







Health Indicator Modeling and Association Rule Mining for Stoppages Prediction in a Refinery Plant

Giovanni Mazzuto , Sara Antomarioni ,
Filippo Emanuele Ciarapica , and Maurizio Bevilacqua 

Department of Industrial Engineering and Mathematical Science,
Università Politecnica Delle Marche, Ancona, Italy
{ g.mazzuto, s.antomarioni, f.ciarapica,
m.bevilacqua}@univpm.it

Abstract. Predictive maintenance practices represent a relevant feature for improving the useful life of the systems and decreasing costs. Hence, the modelling of the Health Indicator is a useful support in order to define the Remaining Useful Life of a system. In this work, an approach for the Health Indicator modeling of an oil refinery sub-plant is proposed; in addition, the Association Rule Mining is applied, in order to identify the components frequently requiring a work order prior to a stoppage of the plant: in this way, the Remaining Useful Life determined via the Health Indicator is used to inspect such components and, possibly, avoid the stoppage.

Keywords: Predictive maintenance · Health indicator · Association rule

1 Introduction and Background

The emerging Industry 4.0 technologies nowadays provide reliable and accessible smart systems that enable the spreading of predictive maintenance (PdM) practices [1]. The importance of PdM is testified by its ability in improving the useful life of the machine and in decreasing the maintenance costs [2]. The approach at the basis of the development of a reliable PdM system regards the collection and analysis of large amount of data, belonging to relevant time frames [3] as well as the definition of an indicator of system's health [4]. Assessing an appropriate Health Indicator (HI) allows the understanding of the deviation from the regular operating performance of the system in terms of its Remaining Useful Life (RUL). HI definition supports in increasing the knowledge of the system by focusing the analysis on the most relevant information sources. In this sense, having a precise HI enables the possibility of predicting the RUL of a system confidently [5] and is thus the main focus of many researches (e.g., [6, 7]). Recent literary contributions focused on the development of a HI using several different techniques. For instance, some focused on the development of multi-objective models to derive the health of the system for fault diagnostics and prognostics [8], while other implemented artificial neural networks and K-means clustering [9] and genetic algorithms [10]. Even in terms of application areas, there is a

certain heterogeneity: in some works, the focus is posed on semiconductor equipment [11], other focus on the vibration analysis of wind turbines [12]. HIs can indeed be applied for the definition of the remaining useful life of mechanical machinery, as testified by several works (e.g., [13, 14]). Given these assumptions, this work proposes an approach to model the health indicator for a sub-plant of an oil refinery and identify the component causing the performance loss. Predictive maintenance interventions on specific components are performed to avoid system stoppage, prioritizing them through implementing the Association Rule Mining (ARM). Indeed, the Association Rules (ARs) among component failures before the occurrence of a plant stoppage are used as a guide to determine, with a certain probability level, which are the components that caused the HI worsening. In this way, the ARs help identify relationships within a dataset when they are not immediately identifiable [15]. In recent literature, ARM is applied to different fields, ranging from the behavior description [16] to the sub-assemblies production scheduling [17]. In a predictive maintenance perspective, ARM has already been applied to detect the relationships among component failures [18]. Despite the valuable implementations proposed, among the others, by the named authors, there is a lack of research in terms of joint application of HIs definition and ARM.

2 Methodology

In the following, the proposed procedure to define the Health Indicator is described. The procedure is general so that it can be applied to various equipment as long as sensor readings are available. The input dataset contains the readings of one or more system sensors. For simplicity of explanation a single sensor is considered. Seven fundamental steps are performed in order to model the HI:

