**Indonesian Journal of Electrical Engineering and Computer Science**
Vol. 26, No. 3, June 2022, pp. 1736~1746
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v26.i3.pp1736-1746 

# Knowledge discovery in manufacturing datasets using data mining techniques to improve business performance

**Amani Gomaa Shaaban[1], Mohamed Helmy Khafagy[2], Mohamed Abbas Elmasry[3], Heba El-Beih[4], Mohamed Hasan Ibrahim[2]**
[1]Department of Information Systems, Faculty of Computers and Information, Fayoum University, Fayoum City, Egypt
[2]Department of Computer Science and Information Systems, Faculty of Computers and Information, Fayoum University, Fayoum City, Egypt
[3]Department of Management Information Systems, Faculty of Business Administration, National Egyptian E-Learning University, Giza City, Egypt
[4]Faculty of Computers and Information, Menoufia University, Menoufia City, Egypt

**Article Info**

***Article history:***

Received Aug 23, 2021
Revised Mar 8, 2022
Accepted Mar 29, 2022

***Keywords:***

Apriori algorithm
Association rules
CHAID tree analysis
Crises management
Data mining
Knowledge discovery in databases technique
Manufacturing datasets

**ABSTRACT**

Recently due to the explosion in the data field, there is a great interest in the data science areas such as big data, artificial intelligence, data mining, and machine learning. Knowledge gives control and power in numerous manufacturing areas. Companies, factories, and all organizations owners aim to benefit from their huge; recorded data that increases and expands very quickly to improve their business and improve the quality of their products. In this research paper, the knowledge discovery in databases (KDD) technique has been followed, "association rules" algorithms "Apriori algorithm", and "chi-square automatic interaction detection (CHAID) analysis tree" have been applied on real datasets belonging to (Emisal factory). This factory annually loses tons of production due to the breakdowns that occur daily inside the factory, which leads to a loss of profit. After analyzing and understanding the factory product processes, we found some breakdowns occur a lot of days during the product lifecycle, these breakdowns affect badly on the production lifecycle which led to a decrease in sales. So, we have mined the data and used the mentioned methods above to build a predictive model that will predict the breakdown types and help the factory owner to manage the breakdowns risks by taking accurate actions before the breakdowns happen.



***Corresponding Author:***

Amani Gomaa Shaaban
Department of Information System, Faculty of Computers and Information, Fayoum University
63514 Fayoum City, Egypt
Email: amanygom3a@yahoo.com

## 1. INTRODUCTION

The growing amount of data has caused progress in data science and machine learning fields and their algorithms [1]. Its assets are to solve complicated categorization challenges, illuminate complex issues, develop core competencies, learn new things, and make important managerial decisions for organizations and people right now and within the future [2], [3]. Industry 4.0 has grown and generated enormous attention in data analytics and automation in the manufacturing technology field [4]. Information technology has a great contribution in many organizations that collect, manipulate, and analyze data in their huge databases [5]. Previously, when the business was entirely based on manual procedures, it was common to analyze and keep updates of the business status, but the main problem was that even after a lot of hard work, it was very difficult to apply analytics processes and make any useful decision regarding future business [6]. The rising digitalization of manufacturing opens opportunities for intelligent manufacturing [7]. It was recently said, “There is Gold in those mountains of data”. Due to the major development in the industrial environment and the existence of a large amount of data collected and recorded in data stores such as data warehouses and database management system (DBMS) from all interesting areas in factories or companies such as process design, product lifecycle, the materials used, marketing, scheduling, quality control, maintenance, sensors on machines, and selling processes became from the most important goals for companies and factories owners is to improve their business performance and get benefit from these huge data to achieve the best quality in their products, optimize the time process, reduce (running cost and waste time) and detect the reasons behind the machine’s breakdowns. Data mining is a powerful tool as it is an emerging and promising technology, used to discover hidden knowledge and relations automatically from huge complicated and complex which stored in databases and datawarehouse [8], [9]. Data mining is a method that explores the boost value of information that we didn't know exist in a database before [10].

