Classification system for the predicting of psychosocial risk level in public-school teachers based on Artificial Intelligence

Rodolfo Mosquera Navarro,Ph.D(c)
Industrial Engineering Department, Faculty of Engineering
National University of Colombia
Manizales, Colombia
rmosqueran@unal.edu.co

Director: Omar Danilo Castrillon Gómez, Ph.D
Industrial Engineering Department, Faculty of Engineering
National University of Colombia
Manizales, Colombia
odcastrillong@unal.edu.co

Co-director: Liliana Parra Osorio, Ph.D
Psychosocial and Work Organization Department
Universidad Libre, Bogotá
Bogotá, Colombia
lilianaparraosorio@hotmail.com

Co-director: Andrés Carrión García, Ph.D
Director - Master Program in Data Analysis Engineering
Polytechnical University of Valencia
Valencia, Spain
acarrion@eio.upv.es

Abstract— **A new method to identify psychosocial risk level classification in Colombian public-school teachers, based on new hybrid approach on Artificial Intelligence. In this paper, we propose a first approach to hybrid prediction method using the algorithm hill climbing and Support Vector Machines (HC-SVM) to predict the psychosocial risk level on Colombian teachers public-schools in the metropolitan area of a Colombian city, where uses HC to optimize the input parameters of the prediction model based on SVM. A database with 5,340 epidemiological records that correspond to the psychosocial evaluations of public-school teachers in the metropolitan area of a Colombian city (five municipalities). In the best classification performance, general precision and stability will obtained with 95% effectiveness compared to the psychosocial clinical diagnosis. Three new physiological variables will be implemented to be associated with psychosocial risk levels (heart rate, skin electrodermal activity, and electromyography). To validate the proposed prediction method, a 480 individuals who were representative of the sample will randomly chosen and evaluated with three new physiological variables is selected as a case study. Use of the model as an instrument for prediction of psychosocial risk level in public-school teachers is suggested as a tool for occupational psychologists to improve psychosocial risk prevention programs within administrative and occupational safety systems.**

Keywords— Support vector machine; hill climbing; psychosocial risk level; management; work organization; public-school teachers.

I. Introduction

The International Labor Organization (ILO), defines stress as the physical and emotional response to an event caused by the lack of balance between the perceived demands and perceived resources and capabilities that an individual has to cope with said demands. Work-related stress is determined by the work relationships, job design, and work organization. It presents when work demands are not commensurate with, do not coincide with the company's organizational and functional culture expectations [1].

The problems, which have been reported in the workplace, as well as the occupational hazards that require the most urgent attention include: injury and accident prevention, followed by psychosocial risks, work-related stress and musculoskeletal disorders. Action priorities regarding workplace health and safety point clearly to the need to build capacity for and generate instruments, which permit epidemiological surveillance of psychosocial risk and work-related stress. Proposed solutions include a redefinition of tools that facilitate risk identification and control [2].

In Colombia, relating to psychosocial risk, the second national survey for workplace health and safety conditions in the Colombian general occupational risk system indicates that almost 12% of workers' cognitive performance and confrontation abilities are affected, which demonstrates the need for further research on development strategies for accident prevention, health promotion and illness prevention [3].

Certain studies carried out in human resource management areas, in fields where psychosocial risk appears in particular variables as a part of organizational management and the work environment, show the effectiveness of using support vector machines for detection of an employee's intention to quit, associating work motivation, job satisfaction, and stress levels as predictors [4]. They have also been useful in studies related to prediction of voluntary employee turnover associated with job performance [5].