1. Standardization of the signals (S_s) from system sensors and partitioning of the initial dataset into training and testing sets.
2. Modelling of the Health Indicator (HI): the mean time between two stoppages represents the life duration of the system. The objective at the basis of the HI is creating a degradation profile considering that at the beginning of the HI the reliability of the system is equal to 1 (hence, maximum) while, at the moment of the stoppage, it is minimum (hence, equal to 0). The behavior of the HI is described by Eqs. (1)–(3), being $DUR_{i,m}$ the time between two stoppages of category m considering the i -th machine, $TI_{i,m}'$ indicates the remaining hours before the stoppage, while $TI_{i,m}''$ is the normalized value.

$$TI'_m = [DUR_m - 1 \quad DUR_m - 2 \quad \dots \quad 0] \tag{1}$$

$$TI_m(t)'' = \frac{TI_m(t)'}{DUR_m} \tag{2}$$

$$HI_m(t) = TI_m(t)'' + (1 - TI_m(t = 1)'') \tag{3}$$

Equation 3 is such that the first value of $HI_{i,m}$ is equal to 1. In particular, $HI_{i,m}$ represents an initial HI for the considered machine and stoppage.

Once HIs have been calculated for each machine and stoppage, through a linear interpolation, HIs and S_s can be correlated in order to find the transformation coefficient (Eq. 4) able to translate the information from the measures space to the HI space.

$$HI_{total} = b \cdot S_{s,total} \tag{4}$$

HI_{total} represents the array composed of all the determined HIs ($HI_{total} = [HI_{1,m} HI_{2,m} \dots HI_{n,m}]$) and, at the same way, $S_{s,total}$ is the array composed of all the standardised signals ($S_{s,total} = [S_{s,1} S_{s,2} \dots S_{s,n}]$). The parameter b can be in the form of an array if more than one sensor readings are available. In the present paper, it is a scalar since just one sensor readings are available.

Once the transformation “ b ” has been identified, the transformed HI* is calculated for each machine and stoppage according to Eq. 5:

$$HI_{i,m}^* = b \cdot S_{s,i} \tag{5}$$

3. Once all the transformed $HI_{i,m}^*$ have been calculated, a non-linear interpolation has been performed to correlate the $HI_{i,m}^*$ with time (in form of Eq. 6).

$$HI_{i,m}^* = f_{i,m}(t_{s,i}) \tag{6}$$

In particular, function $f()$ can be chosen in the same form for all the stoppage categories or differently to define different profile for different categories. At this point, function $f()$ is stored to be used in the K-NN algorithm. Thus, the system training is completed.

4. During the testing phase, the testing signal is transformed as well, using the weights b determined at step 3. Even the duration of the standardized testing signals ($S_{s(test)}$) is assessed ($t_{s(test)}^*$) and the functions $f(t_{s(test)}^*)$ are evaluated for all the $S_{s(test)}$.
5. $HI_{j,m,test}^* = f(t_{s(j,test)}^*)$ and $HI_{i,m}^* = f(t_s^*)$ are compared through the KNN algorithm in order to identify the closest similarity profile. K-nearest neighbours (KNN) algorithm is a machine learning technique applicable for solving regression and classification problems. The main idea at its basis is that the closer an instance is to a data point, the more similar they will be considered by the KNN [19]. Reference functions $-HI_{i,m}^* = f_{i,m}(t_{s,i})$ - are considered as models to be compared to newly defined ones - $HI_{test}^* = f(t_{s(test)}^*)$. The distances d_{ij} (e.g., Euclidean distance) among $HI_{j,m,test}^*$ and $HI_{i,m}^*$ are used to calculate the similarity weights between the testing and the training. The similarity weight sw_{ij} is determined as reported in Eq. 7. Then, the weights are ranked in descending order and, finally, determine the number of similar units (Eq. 8). Specifically, k refers to the number of function to be selected for the comparison, while N is the number of training units.

$$sw_{ij} = \exp(-d_{ij}^2) \quad (7)$$

$$SU = \min(k, N) \quad (8)$$

6. Starting from the KNN results, the Weibull distribution is fitted considering the k-similar profiles and the median is determined.
7. Subtracting the $t_{s(test)}^*$ from the median determined at step 6, the RUL is assessed.

