A Machine Learning with an Inlier/Outlier Separation Approach for the Prediction of Wagon Maintenance Times

Josemar Coelho Felix[1], Vanessa Miranda Oliveira[1], Rodrigo Silva[2]

[1] Graduate Program in Computer Science, Universidade Federal de Ouro Preto, Brazil
josemar.felix@aluno.ufop.edu.br,vanessa.miranda@aluno.ufop.edu.br
[2] Department of Computer Science, Universidade Federal de Ouro Preto, Brazil
rodrigo.silva@ufop.edu.br

Abstract. Time spent in wagons maintenance consumes a significant part of a rail freight company's budget. Thus, knowing how much time it is going to be spent in a maintenance procedure is critical for their management and planning. A common approach used to predict these time expenditures is the so called chronoanalysis. Despite their wide spread use, they may be inaccurate in some scenarios. Thus, in this paper, we try to replace it with machine leaning models which did not work at first. Then we propose a methodology that uses the chronoanalysis to divide the maintenance procedures into outliers and inliers. Hence, we were able to create independent models for each class. With this approach, the average mean absolute error was reduced from about 6 man-hour to a little above 2 man-hours. The best tested configuration presented an average mean absolute error of 0.417 man-hours compared with a 4.490 man-hours from the chronoanalysis.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; I.2.6 [**Artificial Intelligence**]: Learning

Keywords: machine learning, maintenance, outlier detection, regression

1. INTRODUCTION

Maintenance is a fundamental activity in the industry and its main goal is to retain materiel in a serviceable condition or to restore it to serviceability [Office of the Chairman of the Joint Chiefs of Staff 2021; Eur]. Thus, the ability of predicting, with accuracy, maintenance times is critical for the production planning in many types of industry.

The study of times and movements is the systematic study of work systems to, among other things, determine the time spent by a qualified and trained person, at an average pace, to perform a specific task [Moura and Liu 2014]. Broadly speaking, it consists of timing multiple people performing a specific task to establish the standard time for that task. With these standard times, managers can predict the duration of each task involved in the maintenance and estimate the overall maintenance time, given the amount of available workforce.

This methodology is known in industrial practice as chronoanalysis [Coelho et al. 2021]. Even though it is largely used, it comes with criticisms. For instance, the fact the people performing the task know that they are being timed, may affect their performance. Besides, as it will be shown in Section 3, the computation of standard times depends on subjective knowledge, which may lead to even larger errors.

Nowadays, with the widespread use of information technology, most organizations have a historical record of all performed tasks, including information about when the task was performed and how long

it took. The existence of such databases opens the door for the application of machine learning models for the completion time prediction of many activities, including maintenance and transport assistance [Liyanage 2007]. In addition, the use of machine learning models allows the use of attributes other than the tasks themselves which may improve the prediction quality.

In this context, in this work we propose a machine learning based approach for the specific problem of estimating wagon maintenance times. The modern railway system is widely recognized as one of the most sustainable modes of transport [Cipolletta et al. 2021] and a typical train may pull from 5 to 150 wagons. Hence, time spent in wagons maintenance represents a significant part of what rail freight companies do. The data used in this work was provided by the Brazil-based company MRS® which is is responsible for the maintenance of approximately 4,500 wagons per year.

Initial analysis of the data shows that the maintenance times distribution presents a long tail which tends to increase the complexity of the regression problem. To alleviate this, in addition to the use o machine learning models, we propose a pre-processing method that classifies maintenance procedures into inliers or outliers. Thus, we can create independent models for each class improving the prediction accuracy significantly. The main contributions of this work are listed below:

—An Inlier/Outlier separation method based on chronoanalysis which simplifies the learning task.
—A machine learning based methodology for the prediction of wagon maintenance times;
—A study about the influence of additional variables in the prediction process.

