Fault predictive model for HVAC Systems in the context of Industry 4.0

Inés Sittón-Candanedo
BISITE Digital Innovation HUB
University of Salamanca
Salamanca, Spain
isittonc@usal.es

Sara Rodríguez-González
BISITE, Digital Innovation HUB
University of Salamanca
Salamanca, Spain

Juan Manuel Corchado-Rodríguez[1,2,3]
[1]Bisite Digital Innovation HUB
University of Salamanca, Spain
[2]Osaka Institute of Technology,
Osaka, Japan
[3]University Malaysia Kelantan,
Kelantan, Malaysia

Elena Hernández-Nieves
BISITE Digital Innovation HUB
University of Salamanca
Salamanca, Spain

Fernando de la Prieta-Pintado
BISITE Digital Innovation HUB
University of Salamanca
Salamanca, Spain

Abstract—**Fault prediction has gained importance with Industry 4.0 paradigm. Nations, universities and several companies be research in this topic with the aim to optimize manufacturing process and reduce cost of equipment maintenance through create predictive models. The Heating, Ventilation and Air Conditioning Systems (HVAC) control in an important number of industries: in-door climate, air's temperature, humidity and pressure, creating an optimal production environment. In this paper, the use of some machine learnings algorithms applied to HVAC System data set are review as an instrument for evaluate a fault predictive model.**

Keywords—Fault prediction; HVAC Systems; Industry 4.0; Machine Learning, Logistic Regresion, Decision Trees.

I. Introduction

The development of automated models for the detection and diagnosis of faults in Heating, Ventilation and Air Conditioning Systems (HVAC) has been the subject of continuous research for many years. These researches have been developed for different reasons such as [1]:

- The continuous increase in energy costs;
- The interest of organizations in reducing their operations budgets;
- Reducing maintenance costs and equipment downtime;
- Avoiding production stoppages in factories that must maintain a standard temperature due to quality controls.

Heating, Ventilation and Air Conditioning Systems (HVAC) are an important component not only in residential buildings, also in industrials where it have an important role because their faults can produce an industry downtime. [2] Therefore, an opportune fault conditions recognition such as: (i) malfunctions in air distribution; (ii) very high, low or extreme temperature; (iii) inappropriate air distribution for several periods of time and others can help to prevent the increase of maintenance, energy or production costs.

In 2000, Norford, Wright, Buswell, and Luo developed important researches about the importance of implementing fault detection and diagnosis system in industrial HVAC equipment's to increase equipment performance and durability. Norford, L., *et al.*, indicate that the use of automated algorithms can reduce the delay in fault detection, prevent downtime and correct equipment inefficiencies [3].

Since 2011 Industry 4.0 is using to define the new concept of factories in which manufacturing process is supervised by sensors and autonomous systems. These news ideas are affected important disciplines such as: (i) system, mechanical, industrial and electrical engineering; (ii) computer science; (iii) business information and administration [4]

The strengthening of these disciplines and emerging technologies have renewed the interest of researchers in universities, companies and government institutes. Currently, there is a great interest in the development of projects on automated systems that incorporate technologies for predictive analysis, oriented to the industrial sector.

One of the indicators of Industry 4.0 is the sensorization of the machines or equipment that are part of the infrastructure of the factory and the production process. In this sense, performance optimization is based on the ability to connect them to a data network. [5], [6].

In this technological context, the purpose of this papers is to use techniques and tools for the development of models capable of predicting future events, failures or behaviors. The prediction models proposed are built using statistical techniques, machine learning (ML) or data mining with which you can extract existing patterns in the data set that is analyzed. Allowing to obtain useful information for decision making processes, improving among other things; predictions and the anticipation of errors or failures in equipment, inherent risks in the production, production and expected sales volume [7]. However, these actions are linked to the processing of an overwhelming amount of data, which constitutes a challenge [8].

This document is divided as follows: the second section provides a summary of Industry 4.0, predictive maintenance, the application of machine learning (ML) in the context of Industry 4.0 and the prediction of failures. Then the case study is described and finally the results that show the accuracy of the algorithms and ML techniques applied to the conclusion and the lines of future work are presented.

