Journal of applied research and technology
ISSN: 1665-6423
Universidad Nacional Autónoma de México, Instituto de Ciencias Aplicadas y Tecnología

Valencia, Andrés M.; Caratar, Jesús; Caicedo, Gladys; Chamorro, Cristian

Proposal for a KDD-based procedure to obtain a set of intelligent systems training applied to the identification of failures in hydroelectric power plants

Journal of applied research and technology, vol. 18, no. 6, 2020, pp. 376-389
Universidad Nacional Autónoma de México, Instituto de Ciencias Aplicadas y Tecnología

DOI: https://doi.org/10.14482/INDES.30.1.303.661

Journal of Applied Research and Technology

www.jart.icat.unam.mx

Journal of Applied Research and Technology 18 (2020) 376-389

Original

Proposal for a KDD-based procedure to obtain a set of intelligent systems training applied to the identification of failures in hydroelectric power plants

Andrés M. Valencia[a] • Jesús Caratar[b*] • Gladys Caicedo[b] • Cristian Chamorro[c]

[a]*Master of Engineering Program: Emphasis on Systems Engineering, Universidad del Valle, Cali, Colombia*
[b]*High tension research group (GRALTA), Graduate Program of the School of Electrical and Electronic Engineering, Universidad del Valle, Cali, Colombia*
[c]*DI&ID Group (Development, Innovation and Design Research), Design Department, Universidad del Valle, Cali, Colombia*

Received 04 23 2020; accepted 06 04 2020
Available online 12 31 2020

Abstract: This paper presents a procedure based on KDD (Knowledge Discovery Data), which allows the analysis of a data set to obtain structured information from the behavior of the system under specific conditions, such as system failure conditions at a hydroelectric power plant. By applying this procedure, the information obtained, it is structured in such a mode so that it can be used on the training of intelligent systems focused on fault diagnosis. The former procedure is necessary in the intelligent systems development stage because obtaining an effective training set requires extreme time and effort. The procedure was applied in the historical records of the Amaime hydroelectric power plant, located in Palmira, Valle del Cauca, Colombia, aiming to obtain patterns of behavior of the protection system which can be translated to different failures. This was possible by integrating a data mining technique such as hierarchical clustering and the statistical technique called the interpolation function. The main achievement of this work is to present a structured procedure that reduces the time to obtain a training set. In this specific case, the training set for mechanical failure of a hydroelectric power station was obtained, which can be used in the development of an intelligent system for failures diagnosis.

Keywords: knowledge discovery data, data mining, intelligent systems, failure diagnosis, training set, hydroelectric power plant

*Corresponding author.
E-mail address: jesus.caratar@correounivalle.edu.co (Jesús Caratar).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

1. Introduction

The processes in hydroelectric power plants are complex due to the large amount energy that must be controlled (Sanz Osorio & Almécija, 2016), for this reason, a large number of variables are monitored and controlled. In case of a failure occurrence, this will represent a danger to the equipment, the surrounding areas and personnel that operate these plants (Dominguez Gavilanes & Logroño Vargas, 2010). In this context, electrical and/or mechanical protections are essential for the equipment and production process safety at the Hydroelectric Power Plant (HPP) (Palacios, Echeverría, & Barba, 2015).

At the HPP, qualified personnel must analyze the information provided by the protection system records to identify the type of failure and thus take corrective measures (Sarkar et al., 2014). This task is complex, due to the large amount of information that must be analyzed. Furthermore, this information generally lacks of a defined structure and is not centralized (Efrén & Alvarado, 2012), therefore, make a diagnosis can take considerable time which is quantified as economic losses by lost revenue.

For this reason, intelligent systems dedicated to fault identification are currently being developed, which take the signals supplied by the protection systems, e.g. the research conducted in (Octavio et al., 2014), in which an intelligent system based on fuzzy logic and neural networks was implemented to identify faults in a hydroelectric plant, the authors developed a training set using databases of load of the electrical system of power in normal operation. Similarly the research developed in (Amaya Simeón, 2008), developed a rule-based knowledge base to identify failures of a HPP, these rules were established by evaluating the behavior of the different signals of the protection system to predictively diagnose the state of the all equipment, in order to support maintenance tasks. A similar procedure can be observed in the investigation presented in (Dorantes, Gonzalez, & Mendez, 2014), Where a fault diagnosis system was developed for the Power Electrical System (PES) using fuzzy logic. In general, in the references consulted, few works applied to the diagnosis of failures in HPP were found. Furthermore, these investigations focus on the electrical variables of the generation systems and not on the mechanical variables, on the other hand, these works do not present in a clear and structured way the methodology used to obtain the training set. For this reason, this work presents a structured procedure to obtain a training set for the diagnosis of mechanical failures, using the signals of the protection system of an HPP. As a case of study, the investigation was applied at the Amaime HPP, located in Palmira, Valle, Colombia (Celsia, 2020).

In this context, this work presents a procedure based on the KDD (Knowledge Discovery Data) process that allows to develop and optimize the time to obtain the correlations between the output signals of the mechanical protection system and the possible causes of failures. This information is called a training set, which can be used in the development of an intelligent system for diagnosing faults in HPP.

This procedure allowed to obtain the training set for the diagnosis of mechanical failures in HPP.

This paper is organized as follows: the methodology used to obtain a training set is presented in section 2. Section 3 describes the process of obtaining the training set applied to an HPP. Section 4 presents the discussion of the results are presented in section 4 and the conclusions are presented in section 5.

2. Methodology

The procedure to obtain the training set by mean of the application of the KDD methodological process is presented in three parts: first, the KDD process is described. The second part describes the types of data provided by mechanical protection system of a HPP. The third part shows the proposed procedure applied on a set of data to determine a training set through the analysis of the behavior of the protection system records.

2.1. KDD process

KDD process was exposed by Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy (1996). This procedure is developed on 5 steps that allow the discovery of knowledge, starting with raw data and ending with practical knowledge (Ristoski & Paulheim, 2016). The process is shown in Figure 1.

The KDD process applies the following steps:

Selection: It begins with the understanding of the domain where the application is developed, that means, it is required to have a basic knowledge about the elements and variables associated with the studied system. Likewise, it is necessary to identify what is the objective to achieve at the end of this process. Once the objective to be achieved is clear, the information of the variables associated with the problem must be structured in such a way that allows to select the data group to be used on later stages (Ristoski & Paulheim, 2016). With the data group defined the data mining technique is selected.