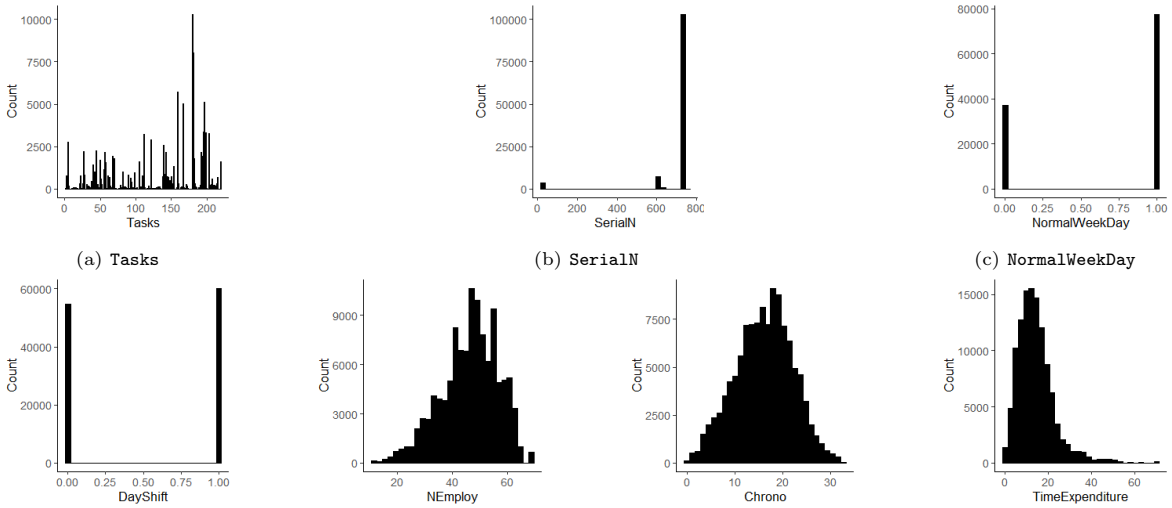
2. THE DATA SET

As mentioned in the previous section, the data set used in this work is real data provided by the company MRS®. It consists of 114,815 maintenance procedures performed on 2,820 wagons in 2019. Each maintenance procedure has the following attributes attached:

—`Tasks` $\in \{0,1\}^{220}$: The tasks performed in the maintenance procedure using one-hot encoding. The company has defined 220 different standard tasks which can be part of a maintenance procedure.
—`SerialN` $\in \mathbb{N}$: The serial number of the wagon which is an indicative of its age and manufacturing conditions.
—`NormalWeekDay` $\in \{0,1\}$: Whether the maintenance was performed in a normal week day, 1, or in a weekend or holiday, 0. In a normal weekday, as opposed to weekends and holidays, the administrative sector of the company is present.
—`DayShift` $\in \{0,1\}$: Whether the maintenance was performed between 7am and 7pm or not.
—`NEmploy` $\in \mathbb{N}$: Number of employees available in the shift.
—`Chrono` $\in \mathbb{R}$: Estimated time in man-hours computed with chronoanalysis. See Section 3 for more details.
—`TimeExpenditure` $\in \mathbb{R}^+$: Time spent in the overall maintenance procedure given in man-hours.

Figure 1 shows the distributions of the variables and the following aspects are worth highlighting: (i) The distribution of tasks is not uniform; (ii) Most of the wagons have a high `SerialN`, indicating that the have about the same age; (iii) As expected, most of the the samples falls into the weekdays category; (vi) `Chrono` seems to have a well behaved bell shape distribution; (v) `NEmploy` also has a bell shape distribution, however, with a negative skew.

Figure 1g shows the distribution of the `TimeExpenditure`, the target variable. It can be seen that has a significant positive skew with a long tail in the increasing direction. From this figure, it can already be seen that the chronoanalysis will not be able to predict `TimeExpenditure` accurately in these tail cases. In the next section, the chronoanalysis is presented in more detail.



(a) `Tasks` (b) `SerialN` (c) `NormalWeekDay`

(d) `DayShift` (e) `NEmploy` (f) `Chrono` (g) `TimeExpenditure`

Fig. 1: Attribute distributions

3. CHRONOANALYSIS

The chronoanalysis is an old technique that comes from the theory of time and motion studies [Hendry 1947]. Despite being old, it has survived the test of time and it remains in use today including at the MRS®.