II. Background

A. *Industry 4.0*

The genesis of the Industry 4.0 concept is linked to the joint efforts that took place in Germany during the year 2011, where the government and the business sector led by the Bosch Company, formed a research group with the purpose of establishing a common framework that allowed the application of new technologies. This would imply significant improvements to the productive sectors of the country. The first group report was delivered in 2012 and was publicly presented in 2013 at the Hannover Fair [9].

With this initiative of the German government emerge what is known as the fourth industrial revolution, which other countries have decided to promote with names such as: Smart Manufacturing, Smart Production, Industrial Internet, i4.0, Connected Industry 4.0, to identify everything that encompasses the paradigm. The adoption of industry 4.0 is related to a set of technological enablers such as: IoT, CPS, Big Data, Wireless Sensor Networks (WSN) and Cyber physical systems (CPS) [2]. The IoT being one of the ones that provides the greatest support and influence.

B. *Internet of things*

About the IoT, there is no standard definition universally accepted. However, the ITU (International Telecommunication Union) and the IERC (Internet of Things European Research Cluster) define it as: "*a global and dynamic network infrastructure with the self-configuration of the capabilities based on protocols of standard and interoperable communication, where physical and virtual "things" have identity, physical attributes, virtual personalities and use intelligent interfaces that integrate seamlessly into the information network*" [9][10].

The IoT or Internet of Things has become a fundamental element for the business sector within the context industry 4.0. This technology is the basis for generating great growth and increasing the indices of productivity and competitiveness [11].

The technological infrastructure that makes up the physical facilities of factories in the industry 4.0 model use control systems that share information among them, and these in turn apply the necessary changes, based on the data generated in the form of alerts or warnings. Subsequently commands or orders are sent to slave teams to modify the production process, thus preventing the production lines from stopping, which would generate losses in time and money. On the other hand, human operators can use the data obtained to properly manage the conditions of the equipment and significantly increase the efficiency and effectiveness of the processes.

Another aspect that brings IoT technology to the technological current of Industry 4.0 lies in the possibilities of communication in real time between the business (factory), suppliers and customers. This will achieve a better business-supplier relationship, which would ultimately guarantee meeting the demands of a market composed of highly critical and demanding customers [12].

C. *Machine Learning for prediction*

The Automatic Learning (ML) is a technology that already has several years of being used in the computer field. However, recently it is when it has gained great importance thanks to the advances in other technological areas such as the IoT. Initially it was used in the research for the search and identification of patterns, to get computers to learn [14].

According to Mitchell (2006) of Carnegie Mellon University, Machine Learning answers the question "*How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes*?" [13]

Therefore, universities, research centers and industries have identify the enormous potential of this technology, transcending its use to a wide variety of areas such as neuroscience, health or economics. This has directly impacted the development of applications that allow us to solve not only the processing of large volumes of data from the Internet of Things and Big Data, but also to provide new information by extracting patterns and building predictive models, with automatic learning algorithms (ML) [15].

There are several types of algorithms based on machine learning (ML) that are grouped in supervised or not. Exist some classification criteria which divide them into others categories such as: decision tree, grouping or regression. However, something that is common to all these algorithms is the existence of components such as: representation, evaluation and optimization, figure 1 [14].

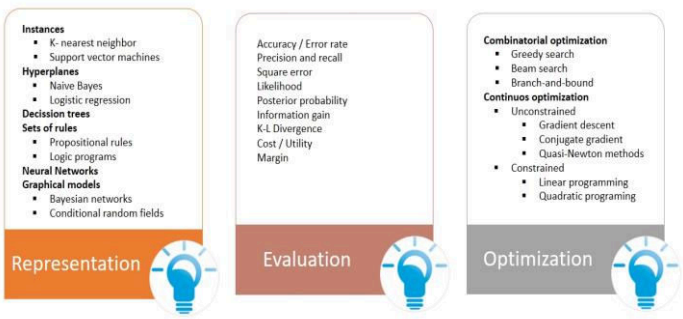

Figure 1. Components of learning algorithms.

