Predictive Analysis Methodology for Industrial Systems: Application in Supplier Delays Prediction

1st Mohamed Aziz Zaghdoudi
FEMTO-ST institute
Univ. Bourgogne Franche-Comté
ENSMM, CNRS
Besançon, France
mohamed.zaghdoudi@femto-st.fr

2nd Sonia Hajri-Gabouj
Laboratoire d'Informatique pour les Systèmes Industriels
Nat. Inst. of Applied Science and Technology
University of Carthage
Tunis, Tunisia
sonia.hajri@insat.ucar.tn

3rd Christophe Varnier
FEMTO-ST institute
Univ. Bourgogne Franche-Comté
ENSMM, CNRS
Besançon, France
christophe.varnier@femto-st.fr

4th Noureddine Zerhouni
FEMTO-ST institute
Univ. Bourgogne Franche-Comté
ENSMM, CNRS
Besançon, France
noureddine.zerhouni@femto-st.fr

5th Feiza Ghezail
Laboratoire d'Informatique pour les Systèmes Industriels
Nat. Inst. of Applied Science and Technology
University of Carthage
Tunis, Tunisia
feiza.ghezail@insat.ucar.tn

***Abstract*—In this paper, we develop a methodology that proposes a clear and reliable approach for collecting, analysing, and then improving the quality of industrial data. In addition, we put forward a machine learning approach based on data of good quality that allows the development of predictive models for decision support and industrial performance improvement. A case study in supply chain will be used to validate the suggested methodology. During the study period, the partner company experienced a high number of delivery delays from its suppliers. We propose using the company's order history, along with an appraisal of its quality, to create a decision-making tool that can predicts supplier delays.**

***Index Terms*—Predictive Analytics, Supply Chain, Machine Learning, Data Quality**

I. Introduction

Massive volumes of data are generated at a high rate by industrial systems. Different types, sources, and forms distinguish these data. Improving industrial performance, however, necessitates predictive data analysis. The latter forecasts the system's future state and anticipates the onset of unfavorable occurrences. The use of prediction algorithms to estimate the future state of systems is known as predictive analysis [1]. Predictive analysis, according to Rajaraman [2], is a method that extrapolates from known data and forecasts what is likely to happen in the near future.

On the other hand, current market trends such as uncertainty, globalization, a constantly changing corporate environment and customer behavior, and the demand for flexibility and security have increased the supply chain's complexity and interdependencies [3]. As a result, data mining and predictive analysis approaches offer a way to create a cost-effective supply plan that reduces overstocking and boosts profits.
Machine learning (ML) techniques are increasingly used to perform complex industrial tasks. However, different methodological approaches have been adopted to exploit industrial data in predictive analysis. Indeed, data exploitation methodologies vary according to several factors, namely the system studied (production, supply, sale, storage etc.), the nature of the data (signals, images, tables etc.) and the objective of the study(prediction or partitioning). In [5], the author proposes a methodology for binary or multi-class classification of supplier delays. His methodology consists of 5 steps, namely data collection and exploration, definition of evaluation metrics, feature selection, data preprocessing and, algorithm evaluation.

1) Data collection and exploration includes system analysis and data collection and focuses on understanding the production system, its data generation, limitations and, predictive purpose.
2) The definition of performance and metrics focuses on transforming the prediction objective into appropriate metrics for evaluating the performance of the model throughout its development.
3) During feature selection, collected data elements are transformed into formats suitable for machine learning algorithms, incorrect data entries or noise are removed, and data features (variables) are added using available raw data, domain knowledge and, experience.
4) Data preprocessing incorporates ML-specific transformations such as data scaling or normalization and the option to apply resampling on each considered subset to reduce the negative impacts of data imbalance on the resulting prediction performance.
5) Algorithm comparison and evaluation is the final step in which the algorithm parameter grids are used to evaluate and compare the prediction performance of the various

algorithms. In addition, post-processing by means of a threshold setting for binary classification is incorporated to shift the prediction performance in the direction of the main performance metric (accuracy) at the expense of various less important metrics.

In [8], the authors suggest adding a data comprehension step that consists of visually inspecting the data using graphs to understand the relationships between variables.

In [9], the author proposes to ensure the quality of the data during its collection. He suggests, therefore, to acquire the data through sources that he calls 'reliable'. We cite as examples the Failure Modes and Effects Analysis, fault trees and, sensors. The author proposes a data acquisition chain for industrial sensors that ensures the collection, selection, transformation, and storage of signals.