The large number of variables utilized in the identification of psychosocial risk (predictors) and various variables (classes) makes important the recognition of the variables that impact on

We extend our sincere thanks to the "Convocatoria Nacional para el Apoyo al Desarrollo de Tesis de Posgrado o de Trabajos Finales de Especialidades en el área de la Salud de la Universidad Nacional de Colombia 2017-2018" via Resolution 21 of 2017 from the office of the Vice-rector of Investigation (21 of December 2017), for selecting this proposal for project financing with identification number 40976.

the risk on people and can be recognized through of the so-called methods for supervised pattern recognition (classification methods). The classification methods are used in different fields, including cellular biology usin SVM´s [6], image analysis using Support Vector Machines least squares (LS-SVM) [7], fault detection and diagnosis of an industrial steam turbine industrial process using support vector machine (SVM) and adaptive neuro-fuzzy inference system (ANFIS) [8], fault detection and diagnosis in process data using one-class support vector machines [9], and in various aspects of economics as time series prediction [10].

In this investigation, the HC-SVM model is used for psychosocial risk level prediction in public-school teachers. The rest of this article is organized as follows: Section 2 summarizes the prediction model, Section 3 addresses a numerical example taken from evaluations performed on public-school teachers from five Colombian municipalities, and finally, conclusions are presented in Section 4.

II Materials and methods

This section presents short descriptions of the models, which compose the new hybrid method, information from the database used, as well as an explanation of implementation of the model in the present study.

A. *Support Vector Machine (SVM) background algorithm*

Support Vector Machines (SVM), proposed by [11], are a supervised machine learning approach. Support vector machines have an advantage over other existing classification approaches, in that they provide a global solution for data classification [12][13]. They generate a unique global hyperplane in order to separate data points from different classes, instead of local limits, as compared to other existing data classification approaches. As support vector machines follow the principal of Structural Risk Minimization (SRM), they reduce the appearance of risk during the training phase [14].

Support vector machines attempt to find an optimum decision limit, or to maximize the margin, by finding the greatest achievable distance between the separating hyperplane and the data points either side [15]. Data classification may be represented through consideration of a group of ordered pair values $D= \{(X\_1,Y\_1),(X\_2,Y\_2),\ldots,(X\_l,Y\_l)\}$, where $i=1,\ldots,5340$. Here, $x \in X$ and $y \in Y$ where *'Y'* represents the class label (Risk level), for example, in the binary classification $Y= \{-1,+1\}$. If the classes are not linearly separable, owing to data out of range (data uncertainty, etc.), it is still possible to use the linear classifier with error tolerance. In this case, the objective is to find a balance between maximizing data accuracy and minimizing error.

B. *Hill Climbing*

The local random restart search is carried out through random selection of an initial solution (in other words, it is uniformly generated in the solution space), and a local deterministic search is applied (moving iteratively to the best neighbor of the current solution) until a local optimal is found.

This process is repeated until k local optimals are obtained. The best of these k local optimals are compared each time that an iteration is performed. The local search (LS) performs a macro iteration (restart), in which the neighbor function η is defined such that the solution space is attainable.

Hill climbing variables $R_k\ (w)i),w)=0$ *for all* $w(i) \in \Omega, w \in \eta\ (w(i))$ and the best neighbor solution are selected in each internal iteration. Once a $G \cup L$ element is found, a new Ω element is randomly generated (uniformly) in order to begin the next group of iterations (internal loop) [16].

C. *Data used*

The group of data comes from a database of information collected from evaluations performed on 5,340 public-school teachers in the metropolitan area of a Colombian city. Said information consists of sociodemographic data, organizational, environmental, physical, and psychosocial variables, which contain 123 predicting variables and one class variable, as shown in Table 1. The model were trained using the same group of training data, 80% training set X_{ij} and 20% validation set V_{ij}.

TABLE I. Variables for domains on intralaboral psychosocial risk factors. Adapted from: battery for evaluation of psychosocial risk factors. [17] annex 1.

Items	Domains	Variables
S_1	Sociodemographic	$S_1 \ldots S_{10}$
D_1	Work demands	$D_1 \ldots D_{50}$
W_1	Work control	$W_1 \ldots W_{21}$
L_1	Leadership and social relations in the workplace	$L_1 \ldots L_{32}$
R_1	Rewards	$R_1 \ldots R_{11}$

D. *Sample data with new physiological variables*

A random sample of 480 individuals, who decided to participate voluntarily, was extracted. New physiological variables were added (Hearth rate, skin electrodermal activity, and electromyography), which consisted of 130 variables, 129 of which were independent, and one which was dependent.