Eventually, the proposed approach requires the extraction of the ARs describing the relationships between component failures and plant stoppages. In this way, when a deviation in the operating performance is detected, the estimated RUL is used to inspect the components likely to be the ones causing the stoppage, so that their normal functioning can be reset and, possibly, the actual stoppage avoided.

Mining the Association Rules from a dataset implies the extraction of non-trivial attribute-values associations which are not immediately detectable due to the dimensions of such dataset [20]. Consider a set of Boolean data $D = \{d_1, d_2, \dots, d_m\}$ named items and a set of transactions $T = \{t_1, t_2, \dots, t_k\}$; a transaction t_i is a subset of items. An Association Rule $a \rightarrow b$ is an implication among two itemsets (a and b) taken from D, whose intersection is null ($a \cap b = \emptyset$). In order to evaluate the goodness of a rule, different metrics can be used, such as the support (Supp) and the confidence (Conf):

- $Supp(a, b) = \frac{\#(a,b)}{\#(T)}$: it measures the number of transaction containing both a and b over the totality of transactions.
- $Conf(a \rightarrow b) = \frac{supp(a,b)}{supp(a)}$: it measures the conditional probability of the occurrence of b, given the fact that a occurred.

For the purposes of this work, an ARs $a \rightarrow b$ is the implication relating the component requiring a work order (a) and the stoppage (b). So, the $Supp(a, b)$ expresses the joint probability of having a failure on a component and a stoppage, while the $Conf(a \rightarrow b)$ represents the probability of having a stoppage given the fact that the component a failed. In this work the FP-growth [21] is applied to perform the ARM.

When the health indicator highlights the risk of a stoppage, the components are inspected to control their functioning, sorting them by decreasing confidence value. If a failure or a malfunctioning is detected on the first component, it is replaced, else the following one is inspected; depending on the maintenance policy adopted by the company, the inspection can involve all the components included in the ARs, can stop when the first failure is detected or can involve the ARs until a certain threshold.

3 Application

The refinery considered for the case study is located in Italy. It has a processing capacity of 85,000 barrel/day. The sub-plant taken into consideration in this application is the Topping unit. Data refer to a three-year time interval. Specifically, the mass-flow

across the plant is collected hourly for each day. Three categories of stoppages or flow deceleration are identified by the company: Non-Significant (NS), if the reduction of the daily mass flow is between 20% and 50%; Slowdown (SLD), if the reduction of the daily mass flow is between 50% and 90%; Shutdown (SHD), if the reduction of the daily mass flow is between 90% and 100%. The dataset containing these data is structured as reported in Table 1: the first column indicates the date of the acquisition; the following twenty-four columns report the hourly mean value of the mass flow registered across the plant, while the last column indicates the kind of stoppage occurred during the day (if any). The mass-flow measures are also used to train and test the proposed approach. In all, 1095 rows and 24 columns are standardized and used to this end. The algorithm is carried out on an approach evaluation Intel® Core™ i7-6700HQ CPU @ 2.60 GHz, using Matlab 2019©. Once the dataset is standardized, steps 1–5 of the proposed approach are carried out. Figure 1 displays the HI profiles obtained through the algorithm in pink for the three stoppage category, while in black the current trend. Evidently, the latter cannot be considered as an anticipator of the NS stoppage, but is far more similar to the SHD one, given its trend. Hence, the through steps 6 and 7, of the proposed algorithm, the RUL can be determined and the relationships among the component failures and SHD stoppages can be enquired.

Table 1. Excerpt of the mass flow dataset indicating the sub-plant, the date of the acquisition, the hourly measurements and the stoppage category.