In addition, [6] proposes a first step to understand the application domain. This step consists of determining the achievable goals based on the situation and requirements from the business perspective in order to obtain potential benefits. They suggest conducting this step based on the history of the company in question and interviews with its managers [6].

It is true that the researchers' work has resulted in a path that adequately describes the actions to be taken (from data collection to evaluation of learning models) and maximizes the performance of predictive analysis. In the literature, and for machine learning projects, researchers always start from the assumption that the data are of good quality, available and adapted to the studied problem. In reality, however, the process of data collection is complicated, and the criteria for judging the quality of the data are not well developed.

It is therefore necessary to assess the quality of the data before starting the predictive analysis because poor quality data can lead to erroneous results.

II. Proposed Methodology

Our methodology combines data quality assessment techniques and machine learning approaches for decision support (Fig. 1).

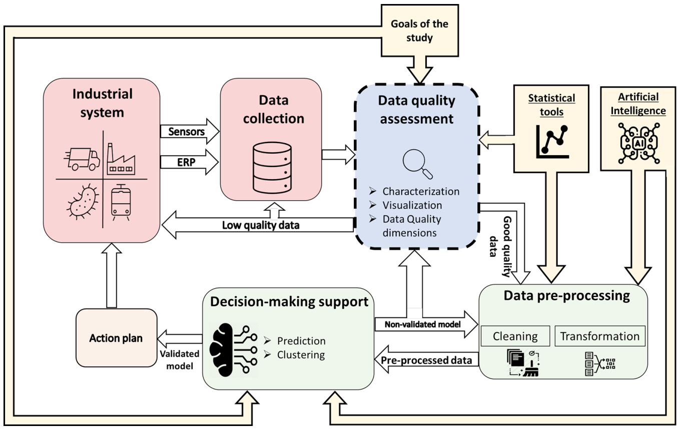


Fig. 1. General methodology for predictive analysis of industrial data.

The studied system represents the source of the industrial data and, we seek, finally, to improve its performance through our methodology.

We start by collecting data from several sources. We cite as examples the industrial sensors, the Enterprise resource planning (ERP) of the company, or the history of the transactions carried out by the company. We then propose to add a new step to evaluate the quality of the collected data. This step is important and requires statistical tools as well as approaches to DQ dimensions. We question the quality of the data and its adaptability to the objectives of the study. The detection of anomalies in the data can be performed in several steps namely characterization, visualization, and study of the dimensions of data quality. Characterization consists of extracting the characteristics of the data set such as the size, type, number of variables, and the time horizon of the study. Visualization helps to better understand the nature of the data and to detect anomalies using different types of graphs (curves, box plots, distribution graphs, etc.) We then propose to proceed to the data preprocessing step. This step aims at manipulating or deleting data before its use in order to guarantee or to improve performance. We also apply transformations on the data in order to adapt them to the study. The next step is the decision support step. We develop learning algorithms oriented towards prediction or partitioning.

- The prediction consists in predicting a missing element of a data set or the future.
- Partitioning aims at determining a class or categorical label for an element of an unlabeled data set.

The learning step is strongly related to the preprocessing and DQ study steps. If good prediction or partitioning performances could not be achieved, we have to revise the previous steps in order to train data able to perfectly describe the studied system. Finally, we use the modeling results to help make appropriate strategic decisions and establish an action plan to improve the performance of the studied system.

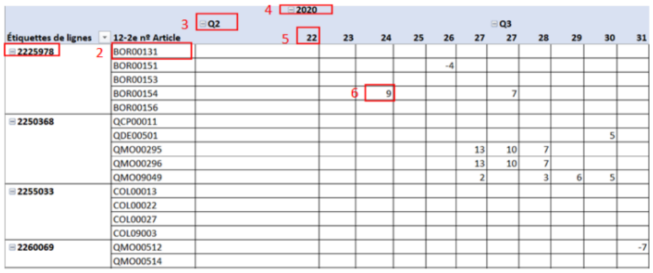
III. Case study

Our partner is an industrial company active in the furniture manufacturing market. To meet its needs in terms of primary and semi-finished materials, it is in constant interaction with dozens of suppliers. However, due to the Covid-19 pandemic, the company has had a high number of supplier delays. We have at our disposal the history of orders made by the company during the year from May 25, 2020 to May 28, 2021. This data is organized in the form of a pivot table. For each order, we have the delta_delivery which is the difference between the delivery date and the promised delivery date by the supplier (1).

$$delta\_delivery = Y - X \tag{1}$$

with Y is the delivery date and X is the delivery date promised by the supplier.

