Advanced Analytics for Non Technical Losses of Energy

G. A. Quintero Rojas, *Researcher of the Innovation Area, MVM Ingeniería de Software and* R. A. Gallego, *Director Innovación Area, MVM Ingeniería de Software.*

***Abstract*-- The first step of a technological solution for the recovery of non-technical energy losses (NTL) in the distribution phase is to use data mining techniques. MVM Ingeniería de Software has developed a technology with high potential, oriented in this direction, for Companies Distributors and Traders of energy (DT), through its advanced analytical capacity in descriptive and predictive levels framed in the context smart grid for achieving a reliable distribution network to failures that are due to technical problems and fraudulent users.**

Keywords— Technological Solution, Non-Technical Losses (NTL), Non-Technical Losses Index (NTLI), Advanced Analytics, Descriptive Level, Predictive Level, Electrical Distribution Network, Distributors and Traders of Energy (DT).

I. Introduction

The study of an imprecise issue requires mechanisms to collecting data. The search ends when appears the possibility to create a structure allowing the identification of relevant factors of the situation to be examined. Descriptive research in which the variables are selected and measured independently is generated from the data exploration and measurements which may be integrated in order to determine how to delimit the phenomenon through: behaviors setting and their frequency, identification of conducts and association of variables. The descriptive study requires two fundamental elements: the size of the sample and the data collection[1]. The need to understand past events has led to the development of the discipline known as Business Intelligence (BI)[2], which consists in a set of strategies and aspects geared to the creation and management of knowledge on the medium, through the investigated data belonging to an organization. These data are also studied using predictive analytics to know what will happen.

This paper focuses in the knowledge obtained from the advanced analytics, which can be classified as descriptive (it lets to know what happened in the past) and predictive (it focuses on what will happen). This last level helps uncover hidden patterns in data of high quality, as result of the mathematics and statistics applied to the subject of study. The purpose of MVM is to harness the information acquired of the consumption of energy of the users in order to support DT through the analytical capacity that owns on both levels: Descriptive, using software that can generate early warnings to recognize irregularities and Predictive, composed of analytical models to address the prevention and detection of anomalies in energy consumption. The technology is framed in two aspects: Information and Communications Technology (TIC acronym in Spanish) and Smart Grids (SG) in order to achieve a network of secure, reliable, efficient and sustainable distribution, composed by infrastructure of information, communication, management, control and protection.

II. Problem

High rates of NTL in Latin America (LA), affect the finances of DT negatively, as well as energy efficiency, quality of utilities, industries and citizens[3]. These losses are related to: badly billed consumption of energy, short circuits, networks and overloaded facilities, forcing to invest in renovation and adjustment of assets. NTL lead to: deterioration of facilities by illicit appropriation of energy with disastrous consequences for public safety; the impossibility of using electrical appliances during the hours of maximum demand of energy because the voltage in these zones is less than the permissible; risk of tragedies due to short circuits; detriment of the meters by improper handling of these which leads to changes or repairs to normalize them. The illegal action occurs in all social strata and even in the industry and commerce where the manipulation of the systems of measurement is technically more qualified. The Fig.1 shows the graph of the behavior (in percent) of the index of energy losses between 2005 and 2011 for some countries in LA, according to the latest report offered by the World Bank (WB), taking as reference the average in Latin America and the Caribbean, East Asia and the Pacific, South Asia and globally.

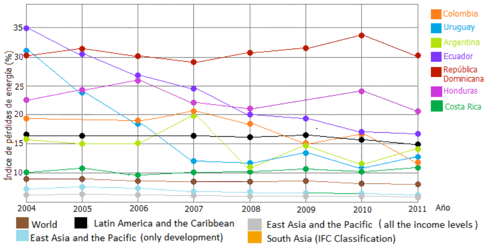
III. Solution

Monitoring of energy consumption and the control of the devices located in the area of the citizen require the presence of appropriate communication technology and early warnings system for eventualities.

G.A. Quintero, Innovación Area, MVM Ingeniería de Software, Medellín Colombia (e-mail: gladys.quinte@gmail.com).

