniques trees decision, support vector machine, Bayesian networks and k-nearest neighbors were applied and obtained that three methods had TpRate 100%, FpRate 0%, Precision 100%, Recall 100%, the fourth method, using Bayesian networks had TpRate 98.4%, FpRate 1.5%. Precision 98.4%, Recall 98.4%. These results clearly show these methods can achieve high percentages in all metrics, and confirms that the proposed method can be efficient and accurate to detect fertility rates in patients, improving the results accomplished by researchers in previous studies [5].

REFERENCES

[1] Villalobos, A. Centro de Especialidades Ginecológicas y Obstetricas. Recuperado de http://infertilidadcr.com/publicaciones/infertilidad-publi.html

[2] Brugo-Olmedo, S., Chillik, C., & Kopelman, S. (2003). Definición y causas de la infertilidad. Revista colombiana de Obstetricia y Ginecología, 54, 227-248.

[3] Auger, J., Kunstmann, J. M., Czyglik, F., & Jouannet, P. (1995). Decline in semen quality among fertile men in Paris during the past 20 years. New England Journal of Medicine, 332(5), 281-285.

[4] Carlsen, E., Giwercman, A., Keiding, N., & Skakkebæk, N. E. (1992). Evidence for decreasing quality of semen during past 50 years. Bmj, 305(6854), 609-613.

[5] Gil, D., Girela, J. L., De Juan, J., Gomez-Torres, M. J., & Johnsson, M. (2012). Predicting seminal quality with artificial intelligence methods. Expert Systems with Applications, 39(16), 12564-12573.

[6] Splingart, C., Frapsauce, C., Veau, S., Barthelemy, C., Royère, D., & Guérif, F. (2012). Semen variation in a population of fertile donors: evaluation in a French centre over a 34-year period. International journal of andrology, 35(3), 467-474.

[7] Bonde, J. P. E., Ernst, E., Jensen, T. K., Hjollund, N. H. I., Kolstad, H., Scheike, T., ... & Olsen, J. (1998). Relation between semen quality and fertility: a population-based study of 430 first-pregnancy planners. The Lancet, 352(9135), 1172-1177.

[8] Gil, D., Girela, J. (2016). Machine Learning Repository: Fertility Data Set. Recupedao de https://archive.ics.uci.edu/ml/datasets/Fertility

[9] Kyoungok, K. (2016). A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree. ELSEVIER, 157-163.

[10] Vapnik, V. N. (1998). Statistical Learning Theory. Nueva York: Wiley-usa

[11] Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Nueva York: Springer-Verlag.

[12] Bakhtiarizadeh, M. R. (2014). Neural network and SVM classifiers accurately predict lipid binding proteins, irrespective of sequence homology. ELSEVIER, 213-222.

[13] Gutierrez, M., & J, F. (2011). Pronóstico de incumplimiento de pago mediante máquinas de vectores de soporte.

[14] Kecman, V. (2001). Learning and Soft Computing. Londres: mit Press-uk.

[15] Cristianini, N. y Shawe-Taylor, J. (2000). An Introduction to Support Vector Machine and other kernel-based Learning Methods. Nueva York: Cambridge University Press.

[16] Jimenez, L. y Rengifo, P. (2010). Al interior de una máquina de soporte vectorial. Revista de ciencias, (14), 73-85.

[17] Cowell, R. G., Dawid, A. P., Lauritzen, S. L., & Spiegelhalter, D. J. (1999). Probabilistic networks and expert systems

[18] Edwards, W. (1998). Hailfinder: tools for and experiences with Bayesian normative modeling. American Psychologist, 53(4), 416

[19] Edwards, W., & Fasolo, B. (2001). Decision technology. Annual review of psychology, 52(1), 581-606.

[20] Nadkarni, S., & Shenoy, P. P. (2001). A Bayesian network approach to making inferences in causal maps. European Journal of Operational Research, 128(3), 479-498.

[21] Nadkarni, S., & Shenoy, P. P. (2004). A causal mapping approach to constructing Bayesian networks. Decision support systems, 38(2), 259-281

[22] Sánchez, A. S., Iglesias-Rodríguez, F. J., Fernández, P. R., & de Cos Juez, F. J. (2016). Applying the K-nearest neighbor technique to the classification of workers according to their risk of suffering musculoskeletal disorders.International Journal of Industrial Ergonomics, 52, 92-99

[23] Fix, E., & Hodges Jr, J. L. (1951). Discriminatory analysis-nonparametric discrimination: consistency properties. California Univ Berkeley.

[24] Mucherino, A., Papajorgji, P. J., & Pardalos, P. M. (2009). Data mining in agriculture (Vol. 34). Springer Science & Business Media.