The chronoanalysis ultimate goal is to establish a standard of efficiency [Hendry 1947]. In this sense, it starts by determining the time spent by a qualified and trained person, at an average pace, to perform a specific task. Mathematically, it determines this, so called, standard time with Equation 1 [Moura and Liu 2014].

$$StandardTime = M_A + M_A * (E_F + E_M + M + CA_T + CA_A + I_F + I_U + V) \quad (1)$$

where M_A is the average of the time spent by different, trained, employees to perform a given task. E_F is the level of physical effort, E_M is the level of mental effort, M represents the monotony of the task, CA_T represents the thermal conditions, CA_A represents the Atmospheric conditions, I_F is the influence of noise, I_U is the influence of humidity, and V is the influence of vibration.

Apart from M_A, the values of all the other terms are defined subjectively by an expert.

It can be said that two factors may affect the accuracy of the chronoanalysis adversely: (i) reliance on expert knowledge, and (ii) the effect of the process itself on the employee performing the task. Since, over the process, employees know they are being timed, that may affect how they perform the task.

4. OUTLIER DETECTION

It is hard to present a general and formal definition of what an outlier is. For instance, the authors of [Ayadi et al. 2017] give 12 different outlier definitions collected from the literature. Broadly speaking, however, an outlier is a point that is significantly dissimilar to other data points or a point that does not present the typical behavior of other points.

Outlier detection methods are getting increasingly popular in the industry in applications such as

identifying faulty equipment [Yun et al. 2016] and in predictive maintenance detecting early signs of possible shutdowns in manufacturing systems [Choi et al. 2022].

In this work specifically, outlier detection methods could be useful to identify wagons with maintenance procedures that notably differ from the regular cases and, therefore, will present different maintenance times. In other words, these methods should be able to identify the cases in the tails of the `TimeExpenditure` distribution shown in Figure 1g. The hypothesis is that, if an independent model is created for the outliers, the problem of predicting maintenance times will be easier to solve. The problem is, however, to know what is an outlier beforehand, i.e., before knowing the actual `TimeExpenditure` for the given maintenance procedure.

One way to solve this problem is to look at the other attributes. This approach would require an outlier detection method for high-dimensional data. More specifically, a method for 224 attributes (one-hot encoding of 220 tasks plus 4 additional variables). Outlier detection in high-dimensional problems, however, suffers from a number of issues such as the combinatorial explosion, the increased sparsity of the data, and the bias of scores induced by the use of different units in each dimension [Souiden et al. 2022].

Therefore, to avoid the issues related to the high-dimensionality of our application, in this work we employ a simple single-dimensional method, commonly used with box-plots, applied to the chronoanalysis estimation. Even though one of the main hypothesis of this work is that the chronoanalysis leads to inaccuracies in the time prediction, it may be possible that it is a good enough estimation to predict outlier maintenance procedures. In this context, the proposed outlier detection procedure works as follows:

(1) Collect a training set, $\mathcal{T}_r$.
(2) Compute the first quartile, $Q1(\mathcal{T}_r)$, the third quartil, $Q3(\mathcal{T}_r)$, and the interquartile range, $IQR(\mathcal{T}_r) = Q3(\mathcal{T}_r) - Q1(\mathcal{T}_r)$, for the chronoanalysis estimations.
(3) Given a new maintenance procedure, $\mathbf{m}_i$, and its respective maintenance time estimated by the chronoanalysis, $chrono(\mathbf{m}_i)$, classify $\mathbf{m}_i$ as an outlier if $chrono(\mathbf{m}_i) < Q1(\mathcal{T}_r) - 1.5 \times IQR(\mathcal{T}_r)$ or $chrono(\mathbf{m}_i) < Q3(\mathcal{T}_r) + 1.5 \times IQR(\mathcal{T}_r)$. Otherwise, $\mathbf{m}_i$ is considered an inlier.

With the outlier detection method defined, in the next section, the proposed approach for the prediction of maintenance times is presented.