Preprocessing: in this step a cleaning is done on the database selected (delete or completing data's) to obtain complete records that allow subsequent analyzes, for this, statistical data cleaning and repair techniques such as interpolation and covariance analysis functions are used.(Ebtehaj, Bonakdari, Zeynoddin, Gharabaghi, & Azari, 2020).

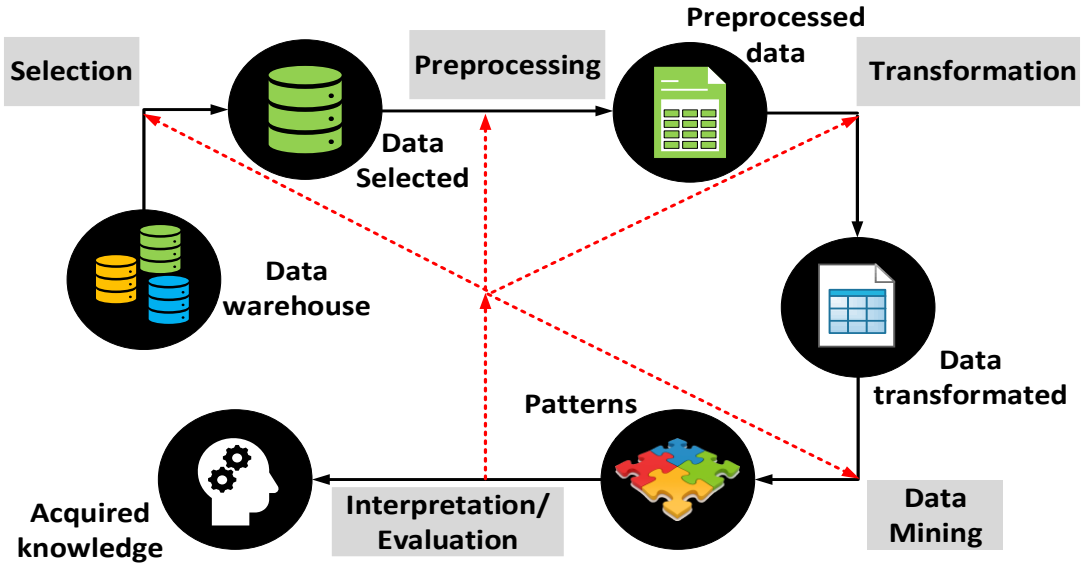

Figure 1. Overview of the KDD process.

Transformation: After preprocessing, the data is adjusted to an appropriate form that allows the implementation of the selected data mining technique, for this, different strategies are applied, such as the binarization of states in a variable or the methods of reducing dimensions that allow optimizing the data extraction algorithms that will be used on later stage, thus reducing the number of variables under consideration (Cibulková, Šulc, Sirota, & Řezanková, 2019).

Data mining: in this stage the data is in an appropriate format to apply the selected mining technique which can be descriptive or predictive (Ristoski & Paulheim, 2016). In addition to these techniques, data mining uses visualization methods that allow to represent data graphically and generate patterns in order to facilitate the understanding of the information obtained (Garcia, Molin, & Berlanga, 2018).

Finally, the level of complexity in which the results can be presented must be established, because it is necessary to choose between information easy to understand with little precision or information difficult to understand and very precise (Han, Kamber, & Pei, 2012).

Interpretation: The patterns and models obtained by data mining techniques are evaluated to determine their validity using the visualization methods, Likewise, the relevance of the knowledge obtained for the fulfillment of the objective set at the beginning of the process is evaluated. In case of not reaching the proposed objective the previous steps should be evaluated to determine what modifications should be made and re-execute the process (Ristoski & Paulheim, 2016).

Having defined the stages of the KDD process, below, an application example of the different stages is performed to obtain the training set for an intelligent system that could be used to identify failures in hydroelectric power plants automatically.

2.2. Description of data types supplied by the protection system of a hydroelectric power plant

Currently, HPP have a digital protection system that monitors the operation of the mechanical and electrical components of the system to react to anomalous situations or failures. In case of eventuality, these protection systems activate some protection functions, which have a logic implemented that allows them to determine whether they should enter an alarm state or issue a trip order that initiates a stop sequence of the generation unit (Carreño-Pérez, Morales-Rivera, & Rivas-Trujillo, 2019).

As shown in Figure 2, the mechanical protection system monitors and protects mechanical equipment using electronic devices called PLC's (Program Logic Control), the actions and records of these devices are displayed through SCADA (Supervisory Control and Data Acquisition). On the other hand, the electrical components are monitored and protected by digital relays, which in case of a fault, activate specific protection functions for the type of fault, these actions are registered in the SOE (Sequence Of Events), which is a binary register that gives information on the electrical and mechanical protection functions that were activated during the fault (Penin, 2007).

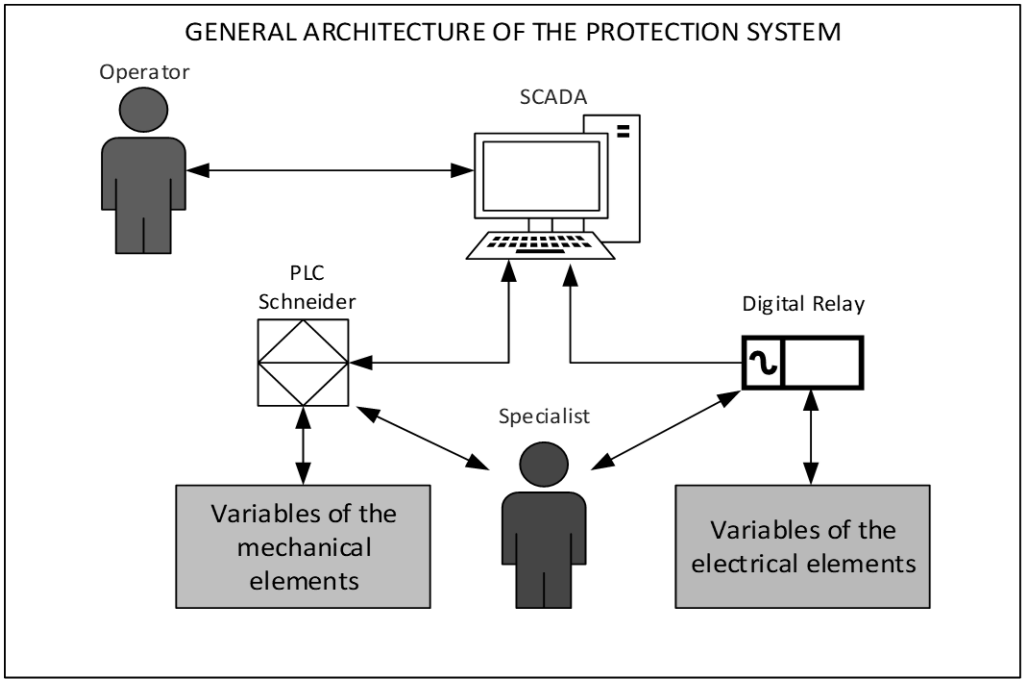

Figure 2. General architecture of the protection system from the Amaime Colombia HPP.