Bhandare *et al*. [11] developed an Android application that enables the managers and helps them to take important decisions available on smartphones, they used data mining algorithms. to extract useful data from the central databases in the company. Massaro *et al*. [12] developed some tools in a project related to industry research on business intelligence, they used data mining tools like Weka, Rapid Miner, and KNIME workflows, and some big data techniques, they checked a good performance for all the outputs of the algorithms, they also developed a model based on big data connection, multi-attribute analysis and neural network workflow that can predict E-commerce sales with a convenient performance, Some of these inputs of this created model are the outputs of other data mining tools like social sentiment analysis. Jantan *et al*. [13] used data mining classification algorithms to detect the “talent employed” in the manufacturing environment. The results for these algorithms showed the highest accuracy of the used model is C4.5 (95.14%, 99.90%, and 90.54%). Vazan *et al*. [14] used data mining techniques and algorithms to predict the future behavior according to of manufacturing system on the production data, based on the expectation of the objective production result, they analyzed different methods, the results showed that the used predictive model markup language (PMML) files of the neural network method (NN) for numerical prediction and classification are convenient for users in the future. The objectives of this research were to design and apply Data mining techniques to facilitate the management of the industrial system control process. Chen *et al*. [15] used the extension data mining method, they applied decision tree algorithms on the products data, to promote product manufacturing quality, the results clarified that the product efficiency, rate of the company reaches the level sometime recently the generation line adjustment, which not as it incremented the company's benefit, but moreover moves forward the quality administration framework of the complete generation prepare, and realizes the low-cost optimization procedure. Reuter *et al*. [16] applied data mining algorithms such as (decision tree and K-nearest neighbour (KNN)) on real data set to estimate missing information around the user workstation, the planned model for an efficient adaptation of data mining (DM) algorithms was used to increase the consistency of data in production control, the results show that the KNN algorithm outperforms the distance transform (DT) algorithm in speed and accuracy. However, implementing the KNN algorithm is also very complex in comparison since an appropriate distance metric as well as the neighborhood’s size must be specified in advance. Khakifirooz *et al*. [17] developed a model that explore the complex semi-conductor manufacturing data for fault discovery to enable intelligent manufacturing, they developed a framework focused on Gibb’s sampling and Bayesian inference and the using of the kappa coefficient algorithm of Cohen to eliminate the effect of foreign variables. Lin *et al*. [18] used association rules, “Arules” package in R and Apriori algorithm improves the measure of preventing the failure from happening again and provide (predictive analyses) to improve product quality, the predictive enhancing the performance of the product and able to maximizes the value that has been captured by the product service system (PSS) offer, they applied these algorithms on “WBGA” product. Munirathinam and Ramadoss [19] conducted research in data analytics, the method of this analysis is modeled based on the CRISP-DM model, they used the Weka platform and R languages to introduce the proposed method and five other techniques of exploration of machine learning, and they built a decision model to help identify any faults in equipment to enhance the production process in manufacturing.

In this research paper, a case study was conducted on a real factory dataset (Emisal factory). This factory extracts 3 types of salts from “Qarun Lake” they are: Anhydrous sodium sulfate (NA2SO4), Sodium Chloride (NaCl), and Magnesium Sulfate (Mg2SO4.7H2O), our study is on Anhydrous sodium sulfate (NA2SO4) dataset. after analyzing the production data, we found that number of breakdowns hours in the year 2017 was about 63 hours and the amount of production loss for this year was about 1193 tons. So, these breakdowns affected badly on the production and the sales cycle. We have used data mining algorithms, and the knowledge discovery in databases (KDD) process to build a predictive model using the chi-square automatic interaction detection (CHAID) tree algorithm which can predict the type of breakdown by finding out the probability of relations between data and the breakdown types in the factory’s dataset.

## 2. RESEARCH METHOD

The proposed method applied and discussed in this research paper is KDD methodology. It's a full process and data mining technique is a step in this process [20]. Data mining techniques (DMT) can be used in various product management and industry fields, including production scheduling, defect analysis, quality improvement, fault diagnosis, and a lot of other applications. We use these techniques to search for valuable and efficient relations, patterns or rules and check for issues and unclear mutation processes, to improve product quality and efficiency more intelligently, exactly, and adjust the production plan timely [21], [22]. A database looks like the tank, its computer system where we can store a collection of data which we can easily access and manipulate it electronically. Big data analytics is related to huge data sets and the size is bigger than the ability of traditional database software tools to select, store, handle, evaluate, and manipulate [23], [24].

