Association Rules in the measurement of air pollution in the city of Santiago de Chile

Santiago Zapata Caceres
Department of Informatics and Computation
Engineering Faculty
Metropolitan Technological University of Chile
Santiago, Chile
szapata@utem.cl

Juan Torres Lopez
Department of Informatics and Computation
Engineering Faculty
Metropolitan Technological University of Chile
Santiago, Chile
jtorres@utem.cl

Received: February 20, 2021 Revised: March 12, 2021. Accepted: March 15, 2021. Published: March 29, 2021.

***Abstract*—some time ago, the use of the computation was operating alone. In the 1970s there is a change in the mindset of companies and organizations. Are recognized continuously recorded data as the raw material that would lead position in the market. The needs changed, requires additional storage capacity, data processing, new tools to address the information available. One such tool is known as Knowledge Discovery in Databases (KDD).**

In this paper we shall describe the main difficulties encountered in the discovery process either inherent fears or make a change that arise in data from different data sources. Finally, we present a practical application using data from Air Quality in the city of Santiago de Chile from the years 2000-2012, recorded at different monitoring stations intended for this purpose, and seek to establish a relationship between materials PM10 and PM2.5 particulate concentrations for 24 hours.

Keywords:* *Association Rules, Data Mining, Causation, Extraction of Knowledge (KDD), Environmental Contamination

I. Introduction

The work is intended to apply the tools of Knowledge Discovery in Databases and emphasizing the importance of the process of acquiring new knowledge in order to support decision-making in companies, organizations and institutions, and create the conditions to improve the decision-making process.

In the project development explains the sequence of steps to be executed for the purpose of exploiting the data, different algorithmic approaches that are part of the process

The work devotes more attention to one of the data mining techniques, which corresponds to the Rules of Association, explaining the different types of rules that exist, ending with Fuzzy Association Rules, indicating its importance at the time of search association rules in databases, in which the values of their attributes are numeric and categorical

KDD application applies to data collected by the measurement stations of air quality in the city of Santiago de Chile.

In the work is made a practical application of KDD process, with data collected from the RED MACAM (Automatic Monitoring Network Air Pollutants), from which we obtained data from the years 2000-2012 of various air pollutants, including particulates, troposphere ozone, and carbon monoxide. With data mining tool Clementine, using a MySQL database, you work with the data to establish a relationship between particulate materials through scatter plots, and other graphical tools.

II. Work Development

A. Objectives of the work

Emphasize the importance of information in today's society and the benefits provided by the use of tools such as Knowledge Discovery in Databases to perform exploratory data process generating benefits studied research field. Similarly, applying the tools provided by data mining in the process of acquiring and generating knowledge.

To achieve these objectives, we made a practical application of data mining on a data set from the Metropolitan Health Service Environmental (SESMA) indices corresponding to level of contamination detected in the city of Santiago de Chile, captured at stations Air Monitoring in the years 2000-2012 using the data mining software called Clementine and MySQL database storage level of detected contaminants.

The purpose of the application is to extract association rules from data stored in a MySQL database that records information related to environmental pollution levels of various pollutants and seeks to establish a relationship between particulate materials capable of determining the level of relative hazard of PM2.5 to PM10.

B. Problem

The pre-emergencies in Chile are established when the pollution levels exceeds the values indicated in the environmental law, this has a serious impact in the commerce, and in some critical cases it could become a sanitary emergency, because it’s dangerous to the people’s health, therefore it’s needed to count with methods capable to study the historical situation of the Santiago’s basin, which have characterized for having trouble with the pollution and the dust in suspension, and so to predict anomalous ventilation situations, that permits the authorities to act efficiently.

C. Current Prediction and measurement model

Currently, a predictive model created by Joseph Cassmassi is used the model was developed from the air quality information measured by the Automatic Monitoring Network of air quality (MACAM II Network) and the tall meteorological information from the central zone of the country.