In their jobs Murphy, 2012 and Michell, 1997 established that machine learning techniques emulate human cognition and learn with the aim to training and predict future events in more complex cases. In a brief review makes by [16] about more

common machine learning methods used in fault detection of HVAC Systems the most frequently are:

- Neural networks (NN) to predict, classify or control future problems (Bansal et al., 2015, Taylor 1996, Mehrotra 1997, Fausett 1994, Freeman 1993).
- Multilayer perceptron (Ruck et al, 1990)
- K-nearest neighbours (Denoeux, 1995)
- Naïve Bayes where the predictor is independent and makes an efficiently classification that is easy to interpret (Panda and Patra, 2007)
- Linear regression (Daniel, 2011) to predict the next failures.
- Artificial Neural Networks (ANN) for faults filters (Delgrange et al., 1998).

III. Industry 4.0 Environment Case Study

In this case study, a structured data set was used columnar. The data includes the optimal temperature and the real values captured by sensors located in buildings, to analyze how HVAC air conditioning systems behave. With this, it can be determined if the analyzed equipment is failing by not maintaining the lower temperatures in an optimal range of values.

Heating, ventilation and air conditioning (HVAC) systems control the internal climate, air temperature, humidity and pressure, creating an optimal production environment in industrial buildings. These equipment's are of great importance for the operation of a factory. However, maintenance routinely does not always clearly identify its faults. In this sense, the purpose of implementing a predictive maintenance for infrastructures of industry 4.0, would be to extend the useful life of the equipment using different tools and techniques to identify abnormal patterns such as: vibration, temperature or balance [17].

Following section describes a free dataset from temperature sensors installed on Heating, Ventilation and Air Conditioning System (HVAC) in 20 buildings [18].

3.1 Dataset Description

This dataset contains a total of 8000 (eight thousand) temperature records (TargetTemp) captured by a sensor network, installed in a set of buildings who were between 0 and 30 years old, their age corresponded to age of the HVAC systems. Table I shows the name of dataset attributes and their description.

TABLE I. DESCRIPTION OF DATASET ATTRIBUTE

Attribute	Description
Date	Date of measurement
Time	Measurement time
TargetTemp	Temperature measured by the sensor
Actualtemp	Optimal temperature for the system
System	System Model
SystemAge	Age of the HVAC System
BuildingID	Building Identifier

3.2 Pre-processed

For preprocessing, a range was established for normal temperatures and two types of alarms that indicate extreme temperatures and, therefore, a possible failure. These are described in figure 2 as follows:

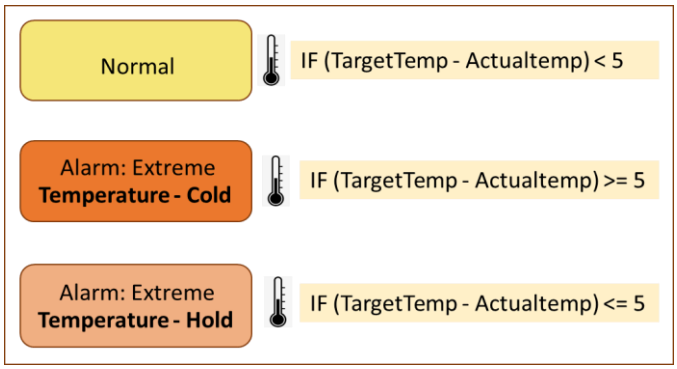

Figure 2. Conditions to determine normal and extreme temperature.

Two labels are added to the dataset 'Difference' and 'FilterDifference', in the first, the values obtained from the difference between 'TargetTemp' and 'Actualtemp' are stored. In 'FilterDifference' the binary conversion is carried out assigning 0 to the normal temperatures and 1 to the alarms for extreme temperature.

3.3 Training and results

Once the data were pre-processed, the extended dataset was used to divide the data into data train and data_test, the former was used to apply Machine Learning algorithms to obtain the prediction model. This model was then validated with the data_test. For the training of the data, two supervised learning algorithms will be used: Logistic Regression and Decision trees to evaluate the accuracy of each one in the prediction.