### 2.1. Knowledge discovery in databases (KDD)

KDD is the significant extraction of potentially useful information and previously unknown from data [25]. KDD process has been followed and applied, see Figure 1 is an outline of the steps of KDD process which includes: Selecting data that are relevant to the analysis and mining task, pre-processing and cleaning data from any missing and annoying data then transforming data into forms appropriate to mining step, choosing the data mining task and technique, then mining step and searching for interesting patterns in a particular representational form. Finally, interpretation and evaluation of interesting patterns in a particular representation form.

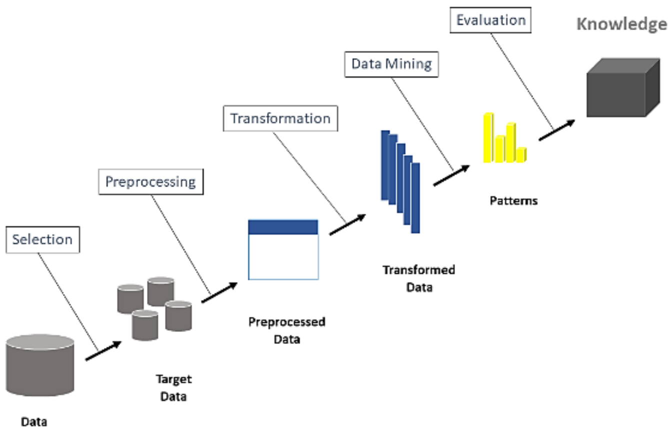



Figure 1. KDD process steps

### 2.2. Applying KDD methodology on the dataset

#### 2.2.1. Data integration

Data with different representations have been put together and the conflict within data has been resolved. Data is not in the correct or valid form, see Table 1. Is sample of the collected data from the sensors databases before applying any processing; all data have been transformed into two main files and divided into two groups: i) factory production data and ii) factory daily reading (sensor's reading).

Table 1. Sample of data on machines' sensors

| Item | N.value | 23:30 | 1:00 | 3:00 | 5:00 | 7:00 |
|---|---|---|---|---|---|---|
| FI-102 | 250 | 167 | 170 | 170 | 179 | 115 |
| FI-202 | 250 | 106 | 106 | 106 | 109 | 70 |
| TI-101 | 25 | 16.7 | 16.7 | 16.1 | 15.9 | 15.5 |
| TI-104 | 23 | ---- | ---- | ---- | ---- | ---- |
| TI-108 | 18 | 14 | 14 | 13.8 | 13.9 | 15 |
| TI-102 | 23 | 16.4 | 16.4 | 15.7 | 15.6 | 15.5 |
| TI-100 | 23 | 11.5 | 11.1 | 9 | 11.4 | 11 |
| TI-103 | ---- | ---- | ---- | ---- | ---- | ---- |
| TI-105 | ---- | 8.2 | 7.6 | 5.8 | 8.9 | 9 |

#### 2.2.2. Preprocessing step

The aim of this step is to make data clean and clear. Transform it into a useful, understandable, and efficient format to be ready for mining step. Table 2 is sample of data before preprocessing step which shows the problems in the data like missing, noisy, and inconsistent.

Table 2. Sample of data before pre-processing step

| Average Ti–102 | 23:30 | 1:00 | 3:00 | 5:00 | 7:00 | 8:30 | 10:00 | 12:00 |
|---|---|---|---|---|---|---|---|---|
| 16.0 | 11.5 | 11.1 | 9 | 11.4 | 11 | 11.7 | 10.8 | 10.1 |
| #DIV/0! | …. | …. | …. | …. | …. | …. | …. | …. |
| #DIV/0! | …. | …. | …. | …. | …. | …. | …. | …. |
| #DIV/0! | …. | …. | …. | …. | …. | …. | …. | …. |
| 16.5 | 11.7 | 10.6 | 12.3 | 11.6 | 9.9 | 4.9 | 11 | 12.5 |
| 14.8 | 9.1 | 5.4 | 11.1 | 10.2 | 6.6 | 11 | 9.5 | 9.6 |
| …. | …. | 11.7 | 8.6 | 12 | 10.6 | …. | 9 | 11.5 |
| 13.0 | 9 | 6.8 | 9.6 | …. | 7.5 | 7 | 10.8 | 9.7 |
| 12.8 | 7.4 | 10.5 | 9.7 | 8.6 | 3.8 | 8.9 | 9.5 | 8.5 |
| …. | 9.8 | 9.2 | 6.8 | 10.9 | 8.7 | 7.1 | 9 | 5.5 |