R. A. Gallego, Innovación Area, MVM Ingeniería de Software, Medellín Colombia (e-mail: ricardo.gallego@mvm.com.co).

978-1-4673-6605-2/15/$31.00 ©2015 IEEE



Fuente: The World Bank. http://data.worldbank.org/ondicator/EG.ELC.LOSS.ZS/countries/1W-ZJ-LA-Z4-C8-4E?display=graph

Fig.1. Performance of the index of energy losses to Argentina, Ecuador, Uruguay, Dominican Republic, Honduras, Costa Rica between 2005 and 2011. Source: WB

Techniques to achieve the quality and efficient distribution of energy are constantly evolving such as evidenced in: The structuration; the equipment and tools used; materials with which the networks are built; working methods of the *cuadrillas*, the methodology of design, the operation and the integration provided by the software used.

MVM has used its knowledge and capability in advanced analytic to develop a technological solution that support the decision making in situations of fraud and failures in the distribution network, using models of quality established by the software industry.

IV. Theoretical Framework

The fraud is associated with the NTL and is defined in the electricity sector as the illegal action carried out by the user when manipulates the meter, in order that this records lower rates of energy consumption than the actuals, action that leads to an erroneous billing of such consumption. Energy losses are defined as the difference between the energy supplied and energy invoiced or consumed by the user. They are divided into: Technical Losses (TL) and TNL[4], these latter are related with the electricity theft and incorrect administrative processes. The fraud also includes the bribe and illegal connections.

Data mining is a process that has as main purpose to analyze, extract and store relevant information from large databases that contain energy consumption records; it is a combination of databases and artificial intelligence technologies[5] and is an evolving discipline. It is a process of searching and extraction of patterns of large amounts of data by using reasoning techniques[6]. The information extraction process involves feedback cycles, which make possible to determine the quality of the selected data and the effectiveness of the technique applied.

This project uses as source of information load profiles of residential customers (consumption patterns) in order to construct and characterize behaviors through actual consumption patterns per time period and quantify results using computational tools that define the normal or suspicious state of the account studied. The extraction process of information involves a feedback cycles variety, because the application of a particular technique can lead to the conclusion that the selected data are of poor quality or that the technique applied is not appropriate. In such cases, the company has to refine and repeat the steps performed or possibly restart the entire process.

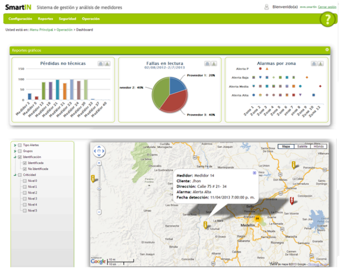
The data mining process consists of several stages: data preparation (sorting, cleaning and processing), exploration and auditing, data mining as such (development of models and data analysis), evaluation, dissemination and use of models (output). In addition, the extraction process incorporates different techniques (decision trees, linear regression, artificial neural networks, bayesian techniques, support vector machines)[7].

V. Technology Developed

DT require to collect, organize, manage and analyze large amounts of data[8], which implies the application of various analytical methods in large volumes of structured and unstructured data[9]-[10]-[11], that provide results in descriptive and predictive levels. In the predictive context companies need to use analytical tools that enable them execute actions to face contingencies in the environments: user needs, market trends, demand management and energy purchase, movements of competition and NTL[12]. MVM has identified aspects of this problem in LA through competitive intelligence and technological surveillance[13].

A. Descriptive Analytic Level

Software that allow generating early warnings to identify irregularities in energy consumption; optimize recovery processes incorrectly billed services in DT, provided they act after alerts; identify communes causes of abnormal energy consumption and analyze behaviors from the collection, organization, analysis and visualization of generated information in the process of measuring consumption of regulated and unregulated users[14]. Their main features are: capture and integration of information from different sources; customized configuration to set parameters and thresholds for the analysis; identifying of suspicious situations; web architecture; multiservice architecture; multiteam and multiprotocol; capture of information from meters that operate under the DLMS / COSEM [11] protocols[15]; visualization of the warning information through a geographic information system (GIS). Fig. 2, illustrates the dashboard interfaces of the software.


