Vol. 22, n.°1, pp. 17-24, 2023
Revista UIS Ingenierías
Página de la revista: https://revistas.uis.edu.co/index.php/revistauisingenierias

Pattern recognition for the modification of characteristics using non-linear techniques
Reconocimiento de patrones para la estimación de características por medio de técnicas no lineales

Fernando Mesa [1a], Rogelio Ospina-Ospina [2], Diana Marcela Devia-Narváez [1b]

[1] Matemáticas aplicadas y educación (GIMAE), Departamento de Matemáticas, Universidad Tecnológica de Pereira, Colombia. Orcid: 0000-0002-3418-5555 [a], 0000-0002-0447-4663 [b]. Emails: femesa@utp.edu.co [a], dmdevian@utp.edu.co [b]

[2] Ciencia de Materiales Biológicos y Semiconductores (CIMBIOS), Escuela de Física, Universidad Industrial de Santander, Colombia. Orcid: 0000-0002-7392-8059. Email: rospinao@uis.edu.co

Received: 20 June 2022. Accepted: 6 November 2022. Final version: 29 December 2022.

Abstract

Traditional data processing applications are unsuitable for handling large amounts of data. To achieve an efficient manipulation and extraction of characteristics or samples that the information represents, it is necessary to know aspects such as data collection and treatment. In this document, a database corresponding to the behavior of electrical energy consumption in a residential load was refined. The debugging and statistical analysis of the samples were carried out using the principal component analysis. The training of the smallest data set to the original database was made using vector support machine techniques and artificial neural networks. Finally, a proposal is presented for the analysis of samples that are within the operating limits or not using updating dynamic patterns for the unsupervised validation of new samples.

Keywords: assertiveness; database; characteristics; estimation; data cleansing; vector support machine; samples; artificial neural networks; validation; treatment.

Resumen

Las aplicaciones de procesamiento de datos tradicionales son inadecuadas para el manejo de una cantidad elevada de datos. Para lograr una eficiente manipulación y extracción de características o muestras que representa la información, es necesario conocer aspectos como la captación y tratamiento de datos. En este documento se depuró una base de datos correspondiente al comportamiento del consumo de energía eléctrica en una carga residencial. La depuración y análisis estadístico de las muestras se realizó por medio del análisis de componentes principales. El entrenamiento del conjunto de datos de menor dimensión a la base de datos original se hizo por medio de las técnicas de máquina de soporte vectorial y redes neuronales artificiales. Finalmente, se presenta una propuesta de análisis de muestras que se encuentren o no dentro de los límites de operación por medio de la actualización de patrones dinámicos para la validación no supervisada de nuevas muestras.

Palabras clave: asertividad; base de datos; características; estimación; depuración de datos; máquina de soporte vectorial; muestras; redes neuronales artificiales; validación; tratamiento.

ISSN Printed: 1657 - 4583, ISSN Online: 2145 - 8456.

How to cite: F. Mesa, R. Ospina-Ospina, D. Marcela Devia-Narváez, "Pattern recognition for the modification of characteristics using non-linear techniques," *Rev. UIS Ing.*, vol. 22, no. 1, pp. 17-24, 2023, doi: https://doi.org/10.18273/revuin.v22n1-2023002

1. Introduction

The term Big Data is frequently used to denote data sets that contain complex and numerous samples from a series of observations of a phenomenon. It often refers to the use of advanced metrology to collect, manipulate and extract values from a collection of data. Traditional data processing applications are inadequate for handling large amounts of data [1]. To achieve efficient capturing and extraction of characteristics or samples that represent a database, it is necessary to know aspects such as information manipulation that includes capturing, searching, storing, transferring, visualizing, analyzing, debugging, exchanging, updating, and privacy of information. The list of challenges is extensive and includes the characterization of uncertainty in complex databases and the extraction of information from large volumes of data, i.e., the origin of the data makes the information not easy to obtain due to the volume, noise, heterogeneity, collection on different temporal and spatial scales. For this reason, the deduction of information from a database becomes a complex task and can be useful to optimize performance and improve the design of classification models or create predictive models [2].