#### 2.2.3. Data cleaning

- Handling the missing
  Some data was missed, this problem handled by using the average of the column, meaning, and the most probable value.
- Handling noisy data
  Noisy data is data which is a meaningless that generated due to data entry errors faulty collection and this handled by binning method and regression function. Table 3 is representing data after data cleaning step. There is no missing, noisy or errors in the dataset.

Table 3. Sample of data after cleaning step

| Date | FI-102 N.value | 11:30:00 PM FI-102 | 1:00:00 AM FI-102 | 3:00:00 AM FI-102 |
|---|---|---|---|---|
| 1-Jan | 250 | 167 | 170 | 170 |
| 2-Jan | 250 | 167 | 170 | 170 |
| 3-Jan | 250 | 167 | 170 | 170 |
| 4-Jan | 250 | 167 | 170 | 170 |
| 5-Jan | 250 | 171 | 171 | 170 |
| 6-Jan | 250 | 170 | 162 | 170 |
| 7-Jan | 250 | 170 | 170 | 173 |

#### 2.2.4. Data transformation

This process transforms data into the most appropriate forms to be ready to apply the mining step by aggregation [22]. This step has been handled by data binning: Data has been transformed into ranges, by reducing the number of categories by binning close bins together, that applied by using excel functions:

- Filter: to determine min and max values.
- IF formula: =IF (logical test, value_if_true, value_if_false).

#### 2.2.5. Data reduction

The process of reducing the capacity of the data storage to increase storage efficiency and reduce costs. Table 4 represents samples of data after preprocessing step. Also, we have classified the breakdown into 4 types. These types explained in Table 5.

Table 4. Sample of data after transformation step

| Date | STOPPING dep. On types | Average Ti–102 | Average Ti-100 | Average Ti–105 | Average Fi–103 | Average Fi–106 | Average Ti–123 |
|---|---|---|---|---|---|---|---|
| 1-Jan | NO | 2 | 1 | 2 | 3 | 4 | 1 |
| 2-Jan | A | 2 | 1 | 2 | 3 | 4 | 1 |
| 3-Jan | A | 2 | 1 | 2 | 3 | 4 | 1 |
| 4-Jan | NO | 2 | 1 | 2 | 3 | 4 | 1 |
| 5-Jan | NO | 2 | 1 | 2 | 3 | 4 | 1 |
| 6-Jan | B | 1 | 1 | 2 | 3 | 3 | 1 |
| 7-Jan | NO | 2 | 1 | 2 | 3 | 3 | 1 |

Table 5. Explanation of the symbols which mentioned above

| Symbol | Meaning |
|---|---|
| Mode (NO) | Means there is no stopping occurred |
| Mode (A) | This is a type of stopping means (Examination of the tubes of the first evaporator exchanger) |
| Mode (B) | This is a type of stopping means (Reducing the Smelter temperature) |
| Mode (C) | This is a type of stopping means (Boiling the evaporator) |
| (M3/h) AverWithdBrine | Average of the withdrawer brine |
| BrineTemp | The temperature of the brine |
| (M3/ton) SpecConsump | Quantity of the qualitative consumption |
| (m3) BrineQuRet | The quantity of the returned brine |
| P_1205 A | The electrical capacity of the pump motor to turn on the brine for the first crystallizer |
| P_1207 A | The electrical capacity of the pump motor to turn on the mother's brine. For the first crystallizer |
| PI-109 par | The vacuum pressure inside the third evaporator |
| P-1303 amp | The electrical capacity of the pump motor to turn on the brine in the Smelter |
| TI-1311B | Oil temperature in centrifuge B to produce anhydrous sodium sulphate second stage |
| P_1207 B | The electrical capacity of the pump motor to turn on refrigerant (glycol) in the second crystallizer |
| TI-127 C | The temperature of the brine inside the smelter |
| P_1205 C | The electrical capacity of the pump motor to turn on the brine in the third crystalline |
| SI-126 C | The speed in screw pump B for pulling globar (Anhydrous sodium sulphate) |