In this way, the information provided by a hydroelectric protection system for the identification of faults is divided into the records presented by the SCADA (mechanical components) and the relay (electrical components) that contain both the analog variables records and the binary variables contained in the SOE.

Therefore, the records from the SCADA and the relay provide categorical data (binary) and quantitative data (analog) which are divided as follows:

Alarms: categorical data, indicates that the component associated to the sensed variable is in a state that represents a potential danger to the operation of the plant (this does not generate a system stop sequence).

Trips: categorical data, are states where the measured variables exceed a set value, this set value is critical and represents a danger to the operation of the system, for this reason, a stop sequence of the generation system is performed.

Analog records: quantitative data measured by the different sensors installed in the system elements (Temperature, pressure, vibration, current, voltage, etc.), which are stored in the SCADA and the relay.

3. KDD process to obtain a training set for identification of failures in hydroelectric power plants

This chapter describes the application of the KDD process, shown in Figure 1, to obtain a training set from raw data of the mechanical protection system of the Amaime HPP located in Valle del Cauca, Colombia.

3.1. Data selection

A brief description is made about the hydroelectric power plant, in such a way that it allows us to understand the structure of the available data and, therefore, facilitate its selection.

The Amaime HPP is listed as a small hydroelectric power station since its operating range is below 20 MVA, it has two turbo generator groups, each consisting of a Francis type turbine and a salient pole synchronous generator with a power of 11.56 MVA.

Each turbo generator group consists of subsystems with specific functions. Figure 3 shows the mechanical elements associated to the generation process and the subsystems that support its operation.

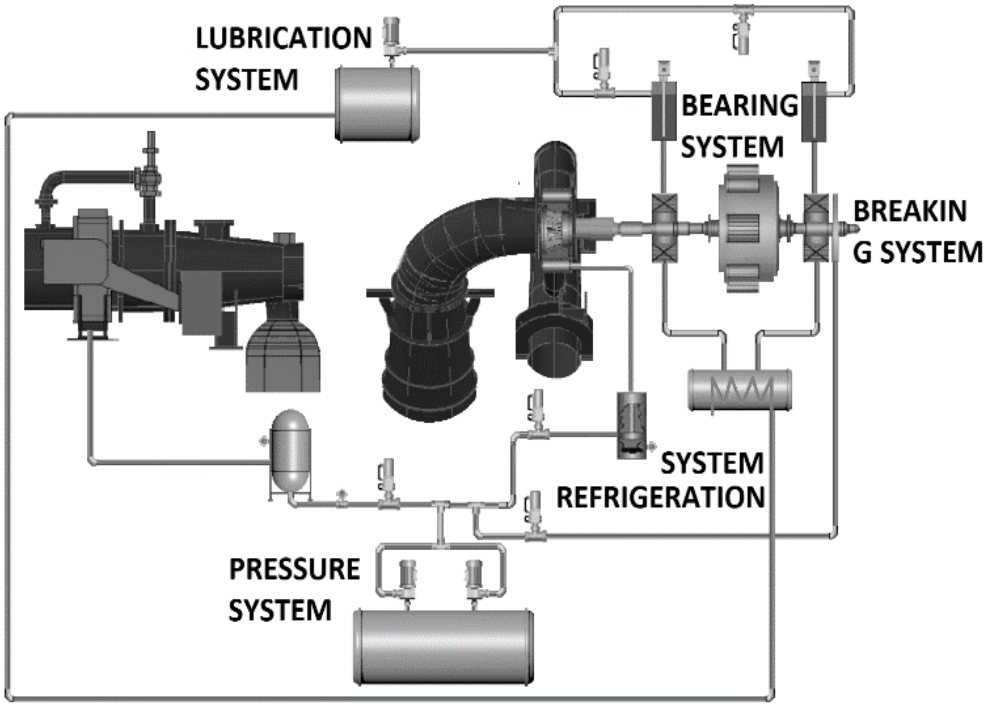

Figure 3. Subgroups of mechanical components of the generation system, image taken from the SCADA System.

The subsystems presented in Figure 3 have a series of sensors installed, which allow monitoring and recording the state and behavior of the different components associated to the mechanical system.

The mechanical variables associated to the bearings, lubrication and pressure system are presented in Table 1, as well as the abbreviation that will be used to refer to these variables in this article.

Once all the analog variables that make up the turbo generator group have been identified, the historical records of these variables are selected.

With this, the selection of the data set is completed and the data mining technique to be used will be selected, which should allow the records to be separated into groups with similar behaviors.

Having defined the data group, the next step is the preprocessing of the available information.

3.2. Data set preprocessing

The preprocessing begins with the search of the records of variables that present anomalous or missing data, as observed in Table 2. where 3 randomly chosen records are shown, among which one presents a missing data. After locating this data, the interpolation function shown in (1) is used to replace the anomalous value or complete the missing value.

$$Y_i = Y_0 + \left(\frac{Y_1 - Y_0}{X_1 - X_0}\right) + (X_i - X_0) \quad (1)$$

The use of this equation in this process is valid since the sampling rate of data are 5 seconds and during this time interval there are no significant changes in the variables of the mechanical process, because the changes in the variables of these systems are too slow and occurs in minute intervals. From the above, a linear behavior of the variables between nearby records can be assumed (Devore, 2016).

Dataset records were adjusted by searching and deleting records when the system is out of service to avoid having fault records that do not exist, for this, histograms such as the one presented in Figure 4 were used, in which it is evidenced that the monitored variable “Accumulator oil pressure” (AOP) indicates a value of pressure zero when the system is out of operation, this could be interpreted by the protection system as a failure.