On the other hand, samples generate random information from an unknown probability distribution. Understanding the behavior of the data means deducing a probabilistic model from the information that allows knowing the relationship between the variables of a model. The accuracy of the model through Big Data can lead to safer and more reliable decision-making that results in greater operational efficiency of an industrial process, reduced operating costs, and mitigation of capital investment risk [3]. The importance of using information collection sources such as social networks, online transactions, and video devices to use the underlying information of patterns that allow improving processes in companies that are interested in research in the field of Big Data with the help of the government [4].

With the developed application and the increasing trend of diagnostic image analysis in Latin America, it was possible to improve the extraction of pathology information from a large volume of patient images but focused on visual computing and image analysis. The Support Vector Machines technique consists of the detection and recognition of traffic signs from image sequences. To obtain a traffic sign recognition system, samples of these signals were collected in various paths made by an automaton. The classification of colors and signs was achieved with the use of the algorithm of linear vector support machines. A non-linear multi-classification support vector machine model was used for the pigmentation segmentation of the signals [5].

Finally, the optimal fragmentation, sub-clustering, and feature extraction of a set of syndromes are considered. Statistically, the relevance of each example pattern is verified, and with the weighting of the subregions, each new sample that enters the analysis is classified in such a way that the utility of each class does not interfere with the other classes obtained in the classification [6].

In this document, a database that relates the information of power consumption measurements in a load was analyzed. The database contains 2075259 samples recollected from December 2006 to November 2010, forming a total of 9 base characteristics such as date, sample time, active power (kW), reactive power (kVAr), voltage (V), current intensity (A), energy sub-measurement (kitchen), energy sub-measurement (lighting circuit) and energy sub-measurement (heater, air conditioner).

2. Content

2.1. Database debugging

For the collected data [7] it was necessary to normalize the measurements according to each type, i.e., to obtain a range for each characteristic. After knowing the ranges of variation, we proceed to extract "hidden" information from the samples by calculating intricately related characteristics, e.g., the variations of each type concerning previous samples. After knowing the debugged database, it is necessary to perform the relevance analysis [8], [9]. This allows for discarding the characteristics that do not provide relevant information to the regression task, as shown in Figure 1.

Equation 1 defines the closeness of the samples:

$$\min_{i \neq j \;\; feature\; k} \left\{ \left(Sample_{ik} - Sample_{jk} \right)^2 \right\} \tag{1}$$

For this practical case, the proximity between samples was delimited with a value of $\Delta = 0.7$, i.e., if the operation of (1) is greater than the parameter Δ, it is considered that $sample_{jk}$ is far from $sample_{ik}$. On the contrary, if the operation is less than the parameter Δ, the sample $sample_{jk}$ is close to sample $sample_{ik}$ and will form a group of patterns with similar characteristics, thus reducing the number of samples in the database.

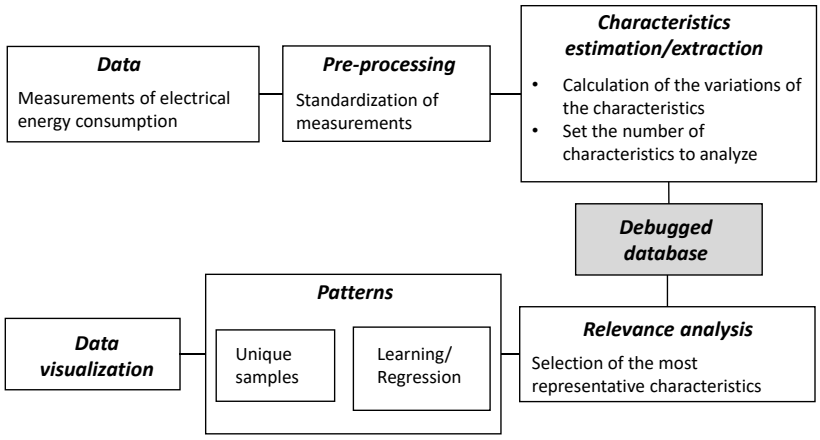

Figure 1. Pre-processing sequence and characteristic debugging.