### 2.2.6. Data mining step

After understanding the factory production process, we found that there are several days when the factory has stopped of working, we have arranged these types into four groups. Mining techniques analyze data and discover the hidden useful relations and patterns among this huge amount of data to predict the potential type of breakdown, we need appropriate techniques and algorithms to achieve our goals. Regarding the data mining tasks, methods and the actual available data, we have applied the following algorithms and techniques: i) association rules techniques (Apriori algorithm) and ii) classification prediction trees technique (CHAID algorithm).

#### a) Association rules techniques

Apriori algorithm is one of the association rule mining algorithms that use the accuracy to determine the appropriate number of indices, it's used to discover the frequent itemset, this algorithm is easy and suitable to find the association rules and relations among the given dataset items [26]. The techniques of association rules are used to extract the hidden relations of the data and discover the rules among those items [18]. So, in each transaction data with multiple items, association rules try to find the rules that govern how or why such items often appear together. There is a famous example on association rules it is (market basket analysis) that discover interesting purchasing patterns among the data transaction in the store. There are important definitions and function that used in solving the problems with association rules, they are: i) confidence: the rule $X \Rightarrow Y$ has confidence c if c% of the transactions in D that contain X contain Y too. Rules that have a c bigger than a user-specified confidence is considered to have the minimum confidence and ii) support: the rule $X \Rightarrow Y$ and support s if s% of transactions in D contain $X \cup Y$. Rules that have s bigger than the specified user support is considered to have the minimum support [27].

#### b) Decision tree analyses

A decision tree consists of nodes and leaf nodes, each decision node matches to a X test on a single attribute of the input data and has several branches, each of which handles an output of test X. Each leaf node represents a class which is the result of a decision for a case [28]. CHAID algorithm is one of the most common statistically supervised learning methods for decision tree development that was implied by a statistical Kass in the late 1970s. The CHAID algorithm is basically one of the methods of multivariate dependence and is used to detect the patterns between the categorical dependent variable and several independent variables that can be categorical [29].

The purpose of decision trees is to model a series of events understand how it affects the results. After defining the problem and preparing data, apriory algorithm and CHAID tree were used to identify and discover the hidden relations among the items in the factory dataset, by understanding the data and the process of producing the "*Anhydrous Sodium Sulfate*", we have tested a lot of items to discover the strong relations between them, we used "SPSS modeler tool" to mine the data.

### 2.3. The conducted models

We have created a predictive model by using "CHAID tree analysis" and the "Apriori algorithm" using SPSS modeler tool. The target is classifying data depending on the type of breakdown. Figure 2 is an outline of the constructed model in SPSS modeler tool.