The forecasting methodology of MP10 concentrations is based in calculus algorithms developed by applying statistical techniques of multiple regression variables, focused in find relations between possible predictor variables and a variable to predict. The possible predictions include observed weather variables, observed weather condition indexes, observed concentrations and expected variations in rates of emissions.

The Cassmassi model forecast the maximum value of average concentration in 24 hours of breathable particulate material (PM10), forecasted for 00-24 h period of the following day, expressed in (ug/m3), in each one of the stations of the MACAM 2 Network classified as PM10 Monitoring Stations with demographic representatively (EMRP). These are: Av. La Paz, La Florida, Las Condes, Parque O'Higgins, Pudahuel, Cerrillos and El Bosque, according with the resolution N°11481 of 1998 from SESMA

The forecasted concentration for the next day is calculated by different equations for each air quality monitoring station. The required variables for the equation solving are obtained from the related information with the expected conditions shifting by day of the week, from the PM10 concentrations measured in the MACAMII Network, from tall meteorological information obtained from the radio probes realized by the Weather Direction of Chile and the weather conditions of synoptic and regional observed and forecasted scale for the region.

The operational application of this methodology considers two prediction algorithms for each monitoring station. A first algorithm includes the index of meteorological potential, forecasted for the next day. The second algorithm is based in observations only (same day and previous day). That way, if the first algorithm cannot be applied, the second one is used.

With the model previously described, environmental measures are enacted according with the following table:

ICAP Level	Air Quality	Enacted Measure
0-99	Good	None
100-199	Regular	None
200-299	Bad	Environmental Alert
300-399	Critical	Pre-emergency
400-499	Dangerous	Pre-emergency
500 or more	Exceed	Emergency

Table II.1: ICAP Levels and Environmental Measures

For example: on the level 299 is enacted "Environmental laws during that day, however, the air quality in level 299 is not very different that level 300. We believe that the different levels must be replaced by linguistic variables with pond rated values by degrees of truth Warning", but in 300 "Pre-Emergency" is enacted (special and more restrictive). All this, so the implemented system can become a decision-making support tool at the time of enacting environmental measures.

III. Application of KDD data to air quality

Application is done using the Clementine tool to apply the methods, generate models and discover the relationships among data, we used the management system MySQL database in order to create a data warehouse that allows storing operational data concerning Pollution Air of the City of Santiago de Chile, from Automatic Monitoring network of Air Pollutants in the period from 2000 to 2010, taking the following metadata: CO - Carbon monoxide, PM2.5 - particulate matter 2.5, PM10 - particulate matter 10 NOX - Oxides of Nitrogen, O3 - Troposphere Ozone, SO2 - Sulfur Dioxide.

A. History

For several years there is continuous monitoring of the air quality both in Chile and in other regions of the country. The first air monitoring stations were established in 1964 in order to calculate the particulate materials blackening and acidity of gases, later joined by measurements of total suspended particulates (TSP), sulfur dioxide and nitrogen dioxide, which were done through a network of semi-automatic quality monitoring. In 1988 he began to evaluate the PM10, CO, SO2, NOX and O3 with automatic monitoring network of five stations and a central data capture. In 1997, the network expands to eight stations online, while monitoring is performed with two portable stations, and based on the Basic Law declares environment Santiago as saturated zone for PM10, TSP, CO and Ozone and SO2 latent area. For this reason they are generating plans that help the decontamination time to time adopted new emission control measures and restrictions on certain activities.

B. Description of Pollutants

- Particulate Matter PM10

  One of the pollutants that cause more damage to health is PM10 Particulate Matter, that reaches the atmosphere through various sources, and whose degree of risk varies depending on the source that emits.

- Particulate Matter PM2.5

  The PM2.5 particulate matter causes more damage than the MP10 due to its smaller size. He is responsible for the deterioration of human health as a percentage of between 50% and 70%. Its small size facilitates entering houses remain suspended a greater amount of time in the atmosphere, entering the lungs, damaging the defense mechanisms of people causing respiratory infections, cardiovascular disease and even causing death.

- Carbon Monoxide (CO).