In Figure 5 the histogram of the filtered variable (AOP) is shown. This procedure was repeated with each of the variables that make up the data set.

Table 1 Mechanical variables sensed by the bearing subsystem, and the lubrication and pressure subsystem.

System	Variable	Abbreviation
Bearing System	Bearing block temperature, radial direction, driver side	BBTRDS
	Bearing block temperature, axial direction 1, driver side	BBTA1DS
	Bearing block temperature, axial direction 2, driver side	BBTA2DS
	Oil pan temperature, driver side	OPTDS
	Axial vibration on the bearing shaft, driver side	AVBSDS
	Radial vibration on the bearing shaft, driver side	RVSDS
	Bearing block temperature, radial direction, opposite to driver side	BBTRODS
	Oil pan temperature, opposite to driver side	OPTODS
	Radial vibration on the bearing shaft, opposite to driver side	RVSODS
Lubrication and pressure system	Tank oil temperature	TOT
	Accumulator oil pressure	AOP
	Oil temperature at the tank inlet	OTTI
	Oil temperature at bearing oil inlet	OTBOI

Table 2. Records of mechanical variables with wrong data.

BBTRDS	BBTA1DS	BBTA2DS	OPTDS	AVBSDS	RVSDS	BBTRODS	OPTODS	RVSODS	TOT	AOP	OTTI	OTBOI
49	80	47	45	1,38	0,49	63	25	1,17	36	141	47	38
50		48	45	1,40	0,46	63	25	1,15	37	141	47	38
49	80	48	45	1,42	0,49	63	25	1,24	37	141	47	38

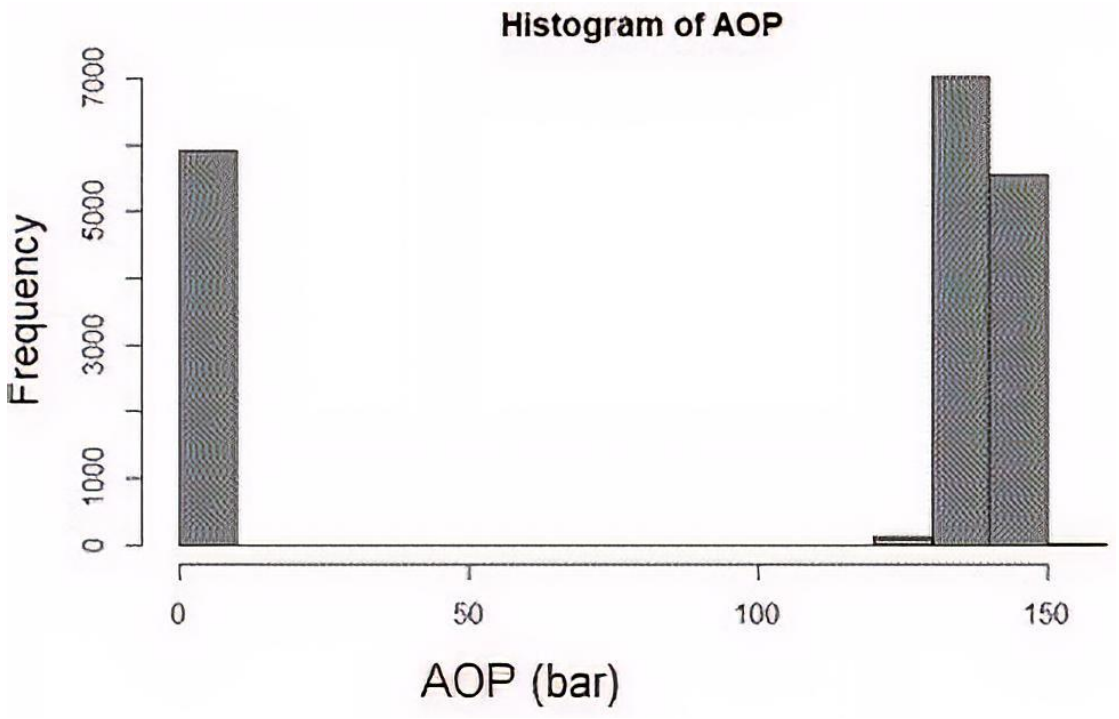

Figure 4. Histogram of the Accumulator Oil Pressure (AOP) variable.

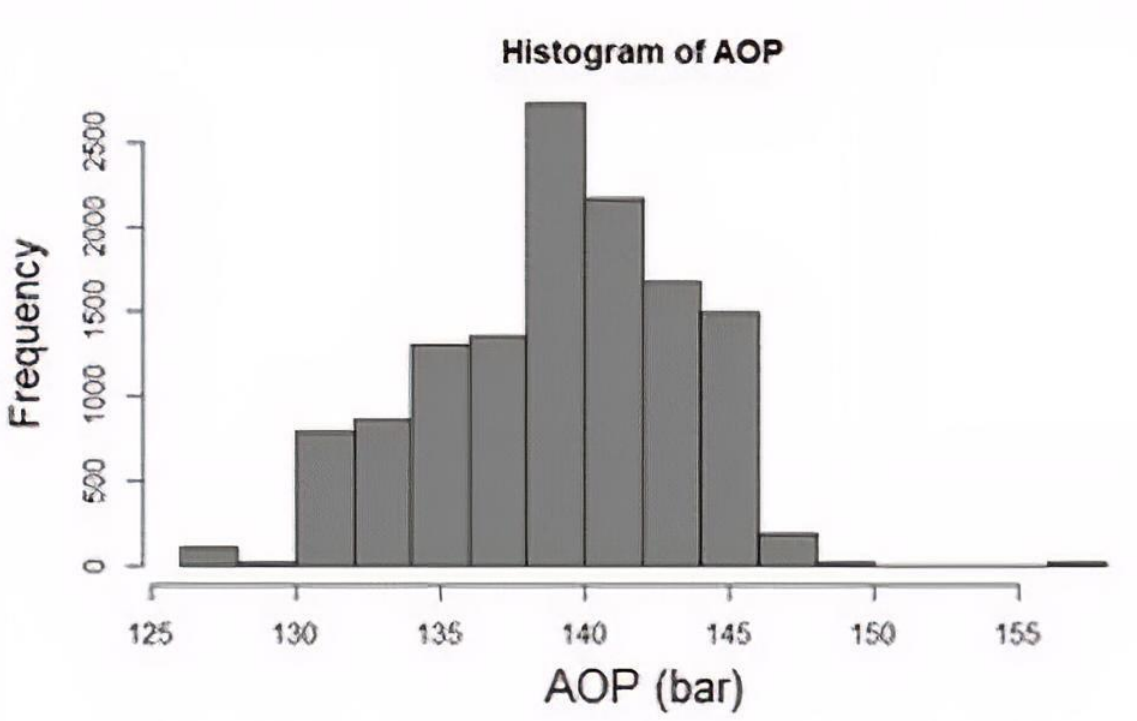

Figure 5. Histogram of the Accumulator Oil Pressure (AOP) variable without records of the system out of service.