E. *Hierarchical cluster*

Redundant variables were filtered through use of a hierarchical cluster procedure, with the goal of grouping the 129 independent variables into different clusters, so as to identify redundant variables in the data sample using the Chebyshev and average methods for metrics and link criteria, respectively.

F. Principal component analysis

In order to carry out the analysis of principal components with this methodology, a method based on correlation matrices is used. This is employed when the data are not dimensionally homogeneous, or the order of magnitude of each random variable measured is not equal. It begins with the correlation matrix, considering that the value of each of the X random variables F_j is present.

G. Proposed hybrid algorithm

A general overview of the proposed hybrid methodology HC-SVM is presented in Fig.1.

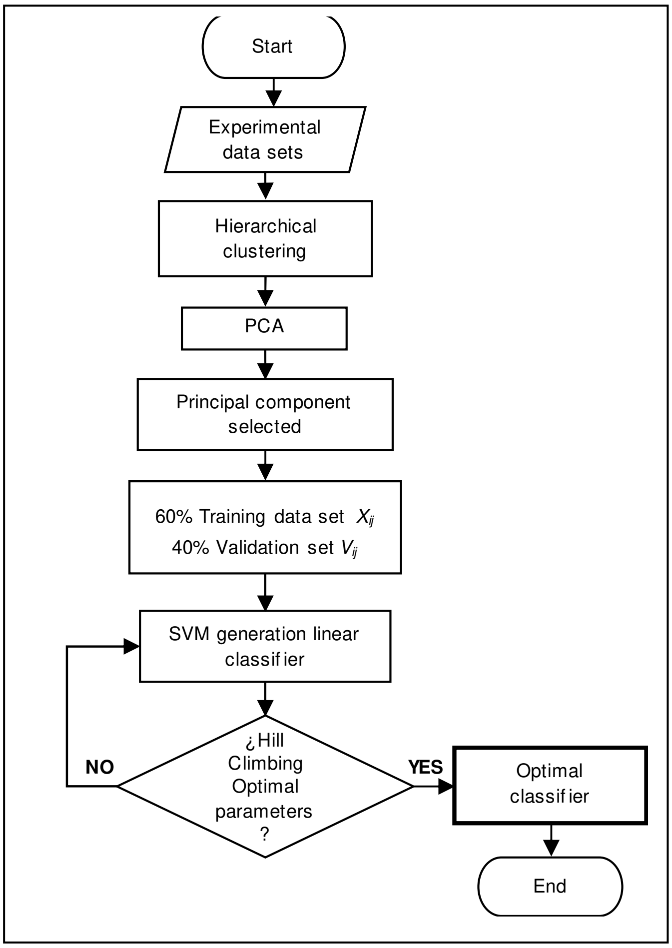

Fig. 1. General overview of the proposed hybrid model (HC-SVM)

III EXPERIMENTS

A. Datasets

There were 124 variables considered, with quantitative and qualitative values, and 5,340 sample registers which corresponded to the application of the psychosocial risk battery [17] on public school teachers in the metropolitan area of a Colombian city.

B. Sample data with new physiological variables

The sample data analyzed comes from a database of the information collected from evaluations performed on 5,340 teachers, from which a random sample of 480 individuals, who participated voluntarily, was extracted. With this information, a database with 480 registers was created. New physiological variables (P_{ij}) were added (Hearth rate (P_{1}), electrodermal skin activity(P_2), and electromyography(P_3). This addition incorporated a total of 130 variables, 129 of which were independent, and one of which was dependent. This is the class composed of four values (Very high risk (Y_4), high risk (Y_3), medium risk (Y_2), and low risk (Y_1)) as shown in Table 2.

TABLE II. INTRALABORAL PSYCHOSOCIAL RISK FACTOR DOMAIN VARIABLES AND PHYSIOLOGICAL VARIABLES. ADAPTED FROM: BATTERY FOR EVALUATION OF PSYCHOSOCIAL RISK FACTORS. [17] ANNEX 1.