Logistic regression is a machine learning technique, statistical-inferential, which dates to the 1960s, used in current scientific research. It is considered an extension of linear regression models, the difference is that is has a categorical variable capable of being binomial (0, 1) or multiple [19][20]. For the development of this research, the dataset was pre-processed so that the categorical variable (y) can be binomial. Applying the logistic regression analysis, we assume that y = 1, when the sensor sends an extreme temperature and y=0 when the measured temperature ('TargetTemp') is within the normal range. Considering the above, the probability that the HVAC system is presenting a failure by recording extreme temperatures is given in equation 1:

$$P(y = 0) = 1 - P(y = 1) \quad (1)$$

$$Y = f(B_0 + B_1X_1 + B_2X_2 + \cdots + B_nX_n) + u \quad (2)$$

Where u is the error term and *f* the logistic function:

$$f(z) = \frac{e^z}{1+e^z} \quad (3)$$

So, that:

$$E[Y] = P = P(Y = 1) = \frac{e^{B_0+B_1X_1 + B_2X_2+\cdots+B_nX_n}}{1+e^{B_0+B_1X_1 + B_2X_2+\cdots+B_nX_n}} \quad (4)$$

$$ln\left(\frac{P}{1-P}\right) = B_0X_0 + B_1X_1 + B_2X_2 + \cdots + B_nX_n \quad (5)$$

Decision trees (DT) can be described as the combination of mathematical and computational techniques that allow us to understand the importance of the attributes of the description, categorization and generalization of a particular set of data. On the other hand, it is one of the most successful techniques for supervised learning through classification [21][22].

Through DT the data that will be analyzed can be expressed in the following way: (x, Y) = (x1, x2,…, x_n, Y). The dependent variable, Y, is the objective variable that must be understood, classified or generalized (FilterDifference, in our case). The vector x is composed of the input variables (or attributes) x1, x2, x3, etc., that are used for that task. The DT is formed through the following steps [21].

a. Define the regression deviation of a node, as expressed in equation 6.

$$D(I) = \sum_{j=1}^{I_k} (y_{I_j} - \bar{y}_I)^2 \quad (6)$$

Where:
y_{I1}, y_{I2}… $y_{Ik,}$ are the values of the target variable that compose the node I, and $\bar{y}_I$ is their average [20]

b. Prune the tree removing divisions from bottom to top, equation 7.

$$D_T(\alpha) = D_T + \alpha|T| \quad (7)$$

Where |T| is the number of terminal nodes and α is a penalty term which ensures the greatest compromise between predictive accuracy and tree size [21].

The evaluation of a prediction model can be done according to different aspects. Therefore, for this research, authors use the accuracy, ROC curve and Spearman's Rho to evaluate the best model. Spearman's Rho is a measure of the relation between two variables. In this case X = SytemAge and Y= FilterDiference as categorical variable. It is defined in equation (8). Spearman's Rho take values between -1 to 1 where higher values indicated a better model.

$$r_s = cov(rg_x, rg_y) / \sigma_{rgx}\, \sigma_{rgy} \quad (8)$$

In other hand, to obtain the accuracy a confusion matrix was made using equation 9

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (9)$$

Where:

- TP = True Positive
- FN = False Negative
- FP = False Positive
- TN = True Negative

In the case of Receiver Operation Curve (ROC) the area under the curve can be interpreted as the probability that before a pair of binary values, one represents the failure and the other does not, the test classifies them correctly [23]. The BigMl platform was used to generate Receiver Operation Curve (ROC) and the results show in the next section. This tool permit analyzes real data set to build machine learning predictive models.

IV. RESULTS AND DISCUSSION

In this research two supervised machine learning algorithms was tested for identify abnormal behavior. They were described in section 3.3., with the training process.

After the training phase, the logistic regression and decision tress algorithms enter in a test phase. The 8000 instances that make up the dataset were divided, taking 70% for training and 30% for test data. The SystemAge column was taken as a characteristic and, as a label, the FilterDifference column, the total of instances was 8000. The evaluation of the model generated an accuracy of 66.9%. To calculate it, equation 9 was used.

After computers, the Spearman's Rho between the ranked actual values of the instances and the model predictions, with logistic regression, the Spearman's Rho was 0.0132. It indicated a better model. Figure 4 is the graph for this model. The ROC AUCH was 0.5557.