  Carbon monoxide is produced by incomplete combustion of natural gas or carbon containing products (kerosene, oil, etc.), Car engines, stoves, and portable heating systems for indoor exhaust pipe cars, trucks or buses etc. It causes thousands of deaths in North America and is the leading cause of deaths from poisoning by inhalation of this gas.Nitrogen Dioxide (NOx)

  Exposure to Nitrogen Dioxide generates acute and chronic effects on the health of individuals, studies confirm the WHO (World Health Organization) and the Environmental Protection Agency (EPA). Damages result in irritation of the lungs and / or reduced resistance to respiratory infections in people with overcoats asthmatic problems.

- Troposphere Ozone (O3)

  This gas is generated on sea level, resulting from the burning of fossil fuels (gasoline, natural gas or carbon, etc.), so a greater volume is in large cities and industrial areas, is injurious to health and the environment, in times of high temperatures generated by the so-called "photochemical smog" when combined with car exhaust and factories.

- Sulfur Dioxide (SO2)

  Exposure to this pollutant acute and chronic affects on the health of people, so their environmental emergencies concentrations defined in an hour, and in areas surrounding the issue. This minimizes lung capacity so that its effects are amplified with physical activity, with hyperventilation when breathing cold, dry air, or in people with bronchial hyperactivity.

C. *Calidad del aire*

Measuring the Air Quality Index for particulate matter (ICAP) in Chile to measure air quality, for a long time was considered the MP10 as the most important source of pollution, in fact, their average concentrations exceed 24 hours in Santiago often considered normal, however, new studies indicate that PM2.5 is the most dangerous due to their small diameter which allows you to have a greater presence in the environment, and also easier to get into the bloodstream.

D. *Data Collection*

The collected data delivery rates air quality referred to particles (ICAP), this comes from different monitoring stations spread over different districts of Santiago de Chile, with temporary archive and a measurement period variable depending on the associated station. Also, the presence of pollutants, determines the impact of contamination in the field of measurement. These indices are not captured by all monitoring stations.

E. *Preparation of the data*

The data coming from different sources, require a treatment to maintain consistency for further manipulation, the process is carried out through Clementine software functionality, which makes access to the data and brings together in a single file, maintaining for each attribute its original data type. Initially the data are recorded in a database buffer, and then a series of transformations to the data, populate the data warehouse created with historical data from environmental pollution levels detected by the various monitoring stations distributed in different districts of the city Santiago de Chile, during the period 2000-2012.

To maintain relative normality of the data, these are subjected to a validation process by the Metropolitan Health Ministerial Secretary, who is running a background check of the network stations (power outages, filter changes monthly or bimonthly telephone network problems, preventive maintenance and corrective maintenances, calibration of analysis equipment, etc.) in order to obtain operational data validated.

F. *Visual Data Exploration*

In the knowledge discovery process is important to have a vision about the information they can provide data prior to handling, which is why using tools to visualize data (scatter plots, histograms, etc.).

1) *Monitoring infrastructure.*

For this purpose, generate graphs with Microsoft Excel tool that allows visual comparison between PM10 and PM2.5 Particulate Matter. Because in 2000 there was no law that regulated the concentration of particulate matter PM2.5, graphics include the current standard for the material PM10 in micrograms per cubic meter, which as explained above has to PM2.5 by considered all aerodynamic diameter less than 10 microns.

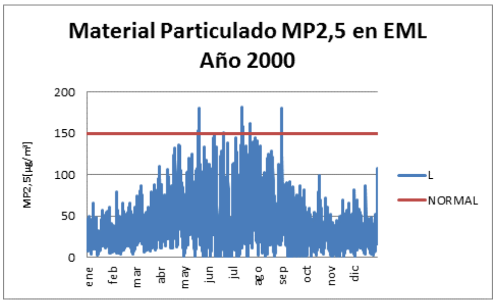


Figure 1: Measurement of Particulate Matter PM2.5 in EML Monitoring Station (L - Florida) for 2000.

This measurement generally remains within the established standard for the pollutant reaching the highest PM10 concentration levels between mid-April and late September, with peaks of overcoming norm between the months of May and September.