People evaluated	**Psychosocial risk Var**					**Pshysio logical Var**
	***S* Domain**	***D* Domain**	***W* Domain**	***L* Domain**	***R* Domain**	***P* Domain**
X_1	X_1S_1… X_1S_{10}	X_1D_1… X_1D_{50}	X_1W_1.. X_1W_{21}	X_1L_1… X_1L_{32}	X_1R_1… X_1R_{11}	X_1P_1… X_1P_3
X_2	.	.	.	.	.	.
X_{480}	$X_{480}S_1$.. $X_{480}S_{10}$	$X_{480}D_1$ $X_{480}D_{50}$	$X_{480}W_1$…$X_{480}W_{21}$	$X_{480}L_1$ …$X_{480}L_{32}$	$X_{480}R_1$ …$X_{480}R_{11}$	$X_{480}P_1$.. $X_{480}P_3$

C. Hierarchical cluster

The goal, when using the hierarchical cluster procedure, is to group the 129 independent variables into different clusters. A representative variable was chosen for each cluster, for a total of 30 representative variables to represent the group of 129.

Thus, variables belonging to each cluster were identified and grouped into each one of the variable group positions, which correspond to clusters. Next, the representative variable of each cluster was chosen. If, in a cluster, there was more than one variable, only one of them needed be selected. In order to make this selection, the distance between pairs of variables in said clusters were measured a new, and the variable with the lowest sum of distances between other variables was chosen, for a total selection of 30 variables from the entire group, as shown in Table 3.

TABLE III. VARIABLES SELECTED USING HIERARCHICAL CLUSTER.

Variables					
S_2	S_6	S_7	S_9	S_{10}	S_{12}
D_{12}	D_{26}	D_{30}	D_{57}	D_{58}	D_{60}
D_{61}	D_{62}	W_{63}	W_{75}	W_{76}	W_{89}
W_{93}	W_{94}	W_{95}	W_{96}	W_{100}	L_1
L_2	L_3	R_1	P_1	P_2	P_3

D. Principal component analysis

The analysis of principal components is a technique employed to reduce dimensionality in a group of data. In this case, each of the 480 individuals was initially represented by 129 points. With the redundant variable filter, each individual was represented by 30 points, using PCA, each person may be represented by up to two variables, thus facilitating their graphic representation. In accordance with the information obtained, the first component explains 99.21% of information variability, and the second 0.41%. With these two components, 99.62% of all variability may be explained. The third component, which corresponds to 0.14% may also be used, but the first two were deemed sufficient, see Fig. 2.

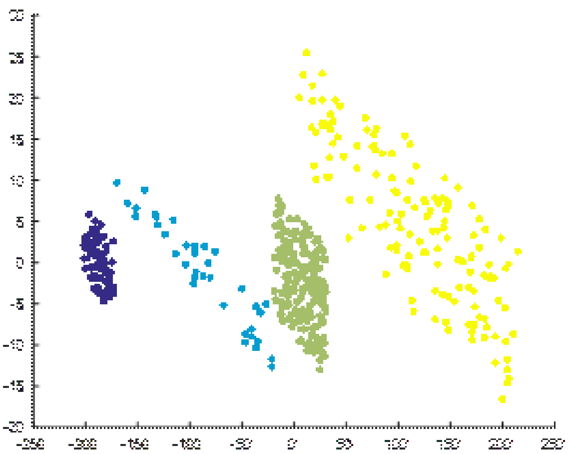

Fig. 2. Principal Component analysis

In the above diagram, each point corresponds to one of the 480 individuals, and each color represents a dependent variable value (risk). Note that separate groups were generated for each risk value.

E. Hybrid algorithm proposed

In support vector machines, finding a separating hyperplane for two groups of data is an optimization problem. The diagram above shows that all of the groups are linearly separable from each other.