5. MACHINE LEARNING WITH INLIER OUTLIER SEPARATION FOR THE PREDICTION OF WAGON MAINTENANCE TIMES

Based on the fact that maintenance times have a long tail distribution, the proposed approach here is based in the hypothesis that if we build independent models for inliers and outliers, we will be able to produce more accurate results with simpler models.

Given a training set, $\mathcal{T}_r$, the training phase of the proposed approach can be describe as follows:

(1) Build an outlier training set, $\mathcal{T}_r^{(o)}$ and an inlier training set, $\mathcal{T}_r^{(i)}$, by applying the outlier detection procedure presented in Section 4 to $\mathcal{T}_r$.
(2) Build the inlier model, $M^{(in)}$, with $\mathcal{T}_r^{(i)}$.
(3) Build the outlier model, $M^{(out)}$, with $\mathcal{T}_r^{(o)}$.

Thus, given a new maintenance procedure, $\mathbf{m}_i$, the predicted time expenditure, $PTimeExpenditure$, is give by Eq. 2.

$$PTimeExpenditure(m_i) = \begin{cases} M^{(out)}(\mathbf{m}_i) & \text{if } chrono(\mathbf{m}_i) < Q1(\mathcal{T}_r) - 1.5 \times IQR(\mathcal{T}_r) \\ M^{(out)}(\mathbf{m}_i) & \text{if } chrono(\mathbf{m}_i) > Q3(\mathcal{T}_r) + 1.5 \times IQR(\mathcal{T}_r) \\ M^{(in)}(\mathbf{m}_i) & \text{otherwise.} \end{cases} \quad (2)$$

The next section presents the experimental setup to evaluate the proposed approach. The main goal is to test whether the outlier detection will in fact lead to more accurate models. In addition, we investigate the performance of 8 different machine learning models and evaluate the effect of the additional variables `SerialN`, `NormalWeekDay`, `DayShift`, and `NEmploy`.

6. EXPERIMENTAL SETUP

As mentioned in the previous section, the goal of the experimental setup, that will be presented in this section, is to evaluate three aspects of the proposed approach. They are:

(1) **The effect of the Inliers/Outliers Separation Strategy:** In this regard, we want to verify whether the proposed strategy of outlier detection and the independent generation of inlier and outlier models leads to more accurate predictions.
(2) **The effect of the Machine Learning Models:** Another factor that may affect the performance of the proposed approach is the used machine learning model. In this context, we will test the following six popular algorithms, all available in the `Scikit Learn` [Pedregosa et al. 2011] and the `XGBoost` [XGB] libraries: (i) Decision Tree (DT), (ii) Bagging (B), (iii) Random Forest (RF), (iv) AdaBoost (AD), (v) Gradient Boosting (GB), (vi) Extreme Gradient Boosting (XGB).
(3) **The Effect of Additional Attributes:** The time estimation given by the chronoanalysis is computed by adding up the standard times of each activity involved in the maintenance procedure. With machine learning algorithms we can test whether the use of the available additional attributes can help with the predictions. In this work, we will test the use of the following additional attributes defined in Section 2: (i) `SerialN`, (ii) `NormalWeekDay`, (iii) `NEmploy`, (iv) `DayShift`.

Given the above factors, a full factorial design of experiments [Mongomery 2017] was implemented. For each combination of factor levels a grid search was performed to find the best parameters for the respective machine learning model in terms of the Mean Absolute Error (MAE).

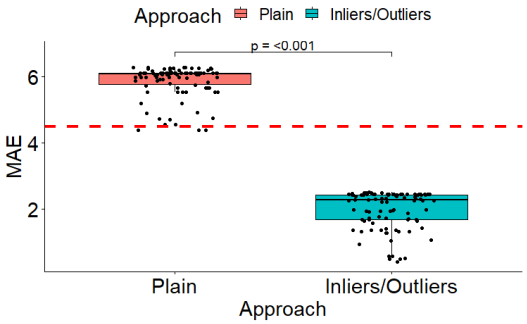
Once the best configuration for each model is found, its 5-fold cross validation MAE is reported. The Friedman test [Friedman 1940] was used to detect differences in groups across the performed experiments. When the Friedman test detects a difference among the groups, for a confidence level of $\alpha = 0.05$, the Dunn's test [Dunn 1964] is used a *post hoc* test. In the next section, the results obtained with these optimized machine learning models are presented.