SmartIN Sistema de gestión y análisis de medidores

Usted está en: Menu Principal > Operación > Dashboard

Medidor: Medidor 14
Cliente: Jhon
Dirección: Calle 75 # 21- 34
Alarma: Alerta Alta
Fecha detección: 11/04/2013 7:00:00 p. m.

Fig.2. Dashboard of the software. Fuente: MVM.

B. Predictive Analytic Level

It is composed of analytical models that generate information about users who are committing some kind of irregularity and at the same time allow establishing efficient actions of intervention in field. Each project is approached using the methodology of the Cross Industry Standard Process for Data Mining (CRISP-DM)[16] which is an iterative process model developed in1996; such methodology was used in this project.

This methodology comprises of the following tasks: business understanding; data understanding; data preparation; modeling; evaluation and deployment. Generally, these tasks follow each other as subsequent phases, but within this stream, many iterative cycles can be observed.

The second CRISP-DM phase or data understanding consists of data formatting, description, exploration and verification of quality.

The data exploration of CRISP-DM supports the analyst in creating hypotheses about the data relationship and scattering. The database query and visualization techniques help to formulate defined goals of DM and reveal interesting patterns.

The phase modeling methodology of CRISP-DM includes the application of different DM methods and of knowledge discovery with wide scale of tunable parameters each other. These methods can be grouped into different categories depending of used algorithms.

The problem of the business is analyzed with the data and the acquired information; strategies to be adopted are defined in each specific case. After this phase, the baseline of preestablished business questions is utilized to determine the scope of the models to be implemented; subsequently, analytical models described in the Table 1, are designed using the baseline of these.

TABLE I

Description of baseline of some analytic models

Source: MVM.

Model	Advanced Analytical Technique
Singular Values [17]	• It allows the cleaning and estimation of missing data.
Synthetic Minority Over-Sampling Technique [18] (SMOTE)	• Balancing classes: Often, the datasets of the real world are predominantly composed of "normal" examples with only a small percentage of "abnormal" examples. The proposal is oversample the minority class by creating "synthetic" examples operating in the feature space instead of the data space by their introducing along line segments that unite any or all the nearest neighbors the minority class k; depending on the amount of over-sampling required, neighbors of the k nearest neighbors are chosen randomly. For example, if the amount of oversampling is 200%, only two of the five neighbors nearest are selected and a sample is generated in each direction. Synthetic samples are generated taking the difference between the feature vector (sample) of study and its nearest neighbor. This difference is multiplied by a random number between 0 and 1 and is added to the function of vector into consideration. This helps the selection of a random point along the line segment between two specific features.
Analysis BI-Clúster [19]	• It has a group approach. Allows the identification of subgroups of rows and columns in the matrix of observations, such that are as similar as possible among them and so different as admissible from the rest. Enables the identification of various types of irregularities in the energy consumption. Inside of categorical data, offers a set of groups linked to pairs of objects and attributes of value (bicluster). It is proposed, in this case, a generic framework for bi-cluster that allows calculating a bi-partition of the collections of local patterns for local capture of strong associations between objects and properties.
Cluster Analysis of K-Means [20]	• It is based on the iterative clustering in terms of the distance. This algorithm takes the input parameter K and allows to partition a set of n objects into groups K, such that the intra-group similarity resulting is high, but between groups is low, which can be measured relative to the average value in a group known as gravity center of the cluster [22]. The purpose of this comparison is determining if the ensemble mean of samples is significantly greater than or less than the average of other, based on a set of variables specified. The solution enables the identification of the users that are committing some kind of irregularity.
Decision Trees [21]	• Empowers the characterization of the users belonging to each of the datasets analyzed. It is based on classification models. Allows the identification of the variable that has the highest weight in the classification of the users that are committing some kind of irregularity. Such models correspond to a set of structured examples of categorical variables with some non-categorical variables called inputs and one categorical variable or class which is the output. Input variables can be continuous or discretes and in general binary type. Output assumes values 1 or 0, which means that it belongs or does not belong, respectively, to that category. The problem is to find a model to classify non-categorical data correctly.
Adaptive Power [22]-[23]	• It allows associating a weight to each set of data studied.
Support Vector Machine [24]	• Feature extraction tool for the detection of fraudulent users in energy residential distribution systems, where the fraud is one of the principals causes of loss of revenue for many DT. The problem is centered in identifying of the fraud. The problem is centered in identifying fraud and is primarily based on supervised learnings that require early stages of training. Classification models used must ensure high precision regardless of the number of load profiles to study
Supervised Method Nearest Neighbor [25] (K-NN)	• It is a method of supervised classification nonparametric to detect irregular patterns of energy consumption by fraudulent users. Learning is based on a training set and prototypes. In the learning process is not made any assumption about the distribution of the predictor variables. In recognizing patterns, the K-NN algorithm is used as method for classifying objects, based in training through examples in the space near to the elements. K-NN is a type of slow learners, where the function is approximated locally and the entire calculation is deferred to the classification. This method assumes It is a method of supervised classification nonparametric to detect irregular patterns of energy consumption by fraudulent users. Learning is based on a training set and prototypes. In the learning process is not made any assumption about the distribution of the predictor variables. In recognizing patterns, the K-NN algorithm is used as method for classifying objects, based in training through examples in the space near to the elements. K-NN is a type of slow learners, where the function is approximated locally and the entire calculation is deferred to the classification. This method assumes that the nearest neighbors offer the best classification using all attributes. The problem is in having many irrelevant attributes dominating the classification, two important attributes would lose weight among twenty irrelevant attributes. The correction of possible bias admits to assign a weight to the distances of each attribute, giving greater importance on the most relevant. Another possibility is to try to determine or adjust the weights with known training examples.