The database is obtained following a manipulation, performed by the company, on the data acquired directly from the ERP. In rows, we have the different products ordered and grouped by their suppliers and in columns, we have the weeks of the study period. We have, for each week, the average delta_delivery of the deliveries received of each product during that week. We can ,therefore, conclude that a positive delta_delivery reflects a delay in delivery while a negative delta_delivery means that the order was received before the promised delivery date. Fig. 2 is an example of a few lines of data received from the company. To further explain the structure of the database, box (6) represents the average delta_delivery of the deliveries of the product ’BOR00154’ provided by the supplier ’2225978’ during week 24 of the year 2020.



Étiquettes de lignes	12-2e nº Article	22	23	24	25	26	27	27	28	29	30	31
		⊟2020										
		⊟Q2						⊟Q3				
⊟2225978	BOR00131											
	BOR00151					-4						
	BOR00153											
	BOR00154			9				7				
	BOR00156											
⊟2250368	QCP00011											
	QDE00501										5	
	QMO00295						13	10	7			
	QMO00296						13	10	7			
	QMO09049						2		3	6	5	
⊟2255033	COL00013											
	COL00022											
	COL00027											
	COL09003											
⊟2260069	QMO00512											-7
	QMO00514											

Fig. 2. Data initially received from the company

1. Supplier ID 2. Product ID 3. Quarter 4. Year 5. week 6. Average delta_deliveries

Tab.I represents a characterization of the studied data

TABLE I
DATA CHARACTERISTICS

Characteristic	Value
Study period	From 22nd week of 2020 to 22nd week of 2021 (May 25, 2020 - May 28, 2021)
Number of Delta_deliveries recorded	22389
Number of suppliers	102
Number of products	3926

A. Studying and improving data quality

In order to understand the trend in the data, we begin by examining the average Delta_delivery (Fig. 3). We notice that the curve of the average shows peaks during the study period for weeks 58, 48, 34 and 2.

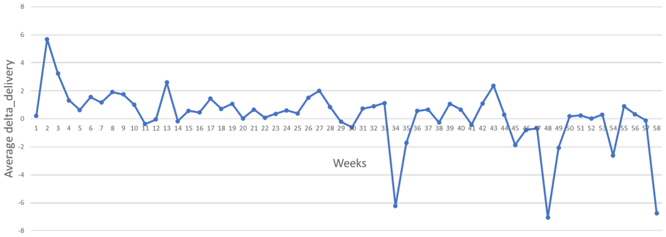


Fig. 3. Evolution of the average delta_delivery per week

We hypothesized that these peaks are caused by supply disruptions during the periods of lockdowns established following the Covid-19 crisis. After verification with the company, this hypothesis was rejected because these disruptions do not correspond to periods of work stoppage.

We then analyzed three dimensions of data quality, namely completeness, uniqueness, and consistency. Indeed, the measure of the accuracy of an attribute is the difference between its measured value and its true value. The actual value measure is often unavailable or difficult to extract. Low actuality does not present a problem for prediction algorithms. Indeed, the more we increase the horizon of the study, the more we will have a variety of learning cases through the history of the studied data. The results of the study of the DQ dimensions are illustrated in Tab.II.

TABLE II
DATA QUALITY DIMENSIONS

Completeness	Uniqueness	Consistency
0.098	2	2

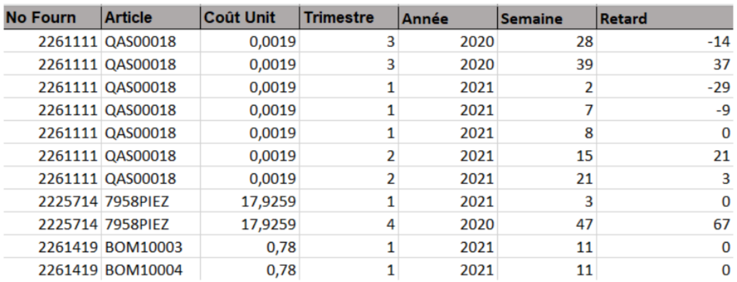
The dataset is 10% complete. In other words, 90% of the data is missing. In addition, we recorded a uniqueness index equal to two. We further inspected the dataset and found a duplicate provider with two different identifiers. Finally, the consistency assessment methodology allowed us to identify two instances of data bins that do not follow the data presentation rules. We conclude that the data has several anomalies in terms of data quality dimensions and structuring. In order to improve DQ, we applied our methodology. We decided to revise the data collection process by acting directly on the raw data extracted from the Enterprise Resource Planning (ERP). We then applied the following modifications to improve data quality.