7. RESULTS

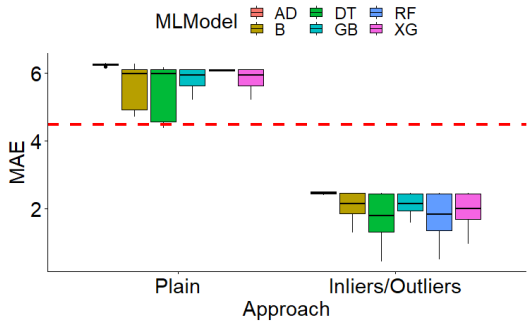
7.1 The effect of the Inlier/Outlier Separation Strategy

Figure 2a shows the boxplots of the results obtained when the Inlier/Outlier separation approach was used against the plain approach. The dotted red line indicates the average error of the chronoanalysis.

It can be seen that the average error for the plain approach was around 6 man-hours. Meanwhile, the proposed approach was able to reduce the average error to less than three hours. Besides, most of the experiments using the plain approach had and average error higher than the the average chronoanalisys error (dotted red line). On the other hand, every tested scenario under Inlier/Outlier approach had a lower error than the chronoanalysis.



(a) Effect of the Inlier/Outlier approach overall in the Mean Absolute Error (MAE). Each dot represents one of the experiments and the dotted red line represents the chronoanalysis average error. The mark with the p-value, p, indicates whether the Dunn's test detected a difference between the groups for a significance level $\alpha = 0.05$



(b) Effect of the Inlier/Outlier approach by machine learning model in the Mean Absolute Error (MAE). The dotted red line represents the chronoanalysis average error

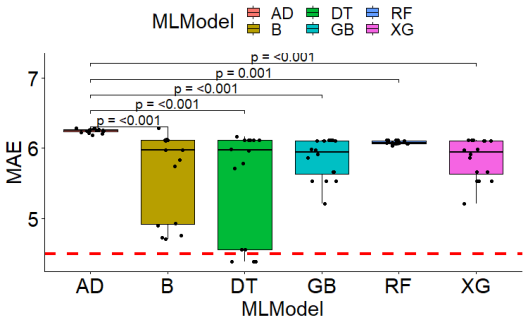
Fig. 2: Inlier/Outlier approach.

These results indicate that dividing the problem in two, one for the inliers and other for the outliers, allow the models to achieve better results. The Dunn's test also attest the superiority of the Inlier/Outlier separation approach. In Figure 2b, we split the results by machine learning (ML) model. It can be seen that the Inlier/Outlier separation approach was beneficial for all the tested ML model.

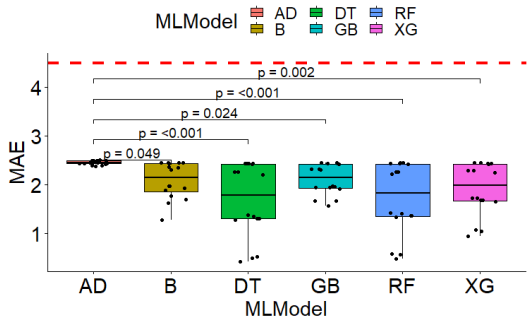
7.2 The effect of the Machine Learning Models

Figure 3a shows the performance of the different ML models under the plain approach. The dotted red line indicates the average error of the chronoanalysis. As it can be seen, the Dunn's test indicates that the AD performance was different from the other ML model but there was no difference among B, DT, GB, RF and XG.

Figure 3b shows the performance of the tested ML models under the Inlier/Outlier approach. It can be seen that, under this scenario, independently from the ML model used, the results were better than the chronoanalysis on average. Again, the AD model presented a worse performance when compared against the other models. The Dunn's test did no detect any difference among the other models.



(a) Effect of the machine learning models in the plain approach.



(b) Effect of the machine learning models in the In/Out approach.