The equation for the sought-after separating hyperplane has the following form:

$$w^T x = 0 \quad (1)$$

Where vector w is a vector that contains hyperplane parameters, and x is a vector which contains the two selected characteristics, after having chosen the principal components.

$$w = w_{0,} w_{1,} w_{3;}\ x = 1,\ x_1 \text{ and } x_2 \quad (2)$$

Where w_0, w_1 y w_2 are the hyperplane components, and x_1 and x_2 are the first and second components, respectively.

If the second component is cleared, the equation below is obtained:

$$x_2 = mx_1 + b \quad (3)$$

This is an equation for a straight line. In other words, the separating hyperplanes will be straight lines. In order to apply the hill climbing to this optimization problem, one must consider the minimization of vector w, which is a normal vector for the separator hyperplane.

Initially, there is separation of just one group from the other three, and later, there is separation of the other groups from each other. In order to separate group one from the others, it is only necessary to separate group one from group two. As such, only these two will be employed. Each pair (x_1, x_2) associated with the independent variable with a class or risk equal to one will be labeled with a value of one, while other pairs associated with a class or risk equal to two will be labeled with a value of -1.

F. Generation of a random individual

The first step in random reset hill climbing is the generation of a random solution, which is called an individual. In order to control the quality of each individual generated, the error variable is used.

These values are random, as it is unknown whether they truly complete the required function—it is merely a starting point. The error is calculated as the number of wrongly classified data, using vector w. In order for a data to be correctly classified, it must comply with the following condition:

Datum i, from the previously defined X group, $X_{(i)}$=[1 $x_{1(i)}$ $x_{2(i)}$] and its label $Y_{(i)} = y_i$, where $x_{1(i)}$ and $x_{2(i)}$ are the first and second components of datum i respectively, and y_i is its label. It is correctly labeled if:

$$Y_{(i)}[w^T X_{(i)}] - 1 \geq 0 \quad (4)$$

Recall that X and Y are vectors within a matrix, and so, firstly, one of their vectors must be accessed to find the information. However, this step is omitted here for ease of reading.

G. *Cycle of number of iterations*

Initially, only the algorithm for hill climbing is explained. The local search is performed with a number of k-local iterations. This value begins at one, and continues to 10,000, by experience.

H. *Generation of a new individual and its new associated error*

The following corresponds to the creation of a new individual and calculation of its associated error. As mentioned previously, when using the hill climbing algorithm, after generation of the initial random solution, an attempt is made to find a better solution, through variation of a single element of the solution. In this case, the new individual is generated by making a random change in the present individual, and the new individual's error is calculated with the wrongly classified function, as explained above.

Initially, a new individual is created, which is equal to the previous one. This is done in order to make changes to a new variable in the individual, without losing its information.

I. *Comparison of error1 and error2*

Following implementation of the hill climbing algorithm, if, after making the change, a better solution is produced, another change is made to the new solution. This process is repeated until improvements can no longer be made. The production of a better solution is confirmed through error: a better solution has a lower error than its individual predecessor.

J. *Implementation*

Application of the new algorithm to the group of experimental data from the sample group of public school teachers, see Table 4.

TABLE IV. RESULTS FROM HC-SVM CLASSIFICACTION ALGORTIHM.

% Training	% Validation	HC-SVM/Err	SVM/Err	ANN/Err
20	80	0.90/0.01	0.89/0.11	0.87/0.13
50	50	0.92/0.08	0.91/0.09	0.89/0.11
80	20	0.95/0.05	0.92/0.08	0.90/0.10

The results show that, in this case, the algorithm improves the classification percentages and rate of error as it improves the model training data.

K. *Evaluation of the data group*

There were 100 replications performed with each algorithm, and the best 5 were selected to be exhibited. The obtained results are shown in Table 5.

TABLE V. REPLICATIONS WITH EACH ALGORITHM AND THEIR GENERAL AVERAGE.