The logistic regression model used in this article compares between the categorical variable and a set of independent or predictive variables. In this phase to identify the attributes that most contribute to predict a failure three variables were exposed to linear regression analysis with the BigMl platform.

In this paper we also show how the use of a decision tree can greatly reduce the occurrence of false positives or negatives.

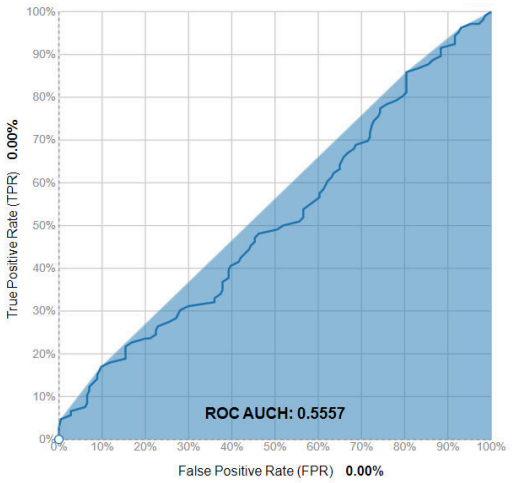

Figure 4. Area under ROC curve with the logistic regression model.

After applying the logistic regression model and obtaining its percentage of precision, the decision tree model was applied to the data set. The results are shown graphically (confusion matrix and the ROC graph) in figure 5. The evaluation of the model provided a precision percentage of 64.2%.

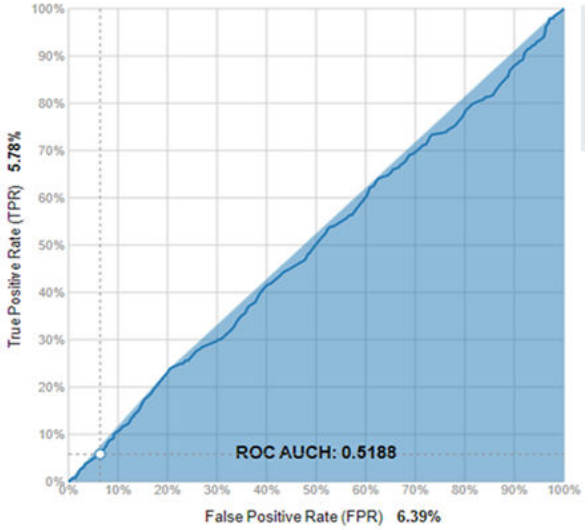

Figure 5. Area under ROC curve with the decision tree model.

V. Conclusions and Future Work

This work presents two machine learnings algorithms used to predict faults in HVAC System.

The proposed prediction model still in its early stage of development. This allows for the implementation of other machine learning techniques and for the use of larger datasets obtained from sensors networks installed in order environment. The results for the dataset used in this case study, show that the precision of the logistic regression model is similar to that of decision tree model, in predicting malfunction in the HVAC system.

The modeling and integration of the large volumes of industrial data that are generated by machines and collected by sensors, is a clear problem that still needs to be addressed in future researches. Thus, testing with other machine learning methods for classification, training and prediction. These test will provide the grounds for the development of algorithms that generate predictive models adapted for organizations, in the context of industry 4.0.

As future work it is possible to conduct additional studies using other techniques to compare and identify the best performing predictive approach for HVAC systems.

Acknowledgment

This work was supported by the Spanish Ministry of Economy and FEDER funds. Project. SURF: Intelligent System for integrated and sustainable management of urban fleets TIN2015-65515-C4-3-R.

I. Sittón-Candanedo has been supported by IFARHU – SENACYT scholarship program (Government of Panama).

References

[1] House, John M.; LEE, Won Yong; SHIN, Dong Ryul. Classification techniques for fault detection and diagnosis of an air-handling unit. ASHRAE Transactions, vol. 105, p. 1087, 1999.

[2] Perez-Lombard, L., Ortiz, J., Pout, C., “A review on buildings energy consumption information, Energy Build, 40(3), 394-398, 2008.

[3] Norford, L. K., et al. Final report of ASHRAE Research Project 1020-RP: Demonstration of fault detection and diagnosis in real a building. Massachusetts Institute of Technology and Loughborough University, 2000.