C. Incorporation of Technology

In the context of AMI, operators have incorporated smart meters in the network infrastructure which generate large volumes of information to be managed, therefore DT must use advanced analytics to transform such information into knowledge. The TS presented in this paper, takes into account the principle of big data from research carried out within the company [26].

VI. Activities for the Validation of the Solution and Results of Their Application in the Context

The Empresa de Energia de Cundinamarca - EEC is carrying out the validation activity of the TS for NTL in the context of its business.

The first result has been the validation and use of business questions to determine needs to be resolved with the TS from its analytical capacity identifying its basic, advanced, predictive and descriptive level and type of analysis to perform (energy losses, billing, revenues, energy consumption, irregularities) which has generated benefits to said company, because it speeds and precise the actions to undertake in the context of NTL. This baseline of business questions is a point of value of the TS because it allows deal quickly and effective, problems facing these businesses. Some of the business questions that are resolved by the ST are presented below:

A. Business Questions about the Analysis of Energy Losses and application of meter data management (MDM).

1): *Energetic Balance:* What is the energy balance per: period, circuits, subsystems and transformer?

2): *Percentage of Total Losses:* What is the percentage of total losses of the company in KWh (and $) per: geographical zone and time period?

3): *Technical and Non-Technical Losses*: Which are the technical and non-technical losses of energy of the company in KWh (and $) per: geographical zone and time period?

4): *Geographical Area:* Which are the geographical areas with higher losses per time period?

5): *Maximum Loss:* What is the maximum loss (KWh) per: transformer (brand) and time period?

6): *Maximum Value of the Losses Index:* What is the maximum value of the losses index per: transformer (lost energy / input energy) and time period?

7): *Type of Alarm:* Which types of alarms have been presented at the macro-measurement level per geographic zone?

B. Business Questions about Analysis of Energy Losses and Data Mining (DM) Application

1): *Fraudulen Customers:* Which customers may be future fraudulent users?

2): *Demand of Energy*: What is the demand of energy (in KW and $) projected for future periods?

C. Business Questions about Analysis of Billing /Revenues; Energy Consumption and Implementation of Bussines Intelligent (BI)

1): *Revenues: What are the revenues regarding to billing per: geographical zone, time period, social stratum and customer type?*

2): *Compliance Indicators:* What is the performance of compliance indicators of billing and revenue per supply of energy?

3): *Accounts:* Which are the accounts that have higher portfolio per: geographical area, macro-meter and customer type? and how is their relation with the level of losses in the zone?