1) Extract the data from the ERP in order to increase the number of orders studied and have more features that will be used for the modeling.
2) Filter the data to eliminate orders that were not received at the time of the study.
3) Change the comparison parameter since, initially, we considered delta_delivery as the difference between the delivery date and the promised delivery date. However, the problem with this last variable is that it is constantly updated by the procurement team in order to inform the system of a delivery that will be received later than expected. To solve this problem, we now use the ”original promised delivery date” field.

$$delta\_delivery = Y - X \tag{2}$$

where Y is the delivery date and X is the original promised delivery date by the supplier.
4) Modify the formulas for extracting weeks and quarters from the promised delivery dates. This manipulation allowed us to eliminate the redundancy of the weeks located between 2 quarters or between 2 years.
5) To solve the problem of duplication of suppliers, we use the variable ’supplier name’ as an identifier of the suppliers.

We obtained a data set with a different structure. All orders placed during the study period are organized in rows, and the variables of each order are organized in columns in the new modified dataset shown in (Fig. 4).



No Fourn	Article	Coût Unit	Trimestre	Année	Semaine	Retard
2261111	QAS00018	0,0019	3	2020	28	-14
2261111	QAS00018	0,0019	3	2020	39	37
2261111	QAS00018	0,0019	1	2021	2	-29
2261111	QAS00018	0,0019	1	2021	7	-9
2261111	QAS00018	0,0019	1	2021	8	0
2261111	QAS00018	0,0019	2	2021	15	21
2261111	QAS00018	0,0019	2	2021	21	3
2225714	7958PIEZ	17,9259	1	2021	3	0
2225714	7958PIEZ	17,9259	4	2020	47	67
2261419	BOM10003	0,78	1	2021	11	0
2261419	BOM10004	0,78	1	2021	11	0

Fig. 4. The dataset obtained after modifications

We chose to use the maximum number of features available for the study at this level as we will proceed to a variable selection work later. Therefore, we included the item code of the ordered item, the vendor number, the unit cost of the item, the week, the quarter, the year, and finally the delta_delivery of the order. We also notice that, with the exception of week 52 of 2020 and week 1 of 2021, which coincide with the year-end vacations, the curve of the progression of the average delta_deliveries per week shows that there are fewer peaks currently. In addition, after the adjustments were implemented, the average curve represents a lagging trend over the entire study period, demonstrating that the data better represent reality.

We re-examined the dimensions of the DQ after making the necessary modifications. In terms of completeness, we obtain a score of 1 for completeness. This is the highest possible score, indicating that the new data set has no missing values. In terms of uniqueness, we looked for duplicate identifiers at the supplier and product level, but found none. Finally, we evaluated the consistency of the modified dataset using the same methods. The errors initially found did not recur, and we found no further inconsistencies.

B. Data preprocessing

Data preprocessing includes actions that create a final dataset that will be fed to the learning models. The data are balanced and do not have missing fields or outliers according to the data quality study. Thus, we use the following preprocessing techniques :

- Normalization: The numerical data used for modeling are scaled to change their range of values and weight all variables equally. These numerical variables are transformed to a scale between 0 and 1 using the Min-Max normalization defined by the equation below.
$$\text{x'} = \frac{x - min(x)}{max(x) - min(x)} \tag{3}$$
with x' is the normalized variable and x is the raw variable before normalization
- Encoding of categorical variables: Most learning models only accept numerical variables. Thus, it becomes necessary to preprocess the categorical variables. We transfer these categorical variables into numerical variables so that the model can process them and extract valuable information. We encode, thus, the categorical variables with the following principle: if the variable has (n) possible values, it is encoded with numerical values between (0) and (n-1) with the conditions that the code is unique for each variable and that the similar variables have the same code.
- Data labeling: Supervised learning algorithms require labeled data. The labels will serve as a reference for our algorithms. For our binary classification, we give the value 1 for orders that arrived early or on time and the value 0 for orders that arrived late.
- Data splitting: The methodology followed requires data for training and data for algorithm evaluation. We adopted the 75-25 strategy by splitting the dataset into 75% for training and 25% for testing while keeping the same data distribution.