2) Scatter Chart 5.6.2 MP10 vs MP2, 5

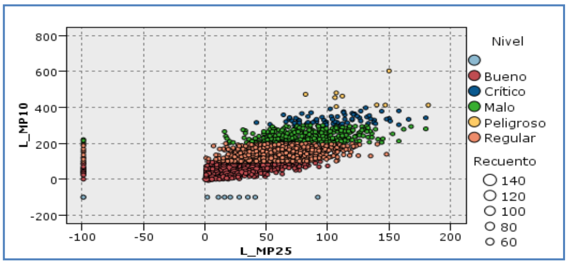


Figure 2: Graphic dispersion of pollution levels PM10 and PM2.5 Particulate Matter in Monitoring Station of Florida (L).

Figure 2 illustrates the greater danger of the PM2.5 particulate material of MP10 opposed reference to the pollution levels regulated latter. It can be appreciated that the agent reaches PM2.5 levels considered bad, with indices below 25 g/m3, while the other requires levels greater than 200 g/m3, and even with an amount close to 50 g/m3 begin to appreciate a MP10 dangerous levels above 300 g/m3, the graph shows the most dangerous of the material against the PM10 PM2.5.

This results, define a trend graph in each of the monitoring stations, lower rates of contaminant compared MP2.5 MP10, but with the same degree of hazard.

3) Particulates relationship Pollution Levels and Monitoring Stations

In Figure 3 shows the graphs on regular days and good days through MP2.5/MP10 rate, which shows that it is higher in the good days unlike what happens in the days when the level quality is considered hazardous, it is repeated that the contamination levels of both particulate materials reach a certain level of similarity. According to the graphs shown, MP2.5 contamination levels are in a lesser degree in the environment, unlike what happens with MP10, therefore when their concentrations are similar levels of contamination are reached critical dangerous bad or exceeded.

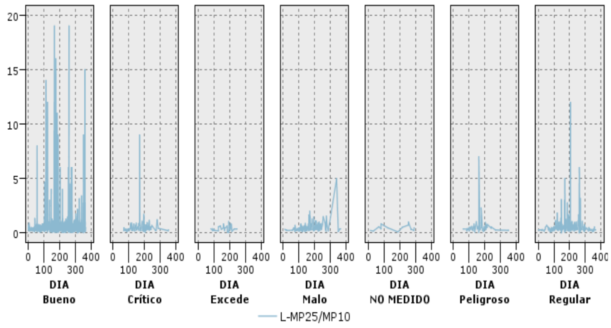


Figure 3: Relationship between particulates for Monitoring Station in Florida

4) Contamination levels Meshes

As can be seen in Figure 4, corresponding to that recorded in the monitoring station of Florida in 2000, the rates achieved during travel the regulated pollution levels, however the good level is what is repeated more frequently, and in the winter and summer months, which can be explained by higher rainfall levels that take place in the winter months, and decreased during the summer polluting vehicles, keeping a lesser extent the concentration of particulates.

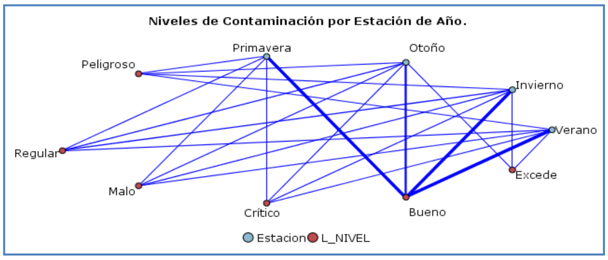


Figure 4: Levels of contamination by season in 2000.

Figure 5 shows the monthly record of concentrations of particulate matter in 2000, shows that concentrations "dangerous", "beyond", "reviews" and "bad" are repeated throughout the year except in the month of November, although in any month was a greater frequency of a certain level of contamination.