Date	v01	v02	v03	...	v22	v23	v24	Stoppage
01/01	411.5	409.6	407.56	...	407.22	407.37	407.56	–
02/01	409.49	410.8	408.03	...	378.41	374.27	372.32	NS
03/01	375.83	376.23	373.42	...	409.72	408.86	409.16	–

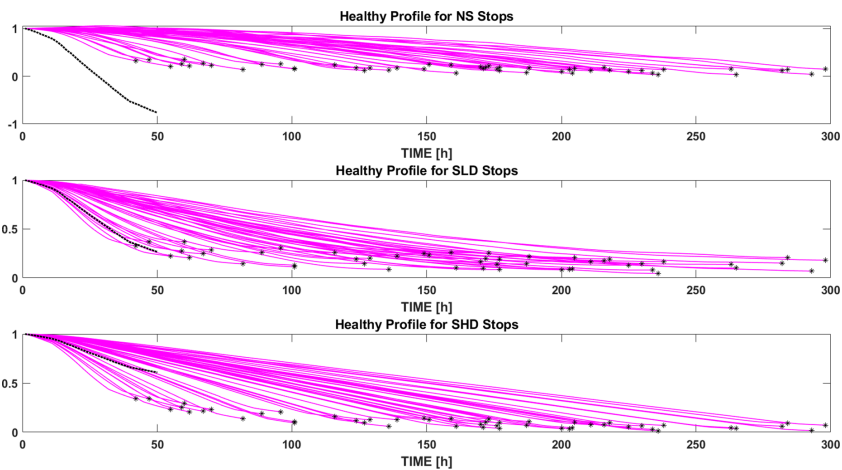


Fig. 1. The HIs comparison.

A second dataset, i.e., the work order list, is taken into account in order to identify the components requiring a maintenance intervention on a specific day (Table 2). This dataset is integrated with the one reported in Table 1 in order to be able to mine the association rules.

Table 2. Work Order (WO) dataset detailed by sub-plant, date and component.

Sub-plant	WO date	Component
Topping	10/06	Valve
Topping	17/11	Controller
Topping	04/02	Drainer
Topping	24/10	Valve

The relationships among component failures and stoppage category are derived through the Association Rule Mining. Specifically, the interest is identifying the work orders historically preceding the stoppages in order to use them as guidelines to understand which failure might cause the stoppage and intervene. The ARM, in this application, is executed using the well-known data analytics platform RapidMiner, that is widely applicable due to its graphical interface. In all, 120 rules have been identified, setting the minimum support threshold to 0 – in order not to lose any potential relation between components and stoppage category (and vice versa). The minimum confidence, instead, is set to 0.1. The set of ARs in the form component_i → stoppage_j is used to identify the components possibly affecting the abnormal functioning of the plant. In the proposed example, it appears that the deviation of the mass flow can be related to a SHD stoppage: 14 rules have been identified, even though only an excerpt is reported (Table 3).

Table 3. ARs relating the WO and the SHD stoppage

Component	Stoppage category	Confidence
Furnace	SHUT_DOWN	0.60
Condensation detector	SHUT_DOWN	0.25
Chiller	SHUT_DOWN	0.25
Chimney	SHUT_DOWN	0.23

This implies that the first component to be checked is the Furnace since, from the actual data of the past events, it requires a work order before the occurrence of a SHD (in other words, the rule Furnace → SHUT_DOWN has a confidence of 0.60). The following components to be checked, i.e., Condensation detector and Chiller, are the ones having the second value of confidence. During the inspection, the technician may detect a failure or a malfunctioning. If a failure or a malfunctioning is detected in one or

more components, they should be fixed in the attempt to avoid the occurrence of the stoppage. Remarkably, the order of the inspection is relevant in terms of time: indeed, according to the profile of the HI, the RUL is determined and the inspection, as well as the preventive replacing of the components should be carried out within the time limit imposed by the RUL.