3.3. Data transformation

In this step the adjusted data is transformed to a format that can be processed by some data mining algorithm. For this, it was decided to transform the numerical data of the records into categorical data containing information on the status (ok, alarm, trip) of each component.

In this first transformation, the setting parameters of the protection functions presented in Table 3 were used, which allow to determine the state of an element of the system.

With the settings presented in Table 3, the numerical values of the variables are compared with the set values. Depending on this comparison, each record assumes a state, i.e. When the value of a variable reaches the alarm set value (value T1 in Figure 6) its status changes to "alarm".

Thus, the data set like that presented in Table 4, which provides information on the state of the system variables is obtained.

Now, the information of interest is focused on the records of the variables in a failure situation, therefore, a filter is done to eliminate the rows that exhibit the normal system behavior (rows without tripping), For this reason, the applied filter removes all rows without presence of trip states.

After filtering, a data set which exclusively contains categorical records in fault states (rows with tripping), e.g. the fragment of records shown in Table 5. The examples shown are only a fraction of the total records, because the volume of records obtained is high and it is not possible to show it in the document.

At this point, the registers have an easy to understand structure, where the state of each variable in the evaluated failures is known., however, it still does not have the format required to apply the selected data mining technique, since the clustering requires that the data be numerical to calculate the distances between the records.

To achieve that, a new table was constructed where each variable is divided into the states that it can assume (ok, alarm, trip), creating for each variable in table 5 at least three new variables in Table 6. After this new classification, for each state of the variables it must be specified whether it is active or not, in this way, the value “1” indicates that the variables are active in that state and the value “0” indicates that the variables are not active in that state. thus obtaining, the data set presented in Table 6, in which it is observed that variables such as “AOP” can assume 5 different states, but only one at a time.

At this point, the data transformation process is completed.

Table 3. List of variables associated with the bearing, lubrication and pressure system with alarm and trip parameters.

Systems	Variable	Alarm setting	Trip setting
Bearing System	Bearing block temperature, radial direction, driver side	80°C	85°C
	Bearing block temperature, axial direction 1, driver side	80°C	85°C
	Bearing block temperature, axial direction 2, driver side	80°C	85°C
	Oil pan temperature, driver side	55°C	60°C
	Axial vibration on the bearing shaft, driver side	2.5mm/s	3.5mm/s
	Radial vibration on the bearing shaft, driver side	2.5mm/s	3.5mm/s
	Bearing block temperature, radial direction, opposite to driver side	80°C	85°C
	Oil pan temperature, opposite to driver side	60°C	65°C
	Radial vibration on the bearing shaft, opposite to driver side	2.5mm/s	3.5mm/s
Lubrication and pressure system	Tank oil temperature	55°C	60°C
	Accumulator oil pressure	127bar	130bar
	Oil temperature at the tank inlet	50°C	54°C
	Oil temperature at bearing oil inlet	42°C	48°C

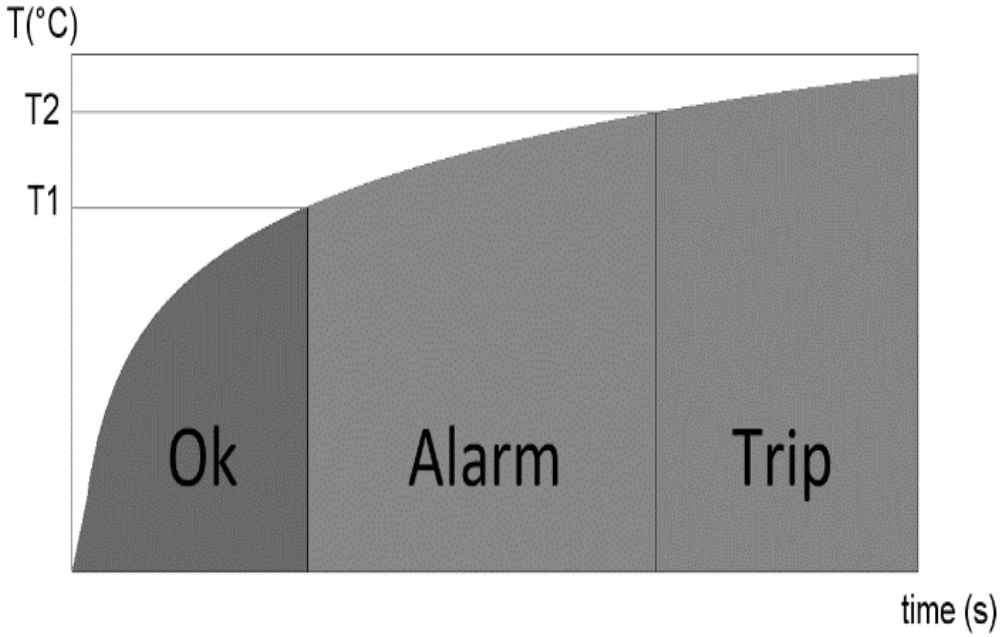

Figure 6. Example of protection function of temperature.

Table 4. Fragment of the transformed data set.

BBTRDS	BBTA1DS	BBTA2DS	OPTDS	AVBSDS	RVSDS	BBTRODS	OPTODS	RVSODS	TOT	AOP	OTTI	OTBOI
Ok	Alarm	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok
Ok	Alarm	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok
Alarm	Trip	Alarm	Alarm	Ok	Ok	Ok	Ok	Ok	Ok	Alarm low	Ok	Ok

Table 5. Data set fragment in fault state.

BBTRDS	BBTA1DS	BBTA2DS	OPTDS	AVBSDS	RVSDS	BBTRODS	OPTODS	RVSODS	TOT	AOP	OTTI	OTBOI
Alarm	Trip	Alarm	Alarm	Ok	Ok	Ok	Ok	Ok	Ok	Alarm Low	Ok	Ok
Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Trip low	Ok	Ok
Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Trip low	Ok	Ok

Table 6. fragment of the binarized data set.