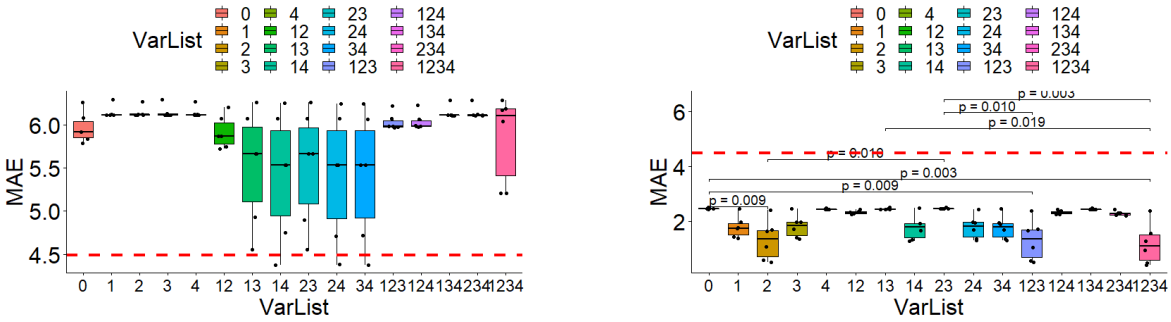
Fig. 3: Machine learning Methods.Each dot represents one of the experiments and the dotted red line represents the chronoanalysis average error. The marks with p-values, p, indicate whether the Dunn's test detected a difference between the groups for a significance level $\alpha = 0.05$

7.3 The effect of the Additional Variables

Figure 4a presents the box plots of the results obtained by the different combinations of the additional variables for the plain approach. The label 0 indicates that no additional variable was used. Label 1 refers to `SerialN`, 2 to `NormalWeekDay`, 3 to `NEmploy`, and 4 to `DayShift`. The dotted red line indicates the average error of the chronoanalysis. Although a performance variation can be noticed for the different configurations, the Friedman test did not detect differences among the the groups of experiments. In addition, the average performance in these scenarios was always worse than the chronoanalysis shown in the red dotted line.

Figure 4b presents the box plots of the results obtained by the different combinations of the additional variables for the Inlier/Outlier approach. The Dunn's test indicates that the configurations using the variable 2 (`NormalWeekDay`), $1, 2, 3$ (`SerialN`, `NormalWeekDay` and `NEmploy`) and $1, 2, 3, 4$ (`SerialN`,`NormalWeekDay`, `NEmploy` and `DayShift`) are better than the configurations that use no addition variable 0.

Since no effect was perceived in the plain case but it was perceived in the Inlier/Outlier case, it is possible that the additional variables affect the inliers and outlier cases differently. Thus, under the plain approach, the models were not able to take advantage of any pattern regarding the additional variables. On the other hand, when we have different models for inliers and outliers, the ability of taking into account other variables became more important.



(a) Effect of the additional variables in the plain approach.

(b) Effect of the additional variables in the Inlier/Outlier approach.

Fig. 4: Additional Variables. The label 0 indicates that no additional variable was used. Label 1 refers to `SerialN`, 2 to `NormalWeekDay`, 3 to `NEmploy`, and 4 to `DayShift`. The dotted red line indicates the average error of the chronoanalysis. Each dot represents one of the experiments and the marks with p-values, p, indicate whether the Dunn's test detected a difference between the groups for a significance level $\alpha = 0.05$

Table I shows the top 10 configurations considering the 5-fold cross-validation MAE. It can be seen that all of them presented errors below 2 man-hours. The models in the top 7, all presented errors below 1 man-hour which is a considerable improvement over the 4.49 man-hours for the chronoanalysis.