After obtaining a subset that represents all the original samples, a supervised classification model is trained with Support Vector Machines and neural network training independently with the probabilistic model equations (2) and (3), respectively.

$$P(x_i, x_j) = e^{\left(\frac{(x_i - x_j)^2}{2\sigma^2}\right)} \quad (2)$$

$$P(\mathbb{X}, \mathbb{W}) = \mathbb{W}\left(\left(\frac{2}{1 + e^{(-2(\mathbb{X}+b1))}}\right) - 1\right) + b2 \quad (3)$$

Equations (2) and (3) allow modeling the entire set of samples with the Gaussian white noise distribution, i.e., the noise allows to separate the samples in such a way that the probability of the samples can be known by knowing the probability function.

The cost function (4) was used for the regression model through a Support Vector Machine (SVM) [10]:

$$\min_{\lambda} \frac{1}{2}||\mathbb{W}||^2 + \sum_{i=1}^{p} \lambda \langle x_i | y_i \rangle \quad (4)$$

To quantify the success by validation, equation (5) was implemented [10]:

$$ACC = \frac{1}{P}\sum_{i=1}^{P}(y_{(px1)_i} - \hat{y}_i) \quad (5)$$

Where:

$\mathbb{W}$ is the Gaussian white noise matrix used to separate the samples and disperse the information. From this matrix, the first two statistical moments, mean and standard deviation, are known. The matrix $\mathbb{W}$ has the same dimensions as the database.

$\mathbb{X}$ represents the database of p samples (rows) and n characteristics (columns).

σ is the standard deviation of the database.b_1, b_2, and λ are constants that vary in the range of 0.1 to 0.4 to make the model proposed by (1), (2), and (3) stricter, i.e., $\lambda \downarrow$ corresponds to the highest level of rigidity of the model and when $\lambda \uparrow$ the level of rigidity of the model decreases for the regression task.

$\langle x_i | y_i \rangle$ is defined as the inner product between each of the samples (i-th sample with n features) multiplied by the i-th output.

$y_{(px1)_i}$ are the original outputs with which the training of samples of size p samples per 1 column is supervised, from which the i -th sample is extracted.

$\hat{y}_i$ is the i -th estimated sample that is compared to the total number of hits between the actual and calculated outputs.

2.2. Relevant samples supervised regression

The pattern recognition task for classification and regression consists of the process of modeling a set of samples, which are considered input samples when they generate a response (output) based on their latent characteristics. These characteristics can be obtained through Principal Components Analysis (PCA) which allows displaying a set that represents the database (database cleanup, samples, and features). That is, going from a database data from a database of complex data (high number of samples and features) to a base of lower dimension.

The regression task is divided into two stages: supervised regression training and regression validation (unsupervised analysis) [11].

For supervised regression, various linear, quadratic, near neighbor models, support vector machines, or neural networks are used to train with the refined data, in such a way that a response can be obtained from the samples that are considered inputs. On the other hand, unsupervised regression validations test with new input samples how the output behaves in real scenarios, hoping that the models fit the observed phenomena [12]. The output of the proposed model is as f $f(\mathbb{X}) = \hat{y} = \sum_{i=1}^{p} \lambda \langle \mathbb{X} | x_i \rangle$ (5), where the values found for λ are multiplied by (3), and the internal product of the database is defined for each of the uncorrelated samples.

For the problem of determining the characteristics to be analyzed, a set of diagnostic images of tobacco membranes was processed to reveal the oxidation chain, resulting in the study of lipids as input samples to the anti-oxidation model [13]. On the other hand, the patterns or characteristics that are considered as inputs were calculated with the combination, weighting, and grouping of samples of electrical energy consumption to predict the dynamics of transmission management [14].

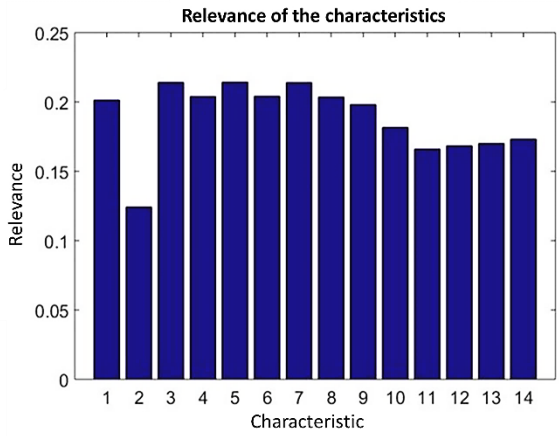

Figure 2. Most relevant characteristics 3, 5, and 7.