AOP Ok	AOP AlarmB	AOP AlarmA	AOP TripB	AOP TripA	OTTI Ok	OTTI Alarm	OTTI Trip	OTBOI Ok	OTBOI Alarm	OTBOI Trip
0	0	1	0	0	0	1	0	0	1	0
0	0	0	0	1	0	1	0	0	1	0
0	0	0	0	1	0	1	0	0	1	0

3.4. Data mining implementation

In the data mining process, it was decided to use hierarchical clustering with the distance function based on the Jaccard index and the "Ward.D" clustering method.

The Jaccard index is a method used in agronomy to determine the similarity between two ecosystems based on the species they present in common (Soler, Berroterán, Gil, & Acosta, 2012). This method measures the similarity between two groups, regardless of the type of variable that governs them (Real & Vargas, 1996), since it calculates the cardinality of the intersection of the two sets and then divides it by the cardinality of the union of both groups as shown in (2).

$$J(A,B) = |A \cap B| / |A \cup B| \quad (2)$$

Next, the number of clusters is specified using the elbow method (Garcia et al., 2018), which allows determining a point where the increase in the number of groups does not generate a significant change in cohesion or relationship between the centroids of each group and on the elements within each group. In the present case study, the estimate of the optimal cluster number is presented in Figure 7.

In it is observed that after cluster number 8 the cohesion between the elements of each group stop increasing, and the cohesion between the centroids of each group ceases to decrease significantly, this indicates that the optimal cluster number is 8.

Once the optimal cluster number is determined, the hierarchical clustering is carried out, implementing a pruning of the dendrogram at K = 8, this results in the distribution presented in Figure 8.

In Figure 8 each division between nodes represents the groups that are formed according to the relationship between the data, in addition, the length of the vertical axis represents the cohesion between the clusters of the same node.

The dendrogram presented in Figure 8 allows to verify that the number of clusters supported by the elbow method is correct, since the distances in the branches after pruning are small compared to the distances before pruning.

After performing the Clustering, the group label is assigned to each of the records presented in Table 5, thus obtaining Table 7 in which the sampled data set for training is presented.

3.5. Evaluation and interpretation of failure cases.

In the evaluation and interpretation process, the clusters obtained after the data mining process are analyzed, with the aim of verifying if the correct clustering was performed.

Next, in Table 8 a fragment of the registers that make up cluster 3 is presented, in which a pattern in the behavior of the signals is observed, composed of a high pressure trigger in the AOP variable and alarm signals in some other variables, however, one record has a different behavior (record in yellow, table 8), since it presents a low pressure trip, which is a significantly different behavior from the real fault. This means that the record must be removed from this group and subsequently reassigned to the cluster that best represents it.

The behavior verification process must be carried out on each of the generated clusters, with the aim of correcting the records that could be poorly classified.

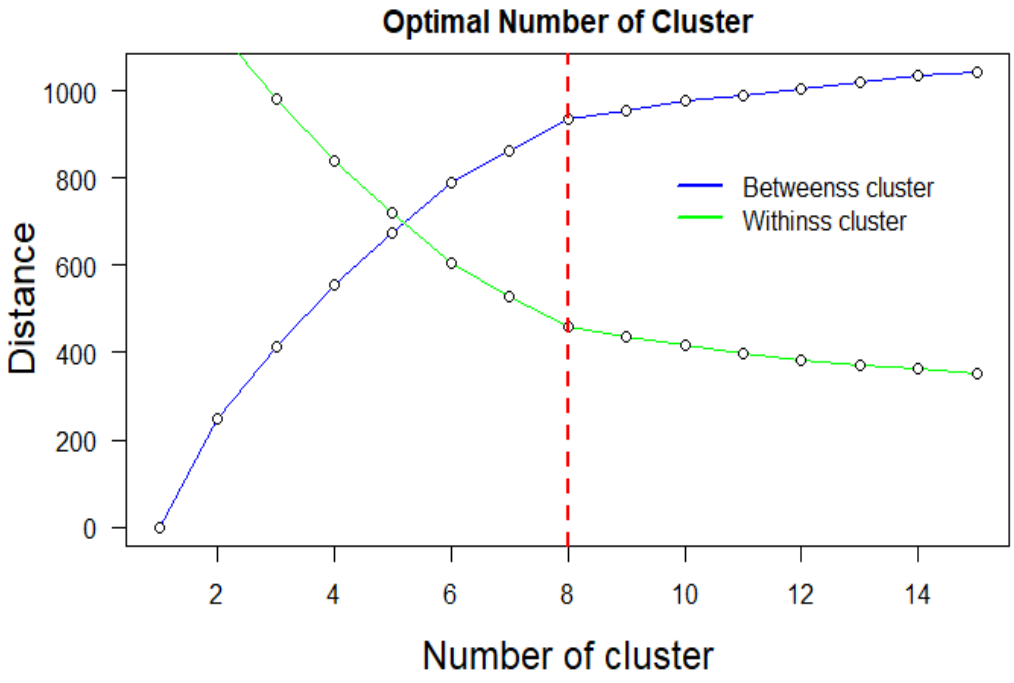

Figure 7. Relation between distance and numbers of clusters.