[4] Kagerman, H., Anderl, R., Gausemeier J., Schuh G., and Wahlster W. “Industrie 4.0 in a Global Context: Strategies for Cooperating with International Partners”, Acatech Study, Munich, Germany. https:// www.acatech, 2016.

[5] Rivas, A., Martín, L., Sittón, I., Chamoso, P., Martín-Limorti, J. J., Prieto, J., & González-Briones, A. Semantic Analysis System for Industry 4.0. In International Conference on Knowledge Management in Organizations (pp. 537-548). Springer, Cham, 2018.

[6] Kuo, C. J., Ting, K. C., Chen, Y. C., Yang, D. L., & Chen, H. M., “Automatic machine status prediction in the era of Industry 4.0: Case study of machines in a spring factory”. In: Journal of Systems Architecture, Vol., 81, 44-53, 2017.

[7] Bishop, C.M., “Pattern recognition and machine learning”. Springer, New York, Vol. 4, doi:10.117/1.2819119, 2006.

[8] Civerchia, F., Bocchino, S., Salvadori, C., Rossi, E., Maggiani, L., and Petracca, M., “In-dustrial Internet of Things Monitoring Solution for Advanced Predictive Maintenance Ap-plications”. In: Journal of Industrial Information Integration, 2017.

[9] Hermann, Mario; Pentek, Tobias; Otto, Boris. Design principles for industrie 4.0 scenarios. In System Sciences (HICSS), 49th Hawaii International Conference on. IEEE, 2016. p. 3928-3937, 2016.

[10] Li, Zhe; Wang, Yi; Wang, Ke-Sheng. Intelligent predictive maintenance for fault diagnosis and prognosis in machine centers: Industry 4.0 scenario. Advances in Manufacturing, 2017, vol. 5, no 4, p. 377-387.

[11] Brettel, M., Friederichsen, N., Keller, M., & Rosenberg, M. (2014). How virtualization, decentralization and network building change the

manufacturing landscape: An Industry 4.0 Perspective. International Journal of Mechanical, Industrial Science and Engineering, 8(1), 37-44.

[12] Xu, L. D., Xu, E. L., & Li, L. (2018). Industry 4.0: state of the art and future trends. International Journal of Production Research, 56(8), 2941-2962.

[13] Mitchell, T. M. (2006). The discipline of machine learning (Vol. 9). Carnegie Mellon University, School of Computer Science, Machine Learning Department.

[14] Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78-87.

[15] Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. "Efficient machine learning for big data: A review". In: Big Data Research, Vol., 2(3), 87-93, 2015.

[16] Tehrani, Mahdi Mohammadi, et al. A predictive preference model for maintenance of a heating ventilating and air conditioning system. IFAC-PapersOnLine, vol. 48, no 3, p. 130-135, 2015.

[17] Verbert, K., Babuška, R., De Schutter, B. "Combining knowledge and historical data for system-level fault diagnosis of HVAC systems". In: Engineering Applications of Artificial Intelligence, Vol., 59, 260-273, 2017.

[18] Hortonworks, "Analyze HVAC Machine and sensor data". https://es.hortonworks.com/ha-doop-tutorial/how-to-analyze-machine-and-sensor-data/#section-2. 2017.

[19] De Menezes, F. S., Liska, G. R., Cirillo, M. A., & Vivanco, M. J., "Data classification with binary response through the Boosting algorithm and logistic regression". In: Expert Sys-tems with Applications, Vol., 69, 62-73, 2017.

[20] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X., "*Applied logistic regression".* Edi-torial: John Wiley & Sons, Second Edition, p. 500, 2013. ISBN 978-0-470-58247-3.

[21] Lin, L. H., Chen, K. K., & Chiu, R. H. (2017). Predicting customer retention likelihood in the container shipping industry through the decision tree approach. Journal of Marine Science and Technology, 25(1), 23-33.

[22] Liu, X., Li, Q., Li, T., & Chen, D. (2018). Differentially private classification with decision tree ensemble. Applied Soft Computing, 62, 807-816.

[23] Bradley, Andrew P. "The use of the area under the ROC curve in the evaluation of machine learning algorithms." Pattern recognition 30.7 (1997): 1145-1159.