In this document, a database of 2,075,259 samples (rows) and 9 characteristics was processed, which were later increased by 5 new characteristics that come from the calculation of the average between samples, i.e., 14 characteristics to be refined. The analysis of principal components is carried out with the graph in Figure 2, where the most relevant characteristics are shown. 3, 5, and 7 will be taken as the most relevant characteristics.

The output to be analyzed is characteristic 6 associated with the voltage. Of the 14 labels, the most relevant of them is around 0.2143 attached to characteristic number 7. The minimum value that represents less interference is 0.1242, which corresponds to label number 2, the average is the sum of all the labels, and the value average corresponds to 0.1884.

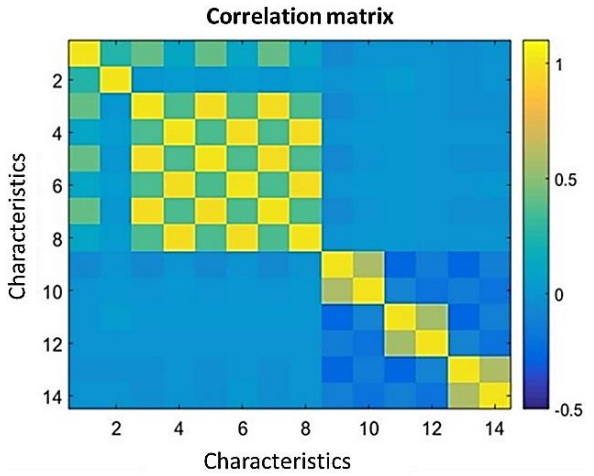

Figure 3. The correlation matrix contains off its diagonal the correlation coefficients between each of the samples.

The correlation matrix quantifies the relationship between pairs of characteristics (Figure 3). This matrix is square and symmetric. The diagonal has one as the maximum correlation of each characteristic with itself, and off the diagonal are the correlation coefficients between the variables.

The distance matrix presented in Figure 4 shows the relationship of the internal product of the database samples obtained by calculating: $\Sigma_{pxp} = \mathbb{X}^t \mathbb{X}$ [15].

Figure 5 shows the unsupervised approach found through principal component analysis, where 3 characteristics are presented in one dimension with no dependency between them. The accumulated variance must be calculated by entering each of the characteristics of the internal product defined as $var(\mathbb{Y}_i) = \mathbb{Y}_i \mathbb{Y}_i^t$ [16]. As characteristics are entered into the calculation, the highest cumulative variance is selected for the number of characteristics entered (Figure 6). The blue normalized output represents the voltage variation in a range from 1 to -1. The validation shows this same output (black color estimation) by the vector support machine technique (Figure 7) and by the artificial neural network technique (Figure 8).

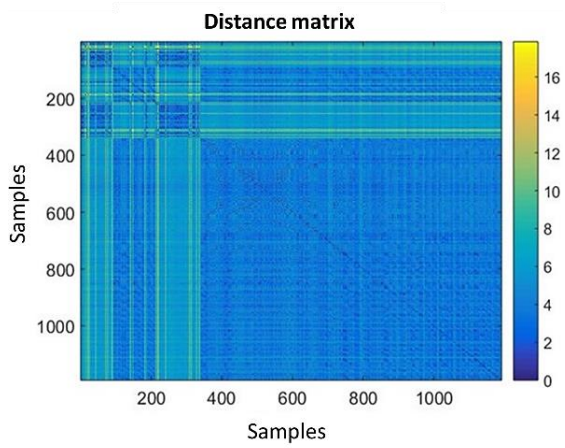

Figure 4. Sample separation matrix.

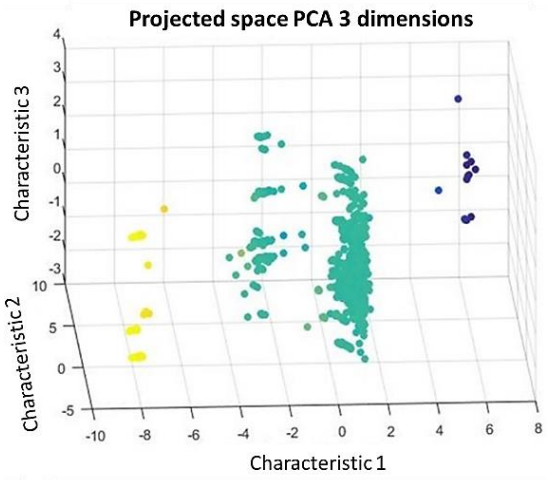

Figure 5. Representation of the samples in three dimensions.