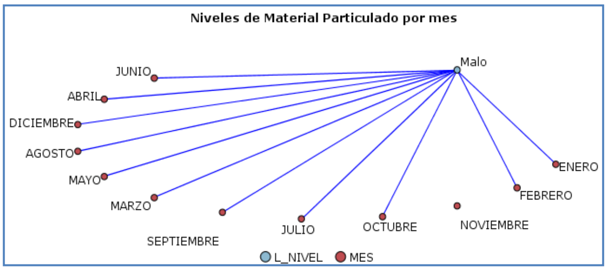


Figure 5: Months of 2000 that are recorded pollution levels considered "bad"

a) Mesh critical levels

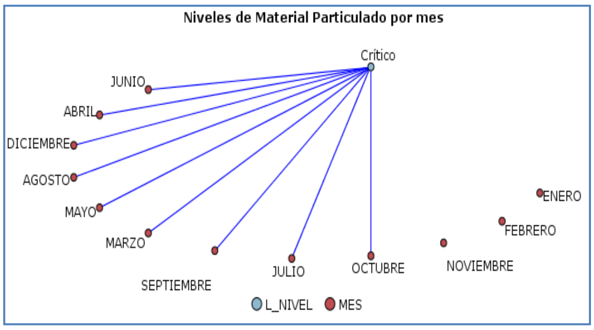


Figure 6: Months of the year that critics are recorded pollution levels.

According to the data recorded for the year 2000, in the only months when there were no critical levels of contamination were in the months of January, February and November.

b) Mesh dangerous levels.

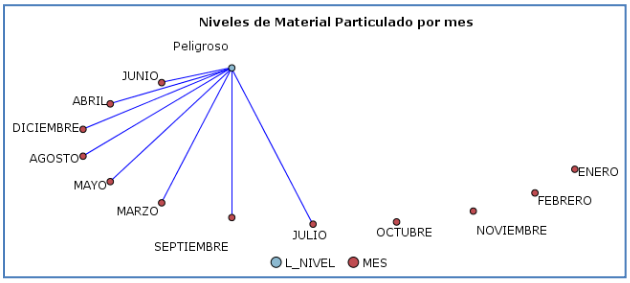
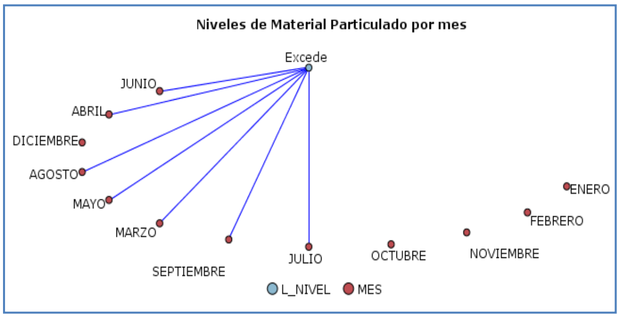


Figure 7: Months of 2000 that are recorded pollution levels considered "dangerous"

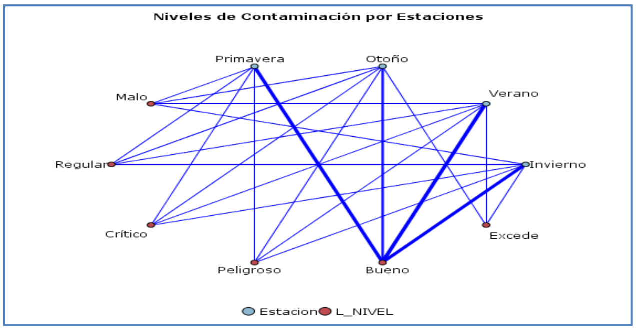
During 2000, are recorded dangerous levels of contamination between the months of March to December except for the months of October and November.



c) Exceeds Levels of Particulate Matter

Figure 8: Level Exceeds Particulate Matter Pollution per month.