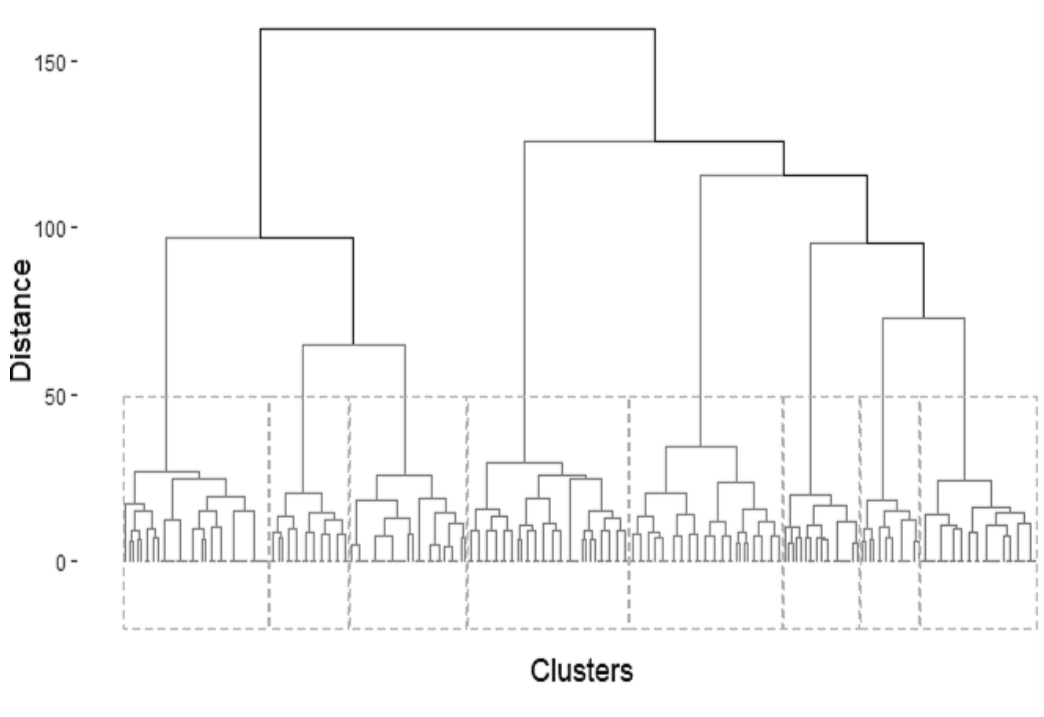

Figure 8. Hierarchical cluster distribution (Distance function Jaccard, grouping method Ward D2).

Table 7. Data set fragment available for evaluation and interpretation.

BBTRDS	BBTA1DS	BBTA2DS	OPTDS	AVBSDS	RVSDS	BBTRODS	OPTODS	RVSODS	TOT	AOP	OTTI	OTBOI	Cluster
Alarm	Trip	Alarm	Alarm	Ok	Ok	Ok	Ok	Ok	Ok	Alarm Low	Ok	Ok	Case 6
Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Trip Low	Ok	Ok	Case 8
Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Trip Low	Ok	Ok	Case 8

Table 8. Records grouped in cluster 3.

BBTRDS	BBTA1DS	BBTA2DS	OPTDS	AVBSDS	RVSDS	BBTRODS	OPTODS	RVSODS	TOT	AOP	OTTI	OTBOI	cluster
Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Ok	Ok	High Trip	Ok	Ok	3
Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Ok	High Trip	Ok	Ok	3
Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Ok	Alarm	Alarm	Ok	High Trip	Ok	Ok	3
Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Ok	High Trip	Ok	Ok	3
Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Ok	High Trip	Ok	Ok	3
Alarm	Alarm	Alarm	Alarm	Ok	Alarm	Alarm	Alarm	Alarm	Ok	High Trip	Ok	Ok	3
Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Ok	Low Trip	Ok	Ok	3
Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Ok	Alarm	Alarm	Ok	High Trip	Ok	Ok	3

After analyzing each of the clusters and adjusting the classification of the records, the interpretation of each of the obtained patterns were made, thus the Table 9 is generated, in which a characteristic pattern of each cluster is presented.

The characteristic pattern of the types of failure presented in Table 9, was determined by extracting the state with the highest frequency of each of the variables contained in the records of each cluster.

To carry out the interpretation of each of the clusters present in Table 9, it is necessary to correlate the knowledge about the configuration of the plant and the types of failure that may occur, it is recommended at this point to consult a specialist in fault identification in HPP that corroborate the type of failure that each cluster represents.

At as a result of this process, Table 10 is obtained, in which each of the type of failure (cluster) showed in the Table 9 is presented with their corresponding interpretation. The information given in the interpretation of Table 10 should be as detailed as possible.

Therefore, the type of failure that each cluster represents and the variations in the signals for each cluster (example in Table 8) define the behavior that the protection system of the HPP can show these failures are presented. This information is called a training data set.

4. Discussion of results

The KDD process applied to the information collected from the protection system of the HPP allowed to clean the information collected, since there may be cases in which the data is incomplete or erroneous, see table 2. Missing values were added using an algorithm that looked for spaces without values and applied the interpolation function (1) to add the calculated value. Therefore, a computational algorithm allows this process to be carried out quickly and efficiently.

Then, the analog values of the variables were categorized according to states (ok, alarm and trip), see table 4. With this structured data, the data mining technique called hierarchical cluster was applied, this technique allows taking a series of signals and grouping them into sets that have a close relationship, table 8. Finally, the clusters that define characteristics of some failures to be considered were obtained, table 9.

The help of protections experts was vital in the association of each cluster with a cause of failure, see table 10. In this context, it is assumed that the interpretation of this information is valid, since it is based on the great knowledge that the experts possess about the system under study.

Table 9. characteristic patterns of failure cases by cluster.

BBTRDS	BBTA1DS	BBTA2DS	OPTDS	AVBSDS	RVSDS	BBTRODS	OPTODS	RVSODS	TOT	AOP	OTTI	OTBOI	cluster
Alarm	Alarm	Alarm	Alarm	Ok	Ok	Alarm	Alarm	Ok	Alarm	Low Trip	Alarm	Alarm	1
Ok	Ok	Ok	Ok	Ok	Ok	Trip	Alarm	Alarm	Ok	Low Alarm	Ok	Ok	2
Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Alarm	Ok	High Trip	Ok	Ok	3
Ok	Ok	Ok	Ok	Alarm	Ok	Ok	Ok	Trip	Ok	Ok	Ok	Ok	4
Alarm	Alarm	Alarm	Alarm	Ok	Ok	Ok	Alarm	Ok	Ok	Ok	Alarm	Trip	5
Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Trip	Ok	Ok	Ok	Ok	6
Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Alarm	Alarm	Low Trip	Alarm	Ok	7
Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Ok	Trip	Alarm	8

Table 10. Interpretation of failure by each cluster.

Cluster	Interpretation
1	oil leakage between the accumulator tank and the bearings, cause overheating on driven side bearings and opposite driven side.
2	oil leakage between the accumulator tank and the opposite driven side bearing, causes overheating only in the opposite driven side bearing. Possibly broken hose or valve leakage.
3	Plugging between the accumulator tank and the bearings, prevents oil from reaching the generator and causes overheating. Possibly strangled hose or clogged filters.
4	Wear on the generator shaft which produces vibrations.
5	Cooling system failure, the hot oil that comes out of the bearings cannot cool down. Check the condition of the heat exchanger.
6	Excessive radial vibration. Check turbine rotor.
7	Leakage between the accumulator tank and the bearings, the system has an imbalance in the generator shaft, possibly due to wear. Check the condition of the bearing or shaft.
8	Cooling system failure, hot oil coming out of bearing is not cooled and returns to bearing at high temperature. Check the condition of the heat exchanger.