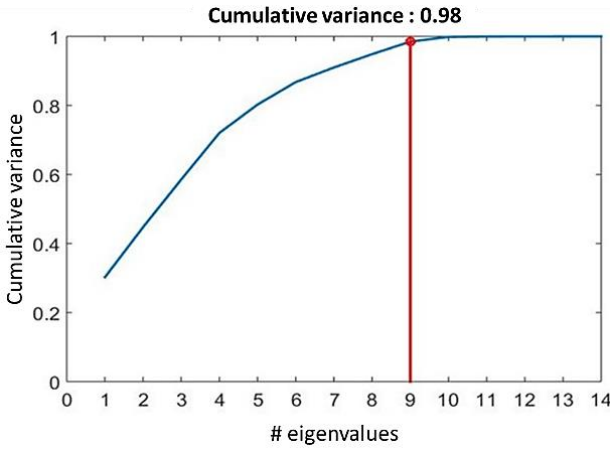

Figure 6. The cumulative variance of characteristics.

3. Analysis and results

From the database described above, a set of 1000 samples were randomly selected to perform the cross-validation or unsupervised process, i.e., show the model new inputs and estimate the output so that they can be compared and find the level of assertiveness through (4).

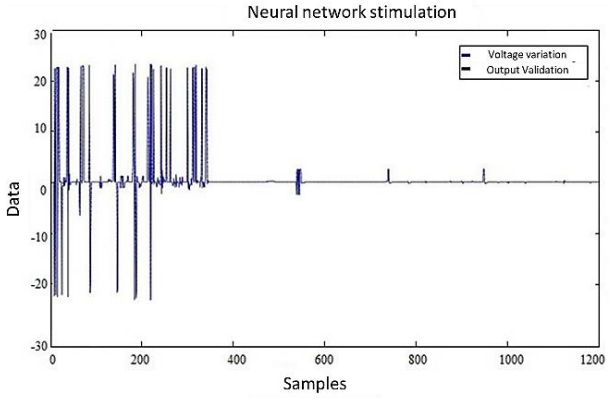

Figure 7. Supervised validation of normalized voltage output by Vector Support Machine.

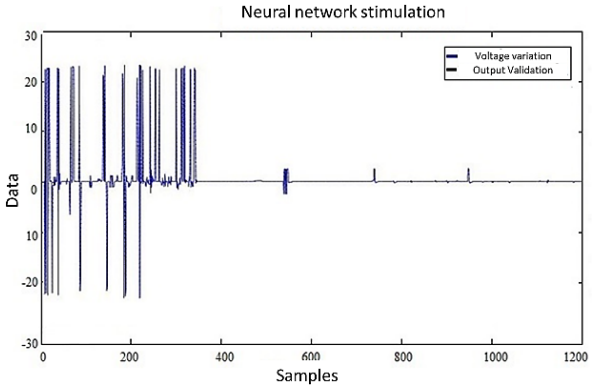

Figure 8. Supervised validation of the normalized voltage output using Artificial Neural Networks.

Figure 9 shows that in the face of a variation in the input of the model obtained by the vector support machine technique, the estimated output in blue tries to resemble it in the first 100 samples. Then, it tries to overlap the mean of the actual output (black line), obtaining an estimated output with a level corresponding to the mean of the data.

With the technique of artificial neural networks, it was possible to estimate a high assertiveness in the dynamics of the output of characteristic 7, as shown in Figure 10.

The estimation shown in Figure 11 proceeded to predict the output by the support vector machine technique with a set of 1200 new samples that were never considered in the supervised training, i.e., it predicted the new input outputs to the support vector machine model.

Finally, in Figure 12, a proposal is made for the quantification of losses or measurements that are outside the permitted limits of operation for the database studied in this document. That is, once the output prediction model uses SVM to artificial neural networks, it can be known in real-time if a measurement varies within limits established by the flowchart to know the dynamics of the behavior of electrical energy consumption in a load determined.