The extreme values of contamination occurred in the period March to December with the exception of the months October and November.



d) Mesh stations contamination levels

Figure 9: Relation between levels seasons Pollution Monitoring Station in Florida

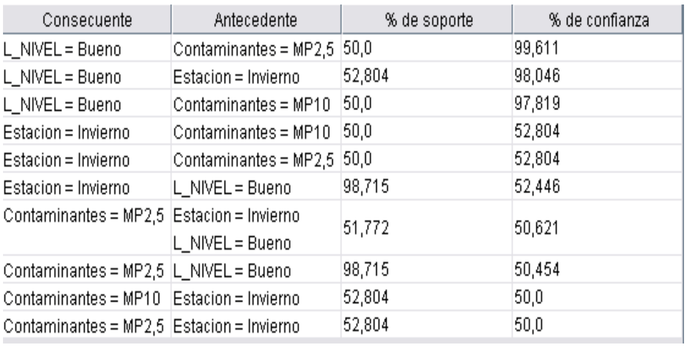
In the commune of Florida levels "good" of contamination found during the summer.

IV. Generation of Association Rules using the Clementine software.

In order to generate association rules use the node "Apriori" Clementine software, also defines a "medium" and "low confidence" of the "rule" of 50%. Recall that is the "support" that defines the frequency at which the items appear together and is "confident" that determines the cohesion of the data.

Minimum Support	= 50% rules
Minimum Confidence	= 50% rules
Maximum background	= 5

The Mobile Media concentrations of 24 hours is available for both Particulate Matter (PM10 and PM2.5) from the year 2001 which will be used as a basis this year to continue until 2012. For testing purposes will apply Air Quality Standard of PM10 to PM2.5 particulate matter, so have a basis to regulate and standardize the latter contaminant approximate.



Consecuente	Antecedente	% de soporte	% de confianza
L_NIVEL = Bueno	Contaminantes = MP2,5	50,0	99,611
L_NIVEL = Bueno	Estacion = Invierno	52,804	98,046
L_NIVEL = Bueno	Contaminantes = MP10	50,0	97,819
Estacion = Invierno	Contaminantes = MP10	50,0	52,804
Estacion = Invierno	Contaminantes = MP2,5	50,0	52,804
Estacion = Invierno	L_NIVEL = Bueno	98,715	52,446

Contaminantes = MP2,5	Estacion = Invierno
	L_NIVEL = Bueno
		51,772	50,621

Contaminantes = MP2,5	L_NIVEL = Bueno	98,715	50,454
Contaminantes = MP10	Estacion = Invierno	52,804	50,0
Contaminantes = MP2,5	Estacion = Invierno	52,804	50,0

Table 1.2: Association rules extracted from data recorded in Florida.

Higher levels of trust achieved in the first three rows. The way to structure the data explains that the support of the first and third rule matching in proportion, and that both have the same consistent because both are governed by the Air Quality Standard for particulate matter defined MP10, if done comparisons between them makes the difference in the percentage of confidence that the Apriori algorithm assigns data.

From the above rules is the third rule that states "If the contaminant is then the level of PM10 Air Quality in Florida is good", which expresses a truth more tangible as it has its own Statement of Regulatory Quality Air. For this reason there is a certainty of more than 97% chance that this is true for the contaminant represented, but "unknown" if this will be as good for particulate matter PM2.5, since we assume that this last and MP10 are governed by the same rules. If the PM2.5 less polluting than the MP10 not generate greater problem because the Trust is very close to 100% and the air quality is good, but if the same level of support, own standard for Particulate Matter PM2.5 define poor air quality, support almost 100% would take steps to reduce pollution levels.

V. CONCLUSIONS

With regard to the results, a relationship is established between PM10 and PM2.5 particulate material, which stresses that when the contaminant reaches a higher index MP2.5 or close to 85 g/m3, the level of contamination of the material PM10 particulate is considered bad, which means that lower levels of this contaminant could be causing health problems in people. Now, if this assumption is compared with graphics relationship between these particulate materials may be accounted for when the rate shown MP2.5/MP10 similar to a straight line, the levels of contamination are classified as "dangerous", "bad" or "exceeds" and this is because the rates of particulate matter PM2.5 to PM10 approach, surpassing their threat and almost reaching the contaminant than 10 microns in diameter.