C. Assessment of learning

Learning evaluation metrics are well developed and studied in the literature [2] [4]. The validation of classification algorithms is done through the confusion matrix which gives

TABLE III
ALGORITHMS EVALUATION METRICS

Metric	Formula	Signification
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Number of correct predictions out of all predictions
Precision	$\frac{TP}{TP+FP}$	Ratio of correctly predicted on-time deliveries to predicted on-time deliveries
Recall	$\frac{TP}{TP+FN}$	Ratio of correctly predicted on-time deliveries to on-time deliveries that must be predicted
F1-score	$\frac{2*Recall*Precision}{Recall+Precision}$	The average of the accuracy and the Recall

the number of correct predictions (True positive and True negative) and wrong predictions (False positive and False negative) compared to the total number of predictions. In order to evaluate the prediction correctness of our models, we will use the following metrics (Tab.III).

D. Learning algorithms

In order to predict delivery delays, we have exploited 3 machine learning algorithms.

- Decision tree: A decision tree is a type of machine learning method that divides data into subspaces. For categorical data, each subspace is assigned a single class label, and for continuous data, a numerical value [10]. Using a top-down strategy, decision trees are built recursively. They have a root, as well as nodes, branches, and leaves. The decision tree technique works by detecting a collection of indications and categorizing the population into n classes. The construction of a decision tree is done according to the following principle: First, it selects the optimum indication to ensure a more accurate population division; then, the divided populations are distributed among the nodes. We repeat the process for each node (population subset) until no further splitting is possible. The population of the terminal nodes is of the same class. The categorization operation entails assigning a person to a terminal node based on a set of rules geared to that node [7].
- Random Forest: Because it is a robust approach in the presence of a large number of input attributes and a small number of samples available for learning, and because tree-based models are very easy to interpret, the Random Forest (RF) classification technique has recently been used for fault diagnosis in several engineering domains. The Random Forest technique is made up of many decision trees that classify the data frame individually. The same data is ranked by each decision tree based on its individual optimal distribution [11]. The usage of random forests can provide us with [11] a minimal (or no) overfitting: Because multiple trees are used, the main risk of overlearning is reduced. In addition, it provides us with high accuracy (the algorithm performs well on large databases, and as the quality of the data used for training improves, so does the accuracy)
- Naive Bayes Algorithm: The Naive Bayes classifier is a probabilistic classifier based on the application of Bayes' theorem with strong independence assumptions between features. A naive Bayes model is easy to build, without complicated iterative parameter estimation, which makes it particularly useful in industrial applications. Despite its simplicity, the naive Bayesian classifier often yields surprising results and is widely used because it often outperforms more sophisticated classification methods [12].

E. Evaluation of the algorithms and feature selection

We exploited the selected evaluation metrics to judge the performance of each model and to compare the models between them.

The tests were performed on a computer with a Windows 10 operating system, an intel i5/2.3 GHz processor, and 12 GB of RAM.

Tab.IV summarizes the metrics recorded for each model for the binary classification.

In terms of accuracy, precision, and F1 score, we see that the Random Forest model outperforms the other models. In other words, this approach is the most capable of avoiding false delivery advance alerts. Given their comparable operating principles, the ”decision tree” and ”random forest” models have almost identical performances. The accuracy of these two models indicates that they can make 76 correct predictions out of a possible 100.

The Naive Bayes model has a Recall of 83%, but its perfor-

TABLE IV
ALGORITHMS EVALUATION

	Decision Tree	**Random Forest**	**Naive Bayes**
Accuracy	75.73%	76,02%	70,09%
Precision	76.35%	76,43%	68,28%
Recall	78.91%	79,55%	83,06%
F1-score	77.61%	77,96%	74,95%
Parameters	Max depth: 42 number of leaf_nodes: 4498 criterion: Gini	Number of trees: 100 Max depth: 100 Criterion: Gini	’alpha’: 1,0 ’class_prior’: None ’fit_prior’: True

TABLE V
FEATURE SELECTION

		Test1	Test2	Test3	Test4	Test5	Test6	Test7
Feature Importance	Supplier ID	7.5%	1.1%	▨	▨	50.8%	51%	50.8%
	Product ID	27.4%	1.7%	1.5%	▨	▨	▨	▨
	Unit Price	1%	▨	▨	▨	▨	▨	▨
	Quarter	1.8%	▨	▨	▨	▨	▨	2.3%
	Year	24.1%	24.4%	23.3%	15%	3.4%	▨	3.4%
	Week	37.9%	62.7%	75.1%	84.9%	45.6%	48.9%	43.3%
Accuracy	Training accuracy	98.2%	97.8%	97.8%	61.3%	77.4%	77.1%	77.4%
	Test accuracy	72.2%	75.1%	74.4%	60.9%	75.7%	75.4%	75.7%

mance in terms of accuracy and precision is poor, making it unsuitable for predicting supplier delays.