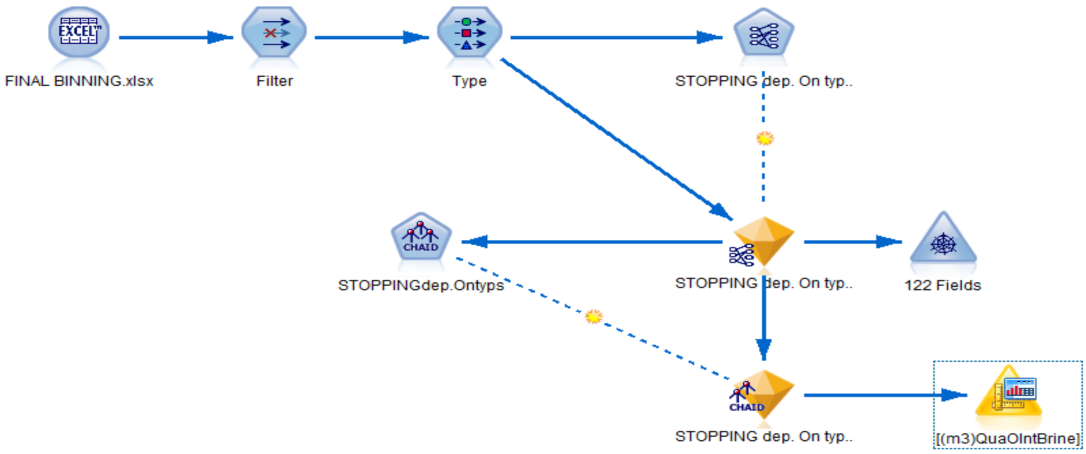



Figure 2. Apriory algorithm and CHAID tree model in SPSS

## 3. RESULTS AND DISCUSSION

### 3.1. CHAID tree analysis

As shown in Figure 3, the most significant independent variable (predictor importance) is “NA2SO4 quantity in brine” which was calculated by the used trained data, it means that the amount of NA2SO4 in the brine is most strongly associated with the dependent variable or target “stopping depending on type” and has the most strength in the distribution of observations into groups. Figures 4 and 5 illustrate the distribution of “NA2SO4 quantity in brine” instances in all ranges of the breakdown types. The results of the CHAID algorithm as shown in the tree in Figure 6 and Figure 7. Specified that the created model contains six levels of the five depth, a total of 32 nodes.

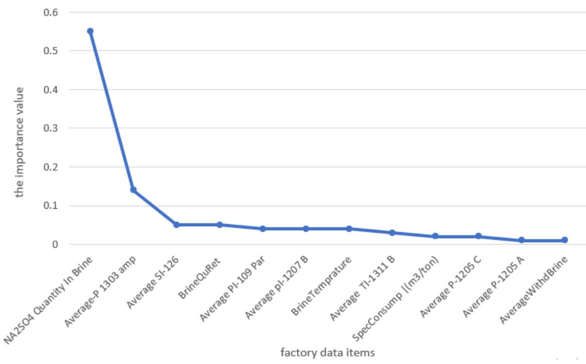
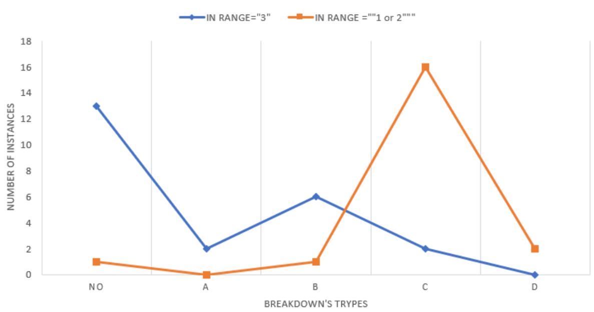



Figure 3. The predictor importance in CHAID tree



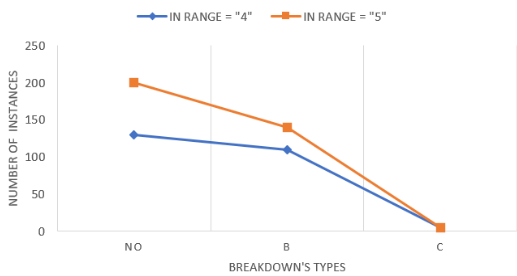




Figure 4. Number of instances when NA2SO4 in ranges 1, 2, and 3 in each type of breakdowns

Figure 5. Number of instances when NA2SO4 in ranges 4 and 5 in each type of breakdowns