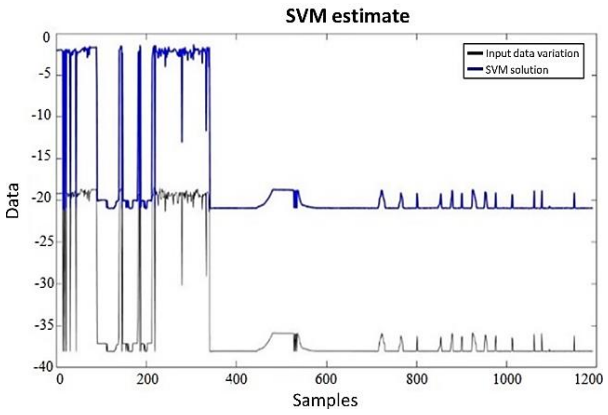

Figure 9. Feature estimation (Normalized Output 6) Actual output (black line), estimated output (blue line) via SVM.

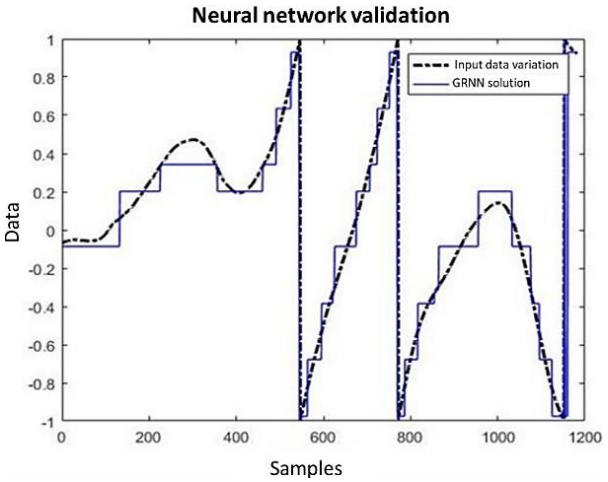

Figure 10. Estimation of the characteristic (Output 7 normalized) Real output (black line), estimated output (blue line) using Artificial Neural Network.

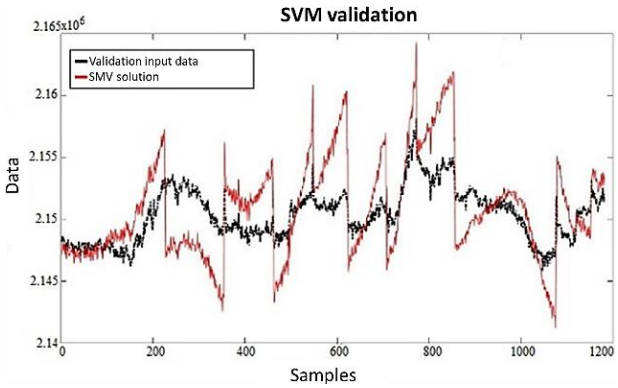

Figure 11. Output 6 feature estimate, actual output (black line), estimated output (red line) by SVM.

4. Conclusions

It was found that for a value of standard deviation σ=0.25, the cross-validation performance was optimal due to the assertiveness percentage of 0.96, i.e., for a total of 2,074,259 samples purified by principal component analysis and trained for validation supervision were right in 1991288 samples. On the other hand, the volume of data for data analysis with the techniques proposed in this document is an advantage compared to other validation and estimation techniques and methodologies.

The values used for the constants b_1, b_2 y λ were 0.1, 0.15, and 0.2, respectively, obtaining strict clustering and prediction results using the support vector machine technique. The level of assertiveness proposed as the weight of the total samples with the difference between the real ones and the prediction showed as a result that the technique to improve learning and pattern recognition that best suits the database refined in this document is the vector support machine over the technique of artificial neural networks. That is, the techniques can vary between one or another or a hybrid between them depending on the nature and the database and the set of samples considered as outputs, and the data set to be estimated.

The data analyzed in this document generates information on the behavior of individual customers regarding requests to reduce electricity consumption. Both, energy supply and demand can be managed more efficiently if utility providers and consumers get accurate information about energy use (for example, consumers could be guided to move some electricity consumption to off-peak hours and reduce the need for non-renewable power plants to be activated at peak times). This information can also be used to improve network reliability, respond to outages, reduce the cost of distribution operations, or measure the impact of a demand response program.