VI. REFERENCES

[1] Aristóteles, "La Metafísica".

[2] Mario Bunge. Causalidad. Editorial Universitaria, Buenos Aires, 1978.

[3] Hobbes, Thomas. 1996 (1651). Leviatán (México, Fondo de Cultura Económica).

[4] Jean Wahl. Introducción a la filosofía .Fondo Cultura Económica, México, 1954.

[5] Fernando Berzal, Ignacio Blanco, Daniel Sánchez. and María Amparo Vila, Measuring the accuracy and interest of association rules: A new framework, Department of Computer Science and Artificial Intelligence, University of Granada, E.T.S.I.I, March 2002.

[6] C. Silverstein, S. Brin, et al. [1998] "Scaleable Techniques For Mining Causal Structures," Proceedings. 1998 International Conference Very Large Data Bases, NY, 594-605

[7] G. Cooper [1997] "A Simple Constraint-Based Algorithm for Efficiently Mining Observational For Causal Relationships" in Data Mining and Knowledge Discouvery, v 1, n 2, 203-224

[8] J. Pearl, J. [2000] Causality: Models, Reasoning, And Inference, Cambridge University Press, NY.

[9] W. Frawley and G. Piatetsky-Shapiro and C. Matheus, Knowledge Discovery in DataBases: An Overview. AI Magazine, Fall 1992, pgs 213-228.

[10] RED MACAM (Red de Monitoreo Automático de Calidad del Aire y Meteorología), www.conama.cl/rm/568/article-1114.html

[11] Universidad de Santiago de Chile, "Normativa Ambiental en Aire. (Fuentes Fijas)",

[12] "Geofísica de la Atmósfera: Pronosticando la Contaminación Atmosférica mediante Redes Neuronales", Departamento de Geofísica Universidad de Chile.

[13] Secretaría Regional Ministerial de Salud Región Metropolitana(SESMA)."Aire:InformaciónGeneral", http://www.asrm.cl/sitio/pag/aire/indexjs3aire.asp

[14] C. Arguedas. "INFORME. Análisis de las Normas de Calidad del Aire en Chile, Estados Unidos, México y la Comunidad Europea", SESMA, Chile, 2002.

[15] Zapata Caceres Santiago, Escobar Ramirez Luis, Reyes Pastore Carlos, Cortez Torres Jhons, Fuzzy Approach to the Management of the Environmental Contamination in Santiago city of Chile, 2007 International Conference on Artificial Intelligence (ICAI 2007), vol.II, 2007, Las Vegas, Nevada, USA

[16] Zapata Caceres Santiago, Escobar Ramirez Luis, Intelligent Analysis to the Contamination in the City of Santiago from Chile.Advances in systems, computing sciences and software engineering, Proceedings of SCSS 2005, Sobh, Tarek; Elleithy, Khaled (Eds.), 2006, XIV, 437 p., T. Sobh, University of Bridgeport, Bridgeport, USA; K. Elleithy, University of Bridgeport, Bridgeport, USA (Eds.)

[17] C. Kuok, A. Fu, M. Wong. "Fuzzy Association Rules in Databases", Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. 1998.

[18] "Geofísica de la Atmósfera: Pronosticando la Contaminación Atmosférica mediante Redes Neuronales", Departamento de Geofísica Universidad de Chile, Dirección de Internet: http://www.geofisica.cl/English/pics3/FUM6.htm

[19] L. X. Wang, J. Mendel. "Generating Fuzzy Rules by Learning from Examples", IEEE Transactions on Systems, Man, and Cibernetics 22, 1414 -1427, (1992).

[20] Zapata C. Santiago, Maruri B. Christian, Rojas B. Ronald, Creation of a Data Warehouse using the F-Cube Factory Software to Resolve Problems with Degrees of Truth, Proceedings of the 2nd European Conference of Computer Science (ECCS '11), WSEAS, Puerto De La Cruz, Tenerife, Spain, December 10-12, 2011

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0
https://creativecommons.org/licenses/by/4.0/deed.en_US