4): *Consumption:* What is the consumption per: customer, geographic zone, stratum, time priod, type of account, measuring equipment?

D. Business Questions about the Analysis of Energy Consumption and MDM Application.

1): *Historical of the consumption:* What is the *historical of the energy consumption* registred per: measured equipment and time period?

2): *Record above Threshold:* Which meters associated with the x macro-meter have exceeded the record threshold of energy energy consumption?

3): *Overconsumption*: Which clients have exceeded consumption rules per time period?

E. Business Questions about Analysis of Irregularities and DM Implementation

1): *Amount of Irregularities*: Which is the amount of irregularities per: account, service number, social stratum, type and model of meter, type of irregularity and geographic zone?

2): *Field Visits:* What is the effectiveness of field visits supplied by the predictive model per: time period, geographical zone, social stratum and customer type?

3): *Type of Irregularity:* Which are the types of irregularities that occur frequently per geographical zone, social stratum and customer type?

4): *Planned Visits:* What is the effectiveness in the implementation of planned visits per time period, geographical zone, social stratum and customer type?

F. Business Questions about Analysis of Irregularities and BI Implementation

VII. Results of the Benefit of the TS in the Validation with a Real Client

A. According to the validation of the ST in the EEC, it identifies that this complies with the following features and are of high value for this company:

1): *Energy Balances:* It is carried out in terms of macro-meters, circuits or subsystems and if possible with georeferencing.

2): *Intelligent Alarms:* With georeferencing to detect anomalies in the network.

3): *Information Analysis Models:* For identification cause-effect relationship in matters such as: consumption, energetic balances, billing, revenues, portfolio and field visits.

4): *Optimization Model:* For the logistics that intervenes in field.

5): *Classification Model of Irregular Users:* Via some variables and with the possible adaptation of others. The classification model delivers a list of possible clients classified as irregulars.

VIII. Activities to implement analytic of losses

A. Analysis of the Energy Balance: With analytical reporting (OLAP) per: time period, geographical zone and node and in the dimensions of: time, geographical zone and node.

B. Analysis of the Percentage of Total Losses of the Company (KWh): With analytic reporting (OLAP) per: geographical zone and time period and in the dimensions of: time, geographical zone and node.

C. Calculation of the Losses Index: With indicator per: transformer (lost energy / input energy) and time period and in the dimension of time.

D. Calculation of the Density Index of Losses: With indicator per transformer and time period and in the dimension of time.

E. Listing of the Upper Limit of Transformers with Losses (KWh): With reporting and consulting per: time period, characteristics of the transformer and geographical zone and in the dimensions of: time, geographical zone and node.

IX. Efficiency indicators

A. Effectiveness Rate of Visits in a 20%: For identification and location of fraudulent customers. To date has achieved an improvement in this indicator.

1): *Description:* Energetic balances, critical of the information in the context of NTL and the integrated analysis of data from various sources such as CRM, ERP, billing systems, smart meters and macrometers, are made through the descriptive and predictive analytics to identify possible fraudulent customers, in order to program, effectively, visits of *cuadrillas* in field.

2): *Sources of Verification:* Analytical models, datamart of losses and cubes: DM, pending payments, field visits and BI loss model.

B. Operational Efficiency: Percentage reduction in operating costs.

1): *Description:* The application of analytical models allows optimizing the profitability of field operations for identifying of fraud suspect users. With the solution, companies can achieve 35% effectiveness in identifying fraudulent users.

2): *Sources of Verification:* Analytical models, datamart of losses and cubes: DM, pending payments, field visits and BI loss model.

C. Reduction of the Index of Losses: Making decisions based on the likelihood of fraud from a customer allows reducing the percentage of the index of NTL.

1): *Description:* Making decisions in order to reduce the percentage of the NTL index takes into account variables: geographic location, transformer and irregularities user.

2): *Sources of Verification:* analytical models, datamart of losses and cubes: DM, pending payments, field visits and BI loss model.

D. Efficiency in the Visualization of events recorded in the meter in real time: The visualization is an attribute adjustable to the solution from the prototype designed to add this characteristic.