4 Discussion and Conclusions

The proposed approach well suits the dynamic environment characterizing the maintenance field. Indeed, it supports the definition of the RUL of a system and, accordingly, defines the roadmap to inspect the possible cause of the performance losses. In this way, it is possible to fix the malfunctioning promptly, so that the stoppage of the system can be avoided or, at least, the flow can be restored shortly. One of the main advantages of the proposed approach, is the fact that part of the analysis is carried out offline (e.g., training and testing of the proposed datasets, association rule mining) while its application can be run online, during the functioning of the system. The datasets on which the analysis is based can be updated without any impact on the approach implementation. In the proposed example, a single sensor is considered. However, the approach is easily extendable to case studies receiving data from more sensors since the proposed algorithm is general. The accuracy of the proposed approach strictly depends on the quality of the collected data. Before starting the implementation of the algorithm, some preliminary activities on data are required: indeed, it is necessary to clean data from inconsistencies, replace the missing values, eliminate disturbances in the sensor system and eventually, filter to limit the boundaries of the study. In this way it is ensured that the starting database is solid, so that the results are reliable too. As shown in Table 4, the prediction error varies with the percentage of the validation data considered: selecting the 70% of the dataset allows the minimum prediction error, if compared to the 50% and 90% cases. These outcomes, however, are not generalizable since they are strictly related to the specific case study, the sampling time and the initial prediction instant.

Table 4. Prediction error ranges and percentiles varying the validation data percentage for Shut-down stoppages

Validation data	Upper	75%	Median	25%	Lower
50%	7.2	5.16	4.22	3.31	1.69
70%	-1.91	-2.03	-2.24	-2.25	-2.43
90%	-2.3	-2.49	-2.79	-3	-3.49

It should be considered that the data used in this study are usually collected in contexts such as process industries and refineries. In fact, they are used to establish the normal operation of plants and some basic maintenance activities. They are not the result of complex reprocessing or acquisition performed specifically for this purpose.

Few reworkings have been performed (e.g., standardization and filtering). The algorithm is therefore based on the data available to the organization. The algorithm provides promising results in terms of prediction and in reasonable times, being it also able to use data stored for other purposes and thus requiring a minimum effort in terms of data collection. From the company's perspective, knowing in advance the type of intervention to be made to avoid stoppages or to intervene promptly represents a considerable benefit. Indeed, the costs related to such stops are saved. In this work, the goodness of the algorithm is verified at theoretical level through simulations. Further development of this work regard a real evaluation of the actual advantages.

References

1. Ran, Y., Zhou, X., Lin, P., Wen, Y., Deng, R.: A Survey of Predictive Maintenance: Systems, Purposes and Approaches. <http://arxiv.org/abs/1912.07383>. Accessed 17 Mar 2021
2. Selcuk, S.: Predictive maintenance, its implementation and latest trends. *Proc. Inst. Mech. Eng. Part B J. Eng. Manuf.* **231**, 1670–1679 (2017). <https://doi.org/10.1177/0954405415601640>
3. Bevilacqua, M., Ciarapica, F.E., Mazzuto, G.: Fuzzy cognitive maps for adverse drug event risk management. *Saf. Sci.* **102**, 194–210 (2018). <https://doi.org/10.1016/j.ssci.2017.10.022>
4. Sotiris, V.A., Tse, P.W., Pecht, M.G.: Anomaly detection through a bayesian support vector machine. *IEEE Trans. Reliab.* **59**, 277–286 (2010). <https://doi.org/10.1109/TR.2010.2048740>
5. Yang, H., Sun, Z., Jiang, G., Zhao, F., Mei, X.: Remaining useful life prediction for machinery by establishing scaled-corrected health indicators. *Meas. J. Int. Meas. Confed.* **163**, 108035 (2020)<https://doi.org/10.1016/j.measurement.2020.108035>
6. Guo, L., Lei, Y., Li, N., Yan, T., Li, N.: Machinery health indicator construction based on convolutional neural networks considering trend burr. *Neurocomputing* **292**, 142–150 (2018). <https://doi.org/10.1016/j.neucom.2018.02.083>
7. Lei, Y., Li, N., Lin, J.: A new method based on stochastic process models for machine remaining useful life prediction. *IEEE Trans. Instrum. Meas.* **65**, 2671–2684 (2016). <https://doi.org/10.1109/TIM.2016.2601004>
8. Baraldi, P., Bonfanti, G., Zio, E.: Differential evolution-based multi-objective optimization for the definition of a health indicator for fault diagnostics and prognostics. *Mech. Syst. Sign. Process.* **102**, 382–400 (2018). <https://doi.org/10.1016/j.ymsp.2017.09.013>
9. Amihai, I., et al.: Modeling machine health using gated recurrent units with entity embeddings and K-means clustering. In: *Proceedings - IEEE 16th International Conference on Industrial Informatics, INDIN 2018*, pp. 212–217. Institute of Electrical and Electronics Engineers Inc. (2018)
10. Laloix, T., Iung, B., Voisin, A., Romagne, E.: Parameter identification of health indicator aggregation for decision-making in predictive maintenance: Application to machine tool. *CIRP Ann.* **68**, 483–486 (2019). <https://doi.org/10.1016/j.cirp.2019.03.020>
11. Luo, M., Xu, Z., Chan, H.L., Alavi, M.: Online predictive maintenance approach for semiconductor equipment. In: *IECON Proceedings (Industrial Electronics Conference)*, pp. 3662–3667 (2013)
12. Gerber, T., Martin, N., Mailhes, C.: Time-Frequency Tracking of Spectral Structures Estimated by a Data-Driven Method. <https://doi.org/10.1109/TIE.2015.2458781>