Algrthms	Replications					Simple Average
HC-SVM/ Error	0.949/ 0.051	0.948/ 0.052	0.948/ 0.052	0.949/ 0.051	0.948/ 0.052	0.9484/ 0.0516
SVM RBF/Error	0.892/ 0.108	0.893/ 0.107	0.892/ 0.108	0.893/ 0.107	0.893/ 0.107	0,8926/ 0.1074
ANN/ Error	0,903/ 0.097	0,903/ 0.097	0,903/ 0.097	0.901/ 0.099	0.901/ 0.099	0.9022/ 0.0978

IV DISCUSSION

In [4] that investigation attempts to examine the feasibility of SVMs in predicting employee turnover. Besides, two other tradition regression models, Logistic and Probability models are used to compare the prediction accuracy with the SVM model. Subsequently, a numerical example of employee voluntary turnover data from a middle motor marketing enterprise in central Taiwan is used to compare the performance of three models. Empirical results reveal that the SVM model outperforms the logit and probit models in predicting the employee turnover based on job performance. Consequently, the SVM model is a promising alternative for predicting employee turnover in human resource management.

In [5] the project developed a Support Vector Machine for predicting nurses' intention to quit, using working motivation, job satisfaction, and stress levels as predictors. This study was conducted in three hospitals located in southern Taiwan. The target population was all nurses (389 valid cases). For cross-validation, we randomly split cases into four groups of approximately equal sizes, and performed four training runs. After the training, the average percentage of misclassification on the training data was 0.86, while that on the testing data was 10.8, resulting in predictions with 89.2% accuracy. This Support Vector Machine can predict nurses' intention to quit, without asking these nurses whether they have an intention to quit.

In Colombia, attempts are currently being made to design and implement an AI algorithm, based on a group of data from evaluations which were previously administered specifically in a population from the educational sector schools [18]. Previous studies compare the classification effectiveness of models such as artificial neural network algorithms, decision trees, and Naive Bayes, and uses genetic algorithms to reduce dimensionality and to improve precision levels, with values of

93% [19] and in [20] compare Support Vector Machines, Naïve Bayes Classifier and Genetic Algorithms for predicting psychosocial risk levels in Colombian teachers school.

In this research, a new approach will adopted, based on the support vector machine algorithm and hill climbing, called HC-SVM, in order to predict psychosocial risk levels in public school teachers in the metropolitan area of a Colombian city.

This hybrid is promising for psychosocial risk analysis in Colombian public school teachers. It is also necessary to implement processes which allow for collection of data groups by geographical region, given that environmental, cultural, and social conditions affect classification results.

V CONCLUSION

In this investigation, the principal objective was to develop a new hybrid classification method which would allow for evaluation and prediction of psychosocial risk levels, through consideration of existing epidemiological data, thus reducing the dimensionality of existing instruments and improving the degree of prediction precision through reduction of redundant variables and increase the accuracy.

Its second contribution was the application of an optimization method which allowed for discovery of the optimal values within the support vector function, so as to classify all existing data in the search space and achieve very high classification values. This space was composed of values, and the algorithm modified itself constantly in order to reduce the risk of false alarms and new missed classes.

The process for evaluation of controlled characteristics proposed a multiple support vector machine with random restart hill climbing optimization, called HC-SVM. The empirical evaluation of the modified algorithm is better than that of the SVM algorithm.

The error rate of the modified algorithm decreases and the precision rate increases, in contrast to the SVM algorithm.

For future investigations, in the field of artificial intelligence, identification of psychosocial risk associated with facial recognition, supported by registered data, is necessary, as is the development of new algorithms which would allow for the association of physiological variables with specific psychosocial aspects for the prevention of occupational illnesses. The development of algorithms which identify musculoskeletal problems associated with psychosocial risks for illness prevention are similarly urgent, as well as the development of algorithms which classify lean manufacturing variables with psychosocial risks, so as to predict the development of occupational illnesses.

REFERENCES

[1] International Labour Office, & Labour Administration, L. I. and O. S. and H. B. Workplace stress: a collective challenge. Geneva: ILO, 2016.