1): *Description:* It may be performed, using the advanced metering infrastructure (AMI), sources and information systems of the objective company. Evaluated results will be: The performance and response time between the reporting of the event, its analysis and visualization.

2): *Sources of Verification:* Dashboards and reporting.

E. Efficiency Percent in Data Analysis: Analytical capacity in balances and critical of the information.

1): *Description:* It facilitates the process of decision-making for routing the actions.

2): *Sources of Verification:* Cubes: DM, pending payments, field visits and BI loss model.

X. Conclusions

This technology, at both levels, strengthens the analytical capacity of DT and enables decision making based on quantitative analysis. The base line of analytical models allows: the examination of consumption and variables of interest to generate effective routings in the field visits; understand what happened, what is happening and what will happen and the relationship between the data; the grouping of the types of irregularities which provide better exploration, generating indicators and value reports to the user; elements for balance analysis or critical to reports; tools for automation of data extraction process; analysis of relevant variables which allow to realize specialized studies and improve the critical to information.

XI. Acknowledgment

For MVM managers, who have had the commitment and entrepreneurial conviction that technological development and innovation can generate high impact solutions to its stakeholders.

XII. Biographies

Gladys Adriana Quintero Rojas was born in Envigado-Colombia. PhD in Physics of the Pontificia Universidade Católica do Rio de Janeiro (PUC - Brasil), M.SC in Physics of the Universidad de Antioquia (UdeA - Colombia), Specialist in University Teaching of the Universidad Cooperativa de Colombia (UCC). Professor at the University of Antioquia and member of the Research Group: Optics and Photonics (GOF), with "A" category of COLCIENCIAS. Researcher in the area of Innovation of the Company MVM Ingeniería de Software by Administrative Commission granted by the Universidad de Antioquia, through the project: "Development of a prototype product or service to support a solution from a smart grid", approved by the call 535 of COLCIENCIAS: Inserting Doctors in a Company.

Ricardo Alonso Gallego Burgos was born in Colombia. MSc. in Technology Management of the Universidad Pontificia Bolivariana (UPB-Colombia), Information Management and Database Systems Specialist, Systems Engineer of the Universidad de Sanbuenaventura (Colombia), Innovation Director at MVM Ingeniería de Software. Professor of Specialization and Master programs at UDEM, UPB, USB (Colombia). Conference Speaker in academic, industrial national and international events, in areas: Knowledge Management and Innovation. Member of Research Groups: Arkadius of the Universidad de Medellín and GTI of the Universidad Pontificia Bolivariana

XIII. References

[1] I. Vásquez. (2005, Dic.). Tipos de estudio. Universidad Nacional Federico Villarreal. GestioPolis. [Online]. Available in: http://www.gestiopolis.com/canales5/eco/tiposestu.htm.

[2] A. Peña. (2006). *Inteligencia de Negocios: Una Propuesta para su Desarrollo en las Organizaciones*. Instituto Politécnico Nacional. México. (Primera Edición).

[3] C. N. I. Casa and C. M. G. Suncha. "Control y reducción de pérdidas no técnicas de energía mediante el método balance de energía por transformador en 19 sectores de la provincia de Cotopaxi designados por ELEPCO S.A.". Tesis. Electromechanical Engineering Universidad Técnica de Cotopaxi. Latacunga. Cotopaxi. Ecuador. Ed. UTC. 135p. 2009.

[4] A. H. Nizar, Z. Y. Dong, and Y. Wang, "Power Utility Nontechnical Loss Analysis with Extreme Learning Machine Method", IEEE Transactions on Power Systems, vol. 23, no. 3, Aug. 2008, pp. 946-955.

[5] L.S. Bora. 2011. "Data mining and ware housing", Technical report, Department of Computer Science and Application, IEEE 2011 Chandrapur, India. 2011. pp 1-4

[6] B. Thuraisingham. "Data Mining for Malicious Code Detection and Security Applications", IEEE Conference on Web Intelligence and Intelligent Agent Technology-Workshop, 2009. Pp 1.