13. Calabrese, F., Regattieri, A., Botti, L., Mora, C., Galizia, F.G.: Unsupervised fault detection and prediction of remaining useful life for online prognostic health management of mechanical systems. *Appl. Sci.* **10**, 4120 (2020). <https://doi.org/10.3390/app10124120>
14. Lei, Y., Li, N., Gontarz, S., Lin, J., Radkowski, S., Dybala, J.: A model-based method for remaining useful life prediction of machinery. *IEEE Trans. Reliab.* **65**, 1314–1326 (2016). <https://doi.org/10.1109/TR.2016.2570568>
15. Buddhakulsomsiri, J., Siradeghyan, Y., Zakarian, A., Li, X.: Association rule-generation algorithm for mining automotive warranty data. *Int. J. Prod. Res.* (2006). <https://doi.org/10.1080/00207540600564633>
16. Rygielski, C., Wang, J.C., Yen, D.C.: Data mining techniques for customer relationship management. *Technol. Soc.* **24**, 483–502 (2002). [https://doi.org/10.1016/S0160-791X\(02\)00038-6](https://doi.org/10.1016/S0160-791X(02)00038-6)
17. Agard, B., Kusiak, A.: Data mining for subassembly selection. *J. Manuf. Sci. Eng. Trans. ASME.* **126**, 627–631 (2004). <https://doi.org/10.1115/1.1763182>
18. Antomarioni, S., Pisacane, O., Potena, D., Bevilacqua, M., Ciarapica, F.E., Diamantini, C.: A predictive association rule-based maintenance policy to minimize the probability of breakages: application to an oil refinery. *Int. J. Adv. Manuf. Technol.* **105**(9), 3661–3675 (2019). <https://doi.org/10.1007/s00170-019-03822-y>
19. Dudani, S.A.: The distance-weighted k-nearest-neighbor rule. *IEEE Trans. Syst. Man Cybern.* **SMC-6**, 325–327 (1976). <https://doi.org/10.1109/TSMC.1976.5408784>
20. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: *Proceedings of the 13th International Joint Conference on Artificial Intelligence* (1993). <https://doi.org/10.1109/TKDE.2011.181>
21. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. *SIGMOD Rec. (ACM Spec. Interes. Gr. Manag. Data)*. **29**, 1–12 (2000). <https://doi.org/10.1145/335191.335372>