[2] Kortum, E., Leka, S., & Cox, T. Psychosocial risks and work-related stress in developing countries: health impact, priorities, barriers and solutions. International Journal of Occupational Medicine and Environmental Health, 23(3), 2010. https://doi.org/10.2478/v10001-010-0024-5

[3] Ministerio del Trabajo, & Organización Iberoamericana de Seguridad Social-OISS. II encuesta nacional de condiciones de seguridad y salud en el trabajo en el sistema general de riesgos laborales. 2da ed., Vol. 1. Bogotá, 2015.

[4] Hong, W.-C., Pai, P.-F., Huang, Y.-Y., & Yang, S.-L. Application of Support Vector Machines in Predicting Employee Turnover Based on Job Performance. En L. Wang, K. Chen, & Y. S. Ong (Eds.), Advances in Natural Computation, Vol. 3610, pp. 668–674, 2005. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/11539087_85

[5] Tzeng, H.-M., Hsieh, J.-G., & Lin, Y.-L. Predicting Nurses' Intention to Quit With a Support Vector Machine. CIN: Computers, Informatics, Nursing, 22(4), 232–242, 2004.

[6] Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., ... & Haussler, D. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences, 97(1), 262-267. 2000.

[7] Molina, V., Vera, M., Robles, H. V., Bejarano, E., & Dávila, H. Lung segmentation using Support Vector Machine in computed tomography images. In Health Care Exchanges (PAHCE), 2014 Pan American (pp. 1-5). IEEE. April 2014.

[8] Salahshoor, K., Kordestani, M., & Khoshro, M. S. Fault detection and diagnosis of an industrial steam turbine using fusion of SVM (support vector machine) and ANFIS (adaptive neuro-fuzzy inference system) classifiers. Energy, 35(12), 5472-5482, 2010.

[9] Mahadevan, S., & Shah, S. L. Fault detection and diagnosis in process data using one-class support vector machines. Journal of process control, 19(10), 1627-1639, 2009.

[10] Sapankevych, N. I., & Sankar, R. Time series prediction using support vector machines: a survey. IEEE Computational Intelligence Magazine, 4(2), 2009.

[11] Vapnik, V. The Nature of Statistical Learning Theory. Ed: Springer-Verlag, New York, 1995.

[12] Cortes, C., Vapnik, V. Support-vector network, Mach. Learn. 20, pp 273–297, 1995.

[13] Vapnik, V. The Nature of Statistical Learning, second ed., Springer, New York,1998.

[14] Cristianini, NJ., Shawe-Taylor. An Introduction to Support Vector Machines, Cambridge University Press, 2000.

[15] Kotsiantis, S.B., Zaharakis, I.D., Pintelas, P.E., Machine learning: a review of classification and combining techniques. Artif. Intell. Rev. 26, 159–190. 2006.

[16] Jacobson, S. H., & Yücesan, E. Analyzing the performance of generalized hill climbing algorithms. Journal of Heuristics, 10(4), 387-405, 2004..

[17] Villalobos, G.; Vargas, A. M.; Escobar, J.; Jiménez, M. L. y Rondón, M. A. Batería de instrumentos para la evaluación de factores de riesgo psicosocial. Bogotá: Ministerio de la Protección Social, 2010.

[18] Mosquera, Rodolfo, Parra-Osorio, L., & Castrillón, O. D. Methodology for Predicting the Psychosocial Risk Level on Colombian Teachers using Data Mining Techniques. Inf. Tecnol, 27(6), 259–272. 2016. https://doi.org/10.4067/S0718-07642016000600026

[19] Mosquera, Rodolfo, Parra-Osorio, L., & Castrillón, O. D. Prediction of Psychosocial Risks in Colombian Teachers Public Schools Using Machine Learning Techniques. Inf. Tecnol, 29(4), 267-280, 2018. http://dx.doi.org/10.4067/S0718-07642018000400267

[20] Mosquera, Rodolfo, Parra-Osorio, L., & Castrillón, O. D. Support Vector Machines, Naïve Bayes Classifier and Genetic Algorithms for the Prediction of Psychosocial Risks in Teachers of Colombian Public Schools. Inf. Tecnol, 29(6)., in press.