Overfitting is a well-known problem with machine learning algorithms. When the model performs well with the training data but not with the test data, it is an overfitting problem. This can be recognized by comparing the accuracy of the model with the test and training data. Since the Scikit-Learn library has a tool for analyzing the importance of variables in this type of model, we chose to conduct a study to e select the relevant variables for our ”decision tree” model.

Table V summarizes the results of this study. We train the model using a subset of variables in each trial, noting the relevance of the variables while keeping an eye on the overfitting problem.

IV. CONCLUSION

In this paper, we are interested in the exploitation of data from industrial systems to predict future events in order to improve their performance. To analyze industrial data, we have developed a general data mining methodology that can be applied to different industrial systems. This methodology aims to design a decision support tool for industrial systems for the prediction of future events. We detail the different steps from data collection and quality assessment to the development of machine learning algorithms for predictive analysis.

In order to validate the proposed methodology, we applied it to a case study in the field of logistics. We conducted a study to characterize and inspect the quality of the data received from the company and then improve it. Indeed, we were able to build a basis for evaluating data quality based on characterization and study according to three dimensions of data quality, namely completeness, uniqueness, and consistency. This study allowed us to detect anomalies in the dataset and then correct them.

In the second part, we addressed the central issues of our project by developing machine learning models for the prediction of delivery delays. These models are then evaluated and compared to derive the best algorithms for our purposes. We found that the ’Decision Tree’ and ’Random Forest’ algorithms are the most efficient for the prediction task.

REFERENCES

[1] Brintrup, A., Pak, J., Ratiney, D., Pearce, T., Wichmann, P., Woodall, P., & McFarlane, D. (2020). Supply chain data analytics for predicting supplier disruptions: a case study in complex asset manufacturing. International Journal of Production Research, 58 (11), 3330-3341. https://doi.org/10.1080/00207543.2019.1685705

[2] Rajaraman, V. Big data analytics. Reson 21, 695–716 (2016). https://doi.org/10.1007/s12045-016-0376-7

[3] Darshit Parmar , Teresa Wu , Tom Callarman , John Fowler & Philip Wolfe (2010) A clustering algorithm for supplier base management, International Journal of Production Research, 48:13, 3803-3821, DOI: 10.1080/00207540902942891

[4] Baryannis, George & al. “Predicting supply chain risks using machine learning: The trade-off between performance and interpretability.” Future Gener. Comput. Syst. 101 (2019): 993-1004.

[5] De Krom, B. (2021) Supplier disruption prediction using machine learning in production environments (PhD thesis), Delft University of Technology.

[6] Huang, M., Bagheri. M. (2019) Predicting deviation in supplier lead time and truck arrival time using machine learning (Master thesis), Chalmers University of Technology.

[7] Karabadji, N.E.I., Seridi, H., Khelf, I., Laouar, L. (2012). Decision Tree Selection in an Industrial Machine Fault Diagnostics. In: Abelló, A., Bellatreche, L., Benatallah, B. (eds) Model and Data Engineering. MEDI 2012. Lecture Notes in Computer Science, vol 7602. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33609-6_13

[8] Kuhn, M. & Johnson. K. (2013) Applied predictive modeling. Springer.

[9] Medjaher, K. (2014) Contribution au pronostic de défaillances guidé par des données (PhD thesis), Université de Franche-Comté.

[10] Olivier J, Aldrich C. Use of Decision Trees for the Development of Decision Support Systems for the Control of Grinding Circuits. Minerals. 2021; 11(6):595. https://doi.org/10.3390/min11060595

[11] Vamsi, Inala Vivek et al. “Random Forest Based Real Time Fault Monitoring System for Industries.” 2018 4th International Conference on Computing Communication and Automation (ICCCA) (2018): 1-6.

[12] Vembandasamy, K., Sasipriya, R. and Deepa, E. (2015) Heart Diseases Detection Using Naive Bayes Algorithm. IJISET-International Journal of Innovative Science, Engineering Technology, 2, 441-444.