[7] C. Pérez López (2008). *Minería de Dato Técnicas y herramientas*. (Primera edición).

[8] S. Kaisler, F. Armour, A. Espinosa, and W. Money (2014). Big Data: Issues and Challenges Moving Forward. System Sciences (HICSS), 2013 46th Hawaii International Conference on Pp 995 – 1004.

[9] M. Adoración de and M.G. (2001). "Diseño de bases de datos. Problemas resueltos". Madrid: RA-MA.

[10] El Rincón del Vago. Conceptos de bases de datos. España

[11] A. Silberschatz, (2000). "Fundamentos de bases de datos". Madrid: McGraw-Hill.

[12] Costa, E. O., Fabris, F., A. L. Rodrigues, H. Ahonen, F. M. Varejao, R. Ferro, 2013. "Using GA for the stratified sampling of electricity consumers. In: Evolutionary Computation (CEC), IEEE Congress on, pp. 261-268, 2013

[13] MVM Ingeniería de Software S.A.S. y Universidad de Medellín (2014). Informe de vigilancia tecnológica e análisis de mercado: Pérdidas no técnicas en Colombia. Dirección Gestión del Conocimiento e Innovación (MVM)-Centro de innovación y Desarrollo Empresarial Vicerrectoría de Investigaciones (UDEM), 1-114.

[14] Enertolima. Mercado Regulado y No Regulado. MedioDigitales.

[15] G. Štruklec and J. Marši (May. 2011). "Implementing DLMS/COSEM in Smart Meters". *2011 8th International Conference on the European Energy Market (EEM) • 25-27 May 2011 • Zagreb,* Croatia. Pp 747-752

[16] Z. Bošnjak, O. Grljević and S. Bošnjak (May. 2009). "CRISP-DM as a Framework for Discovering Knowledge in Small and Medium Sized Enterprises' Data". *In 5th International Symposium on Applied Computational Intelligence and Informatics May 28–29, 2009 – Timişoara, Romania. IEEE*. Pp 510-514.

[17] J. Martínez y C. L. Vergara. *Minería de Datos Para Predicción de Riesgo de! Compras en Retail. Preparación de los Datos.* Universidad de Chile – Departamento de Ingeniería Industrial.

[18]] N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, Vol 16, 321-357.

[19] R. G. Pensa, C. Robardet and J. F. Boulicaut. "Bi-clustering Framework for Categorical Data". pp. 643–650.

[20] A. H. Nizar, Z. Y. Dong and J. H. Zhao, (2006). "Load Profiling and Data Mining Techniques in Electricity Deregulated Market" IEEE. Pp 1-7.

[22] S. Agarwal, "Data mining: concepts and techniques". 2013 International Conference on Machine Intelligence Research and Advancement. IEEE. Pp 203-207.

[21] J. Reis, E. M. Gontijo, A. C. Delaba, E. Mazina, J. E. Cabral y J. Onofre. Fraud Identification In Electricity Company Costumers Using Decision Tree.

[22] J. Reis, E. M. Gontijo, A. C. Delaba, E. Mazina, J. E. Cabral y J. Onofre. Fraud Identification In Electricity Company Costumers Using Decision Tree.

[23] J. Reis, E. M. Gontijo, A. C. Delaba, E. Mazina, J. E. Cabral y J. Onofre. Fraud Identification In Electricity Company Costumers Using Decision Tree.

[24] F. Rios y K. A. Uribe. 2013. "Minería de datos aplicada a la detección de clientes con alta probabilidad de fraudes en sistemas de distribución". Universidad Tecnológica de Pereira Facultad de Ingenierías. Programa de Ingeniería Eléctrica.

[25] Dasarathy B. V., Sánchez J. S. (2000). "Tandem Fusion of Nearest Neighbor Editing and Condensing Algorithms. Data Dimensionality Effects." pp. 692-695.

[26]] J. Giraldo, (2014). "Estrategia de gestión de grandes volúmenes de datos para la medición inteligente en el contexto de Smart Grid: Caso MVM Ingeniería de Software S.A.S." Universidad EAFIT, Medellín, En revisión. Universidad EAFIT.