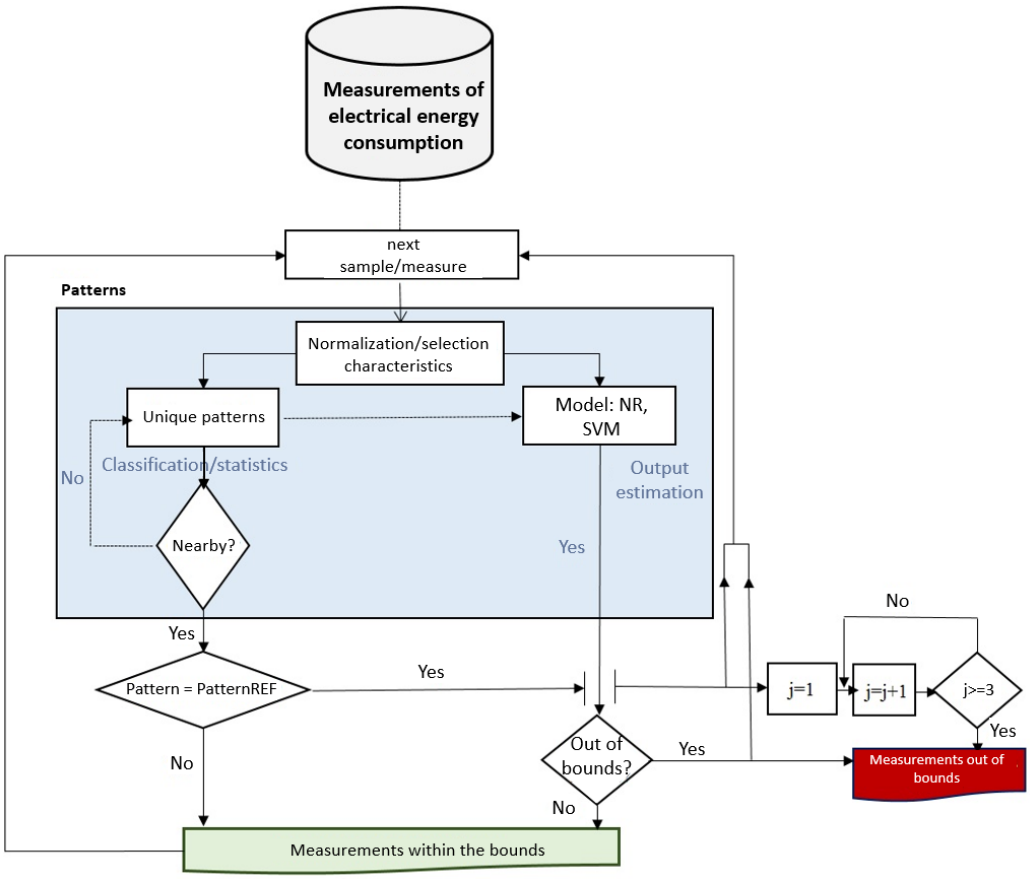

Figure 12. Proposal for predicting measures to generate a warning.

The proposal that appears at the end of the document arises from the need to predict the behavior of electrical energy consumption by analyzing a single voltage output that would allow establishing whether a given load complied with the operating limits.

Funding

This research received no external funding.

Autor Contributions

F. Mesa: Conceptualization, Methodology, Formal analysis, Investigation, Writing - Original Draft. R. Ospina-Ospina: Formal analysis, Investigation, Writing - Original Draft. D. Devia-Narváez: Conceptualization, Formal analysis, Investigation, Writing - Review & Editing.