8. CONCLUSION

In this paper we propose a machine learning with an Inlier/Outlier separation approach for the prediction of wagon maintenance times. The proposed approach was compared against the industry standard, the so called chronoanalys, and against a plain approach which uses machine learning models but no strategy for the outliers. The results show that the ML models alone are not more accurate than the chronoanlysis. Nevertheless, when the proposed inlier/outlier separation approach was aggregated to the method, the results greatly improved. The average mean absolute error decreased from about 6 man-hours in the plain approach to less than 3 man-hours in the Inlier/Outlier approach. These results

Approach	MLModel	SerialN	NormalWeekDay	NEmploy	DayShift	MAE
Inliers/Outliers	DT	1	1	1	1	0.417
Inliers/Outliers	RF	1	1	1	1	0.482
Inliers/Outliers	DT	1	1	1	0	0.500
Inliers/Outliers	DT	0	1	0	0	0.517
Inliers/Outliers	RF	1	1	1	0	0.569
Inliers/Outliers	RF	0	1	0	0	0.583
Inliers/Outliers	XG	1	1	1	1	0.944
Inliers/Outliers	XG	1	1	1	0	1.05
Inliers/Outliers	XG	0	1	0	0	1.07
Inliers/Outliers	DT	1	0	0	1	1.27

Table I: Top 10 configurations an their respective 5-fold cross validation MAE

also improve over the chronoanalysis, which presented an average MAE of more than 4 man-hours. For comparison, the best tested configuration was the Decision Tree, with Inlier/Outlier separation, using all the additional variables. This configuration had an average 5-fold cross-validation error of 0.417 man-hour which is a great improvement over the chronoanalysis.

Acknowledgments

This work has been supported by the Brazilian research agencies, FAPEMIG, CAPES, CNPq, and the Federal University of Ouro Preto (UFOP). The authors also thank MRS® for providing the data.

REFERENCES

European Federation of National Maintenance Societies. `http://www.efnms.eu/`. Accessed: 2022-07-06.

XGBoost. `https://xgboost.readthedocs.io/en/stable/`. Accessed: 2022-07-20.

Ayadi, A., Ghorbel, O., Obeid, A. M., and Abid, M. Outlier detection approaches for wireless sensor networks: A survey. *Computer Networks* vol. 129, pp. 319–333, 2017.

Choi, H., Kim, D., Kim, J., Kim, J., and Kang, P. Explainable anomaly detection framework for predictive maintenance in manufacturing systems. *Applied Soft Computing* vol. 125, pp. 109147, 2022.

Cipolletta, G., Delle Femine, A., Gallo, D., Luiso, M., and Landi, C. Design of a stationary energy recovery system in rail transport. *Energies* 14 (9), 2021.

Coelho, W. L. V., de Souza Matos, I., Gauto, R. F., and Bueno, A. F. Chrono-analysis: Study applied to a processing unit of swine in brazil. *Zeiki-Revista Interdisciplinar da Unemat Barra do Bugres* 2 (1): 4–28, 2021.

Dunn, O. J. Multiple comparisons using rank sums. *Technometrics* 6 (3): 241–252, 1964.

Friedman, M. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics* 11 (1): 86–92, 1940.

Hendry, J. W. *A Manual of Time and Motion Study: A Practical Guide to the Measurement of Human Endeavor in Industry and to the Development of Productive Efficiency.* I. Pitman, 1947.

Liyanage, J. P. Operations and maintenance performance in production and manufacturing assets: The sustainability perspective. *CIRP Journal of Manufacturing Technology Management*, 2007.

Mongomery, D. Montgomery: design and analysis of experiments. *John Willy & Sons*, 2017.

Moura, D. A. and Liu, R. D. Sistemas de produção, o uso de ferramentas adequadas para aumento de competitividade na área de tempos e métodos. *Revista Gestão Industrial* 10 (1), 2014.

Office of the Chairman of the Joint Chiefs of Staff. DOD dictionary of military and associated terms. Tech. rep., Washington DC: The Joint Staff, 2021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research* vol. 12, pp. 2825–2830, 2011.

Souiden, I., Omri, M. N., and Brahmi, Z. A survey of outlier detection in high dimensional data streams. *Computer Science Review* vol. 44, pp. 100463, 2022.

Yun, U., Ryang, H., and Kwon, O.-C. Monitoring vehicle outliers based on clustering technique. *Applied Soft Computing* vol. 49, pp. 845–860, 2016.