Therefore, correlations between the variables that make up the protection system and different cases of failure were obtained, determining in this way a training set or set of patterns of protection system behavior, with this information a user can easily build an intelligent system dedicated to the identification of failures in hydroelectric power plants.

5. Conclusions

The described process allowed obtaining a set of records of the variables that make up the protection system of a hydroelectric power plant, these sets contain information of the behavior of the variables of the protection system that generate some failures. Therefore, the proposed procedure allows to obtain a training set, which can be used in the development of intelligent systems to provide support in the process of identify failures on hydroelectric power plants.

This procedure considerably reduces the time of construction of a training set, since it does not require establishing the fault cases one by one.

The steps shown in the KDD process were enough to obtain the desired knowledge, demonstrating the high impact that data mining techniques can have in the field of fault identification.

The process shown is not automatic and therefore the intervention of the developer is necessary throughout the process, the knowledge of the domain by the developer is vital to obtain good results.

The procedure presented is easily replicable in other industries that present digital control systems such as PLCs and system operation records.

With all of the above, the development of this type of research is considered to significantly reduce the time required for the construction of complex systems such as intelligent systems, moreover, this work demonstrates the importance of storing the functions records in these types of industries.

As a future work, these results will be applied to an artificial intelligence tool to diagnose electrical and mechanical failures at the Amaime hydroelectric power station.

Acknowledgments

The authors would like to thank Colciencias for its support in the framework of the call for national doctorates 727 of 2015 and Universidad del Valle in the framework of the C.I. 105 of 2017.

References

Amaya Simeón, E. J. (2008). Aplicação de Técnicas de Inteligência Artificial no Desenvolvimento de um Sistema de Manutenção Baseada em Condição. Universidade de Brasília, Brasil.

Carreño-Pérez, J. C., Morales-Rivera, J. P., & Rivas-Trujillo, E. (2019). Redundancy in communication networks for the automation and protection of electrical power systems with IEC 61850. *Informacion Tecnologica, 30*(1), 75–86. https://doi.org/10.4067/S0718-07642019000100075

Celsia. (2020). *Centrales hidroeléctricas*. Retrieved January 10, 2020, https://www.celsia.com/es/centrales-hidroelectricas

Cibulková, J., Šulc, Z., Sirota, S., & Řezanková, H. (2019). The effect of binary data transformation in categorical data clustering. Statistics in Transition, 20(2), 33–47. https://doi.org/10.21307/stattrans-2019-013

Dominguez Gavilanes, E. X., & Logroño Vargas, D. O. (2010). *Diseño e Implementación del Control Automático y Monitoreo del Nivel del Embalse en la Central Hidroeléctrica* Agoyán. Escuela politécnica nacional, Quito, Ecuador.

Dorantes, P. N. M., Gonzalez, J. P. N., & Mendez, G. M. (2014). Fault Detection Systems via a Novel Hybrid Methodology for Fuzzy Logic Systems Based on Individual Base Inference and Statistical Process Control. *IEEE latin america transactions, 12*(4), 706-712. https://doi.org/10.1109/TLA.2014.6868873

Devore, J. L. (2016*). Probability and statistics for engineering and sciences* (Ninth edit). California: Cengage learning.

Ebtehaj, I., Bonakdari, H., Zeynoddin, M., Gharabaghi, B., & Azari, A. (2020). Evaluation of preprocessing techniques for improving the accuracy of stochastic rainfall forecast models. International *Journal of Environmental Science and Technology, 17*(1), 505–524. https://doi.org/10.1007/s13762-019-02361-z

Efrén, I., & Alvarado, V. (2012). Algoritmo neuro-difuso para la detección y clasificación de fallas en líneas de transmisión eléctrica del sistema ecuatoriano usando simulaciones y datos de registradores de fallas. Universidad de Cuenca.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery and data mining. American Association for Artificial Intelligence.

García, J., Molina, J., Berlanga, A., Patricio, M., Bustamante, A., & Padilla, W. (2018). Ciencia de datos. *Técnicas Analíticas y Aprendizaje Estadístico.* Bogotá, Colombia. Publicaciones Altaria, SL.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: concepts and techniques, Waltham, MA. *Morgan Kaufman Publishers*, 10, 978-1.

Morales, C. O. H., González, J. P. N., & Siller, E. G. C. (2014). Detección y diagnóstico de fallas en sistemas eléctricos de potencia (SEP) combinando lógica difusa, métricas y una red neuronal probabilística. *Res. Comput. Sci., 72*, 47-59.

Palacios, C., Echeverría, D., & Barba, R. (2016). Estudio del Impacto de la Implementación del Sistema de Protección Sistémica en la Operación del Sistema Nacional Interconectado. *Revista Técnica" energía", 12*(1), 112-120. https://doi.org/10.37116/REVISTAENERGIA.V12.N1.2016.33

Penin, A. R. (2007). Sistemas SCADA (2nd ed.). Barcelona, España: marcombo, S.A.

Real, R., & Vargas, J. M. (1996). The Probabilistic Basis of Jaccard's Index of Similarity. *Systematic Biology, 45*(3), 380–385. https://doi.org/10.2307/2413572

Ristoski, P., & Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Journal of Web Semantics, 36*, 1–22. https://doi.org/10.1016/j.websem.2016.01.001

Osorio, J. F. S. (2008). *Energía hidroeléctrica* (Vol. 139). Universidad de Zaragoza.

Sarkar, S., Sharma, T., Baral, A., Chatterjee, B., Dey, D., & Chakravorti, S. (2014). An expert system approach for transformer insulation diagnosis combining conventional diagnostic tests and PDC, RVM data. *IEEE Transactions on Dielectrics and Electrical Insulation, 21*(2), 882–891. https://doi.org/10.1109/TDEI.2013.004052

Soler, E., Berroterán, P., Gil, J., & Acosta, R. (2012). Índice valor de importancia, diversidad y similaridad florística de especies leñosas en tres ecosistemas de los llanos centrales de Venezuela. *Agronomía Trop, 62*(1-4), 25-37.