All authors have read an agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

References

[1] E. Hernandez-Leal, N. Duque-Mendez, J. Moreno Cadavid, “Big Data: una exploración de investigaciones, tecnologías y casos de aplicación”, *TecnoLógicas*, vol. 20, no 39, pp. 17-24, 2017, doi: https://doi.org/10.22430/22565337.685

[2] R. Alvaro-Hermana, J. Fraile-Ardanuy, J. Merinod, "Algorithm development for night charging electric vehicles optimization in big data applications", *Procedia Computer Science*, vol. 109, pp. 793-800, 2017, doi: https://doi.org/10.1016/j.procs.2017.05.329

[3] X. Zhao, "Research on management informatization construction of electric power enterprise based on big data technology", *Energy Reports*, vol. 8, no 7, pp. 535-545, 2022, doi: https://doi.org/10.1016/j.egyr.2022.05.124

[4] A. Fernández, A. Gómez, F. Lecumberry, A. Pardo, I. Ramírez, "Pattern Recognition in Latin America in the "Big Data" Era", *Pattern Recognition*, vol. 48, pp. 1185-1196, 2015, doi: https://doi.org/10.1016/j.patcog.2014.04.012

[5] C. G. Kiran, V. P. Lekhesh, R.V. Abdu, K. Rajeev, "Traffic Sign Detection and Pattern Recognition using Support Vector Machine", *Seventh International Conference on Advances in Pattern Recognition*, Kolkata, 2009, pp. 87-90, doi: https://doi.org/10.1109/ICAPR.2009.58

[6] O. V. Senko, A. V. Kuznetsova, "Pattern recognition method using ensembles of regularities found by optimal partitioning", *20th International Conference on Pattern Recognition*, Estambúl, 2010, pp. 2957–2960, doi: https://doi.org/10.1109/ICPR.2010.724

[7] Center for Machine Learning and Intelligent Systems, "Individual household electric power consumption Data Set", 2012. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption

[8] J. Li, X. Chu, W. He, F. Ma, R. Malekian, Z. Li, "A Generalised Bayesian Inference Method for Maritime Surveillance Using Historical Data", *Symmetry*, vol. 11, pp. 188, 2019, doi: https://doi.org/10.3390/sym11020188

[9] S. Santoso, *Standard Handbook for Electrical Engineers, Seventeenth Edition*. US: McGraw-Hill Education, 2018.

[10] P.S.P. Ignacio, I.K. Darcy, "Tracing patterns and shapes in remittance and migration networks via persistent homology", *EPJ Data Science*, vol. 8, pp. 1-23, 2019, doi: https://doi.org/10.1140/epjds/s13688-018-0179-z

[11] E. Guzmán, M. Vázquez, D. del Valle, P. Pérez-Rodríguez, "Artificial Neuronal Networks: A Bayesian Approach Using Parallel Computing", *Revista Colombiana de Estadística*, vol. 41, pp. 173-189, 2018, doi: http://dx.doi.org/10.15446/rce.v41n2.55250

[12] J. Calvo-Zaragoza, J. Oncina, "An efficient approach for Interactive Sequential Pattern Recognition", *Pattern Recognition*, vol. 64, pp. 295-304, 2017, doi: http://dx.doi.org/10.1016/j.patcog.2016.11.006

[13] M. Zhang, Y. Liu, Z. Liu, J. Wang, M. Gong, H. Ge, X. Li, Y. Yang, Z. Zou, "Hyper-acidic fusion minipeptides escort the intrinsic antioxidative ability of the pattern recognition receptor CRP in non-animal organisms", *Scientific Reports*, vol. 9, pp. 1-15, 2019, doi: https://doi.org/10.1038/s41598-019-39388-8

[14] A. S. Iwashita, V. H. C. de Albuquerque, J. P. Papa, "Learning concept drift with ensembles of optimum-path forest-based classifiers", *Future Generation Computer Systems*, vol. 95, pp. 198-211, 2019, doi: https://doi.org/10.1016/j.future.2019.01.005

[15] P. Carravilla, J. Chojnacki, E. Rujas, S. Insausti, E. Largo, D. Waithe, B. Apellaniz, T. Sicard, J. Julien, C. Eggeling, J. Nieva, "Molecular recognition of the native HIV-1 MPER revealed by STED microscopy of single virions", *Nature Communications*, vol. 10, 2019, doi: https://doi.org/10.1038/s41467-018-07962-9

[16] M. Zhang, Y. Liu, Z. Liu, J. Wang, M. Gong, H. Ge, X. Li, Y. Yang, Z. Zou, "Online modeling and identification of plug-in electric vehicles sharing a residential station", *International Journal of Electrical Power & Energy Systems*, vol. 108, pp. 162-176, 2019, https://doi.org/10.1016/j.ijepes.2018.12.024