Figure 6. Part 1 of CHAID tree analysis rules

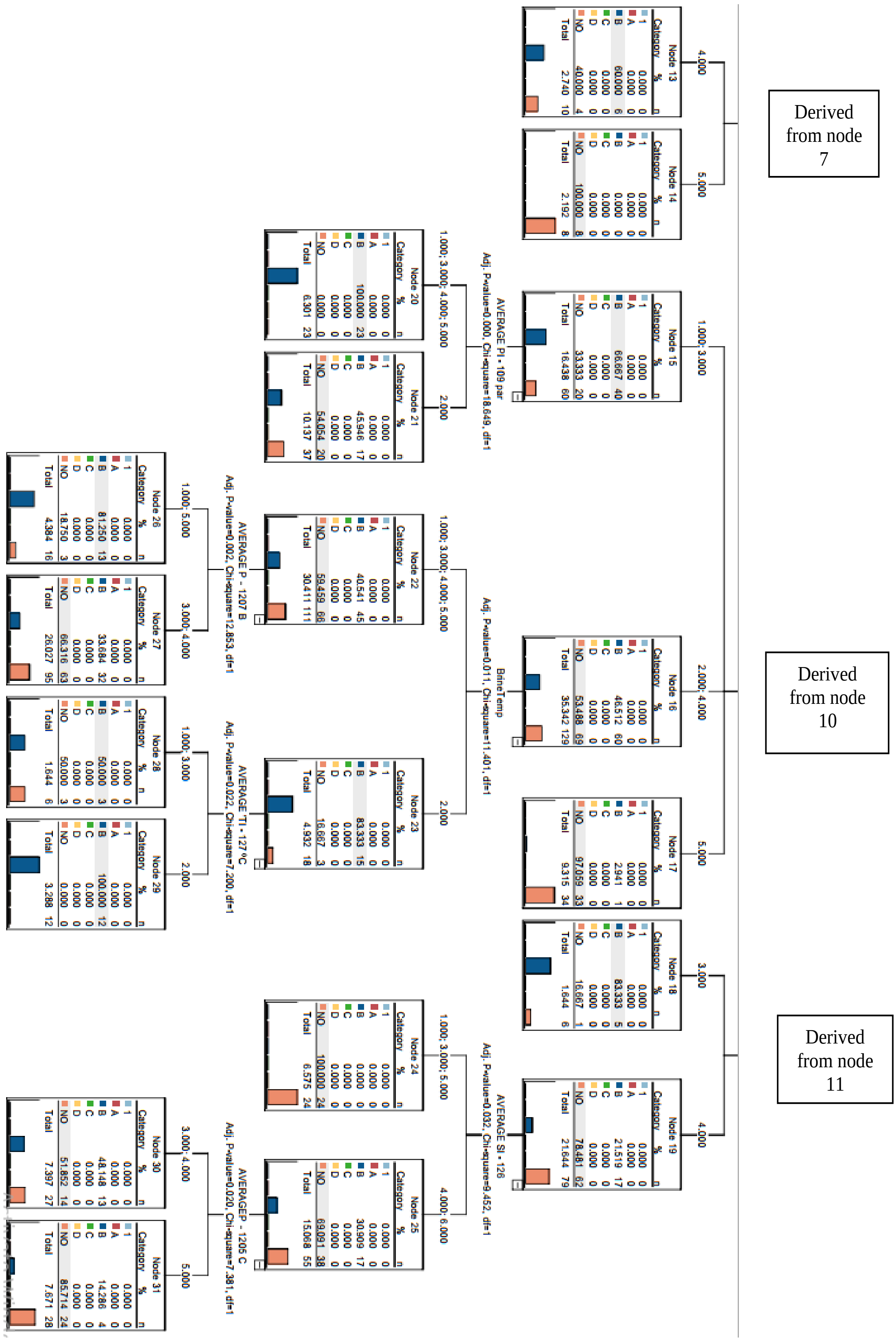



Figure 7. Part 2 of the CHAID tree analysis rules

### 3.2. Explanation of the results in the tree in Figures 6 and 7

The tree below is a predictive model which predicts the type of breakdown depending on the changes in data ranges. The strength of CHAID is that it provides details of the overall satisfaction level at each stage of the "decision tree" [30]. Generally, of the overall number of terminal nodes in the structure of the tree that has been formed:

a) When" "NA2SO4 Quantity in Brine" in range=1 or in range=2: the predicted type is [C].
Then if "NA2SO4 quantity in brine" is in range="1 or 2" and average P–1205 A=(3 or 4), the predicted type is =>[type: C].

b) ""NA2SO4 Quantity in Brine" in range= 3 the predicted type is =>NO
   - Then if "NA2SO4 quantity in brine" is in range=3 and "average withdrawal brine" in range=2, the predicted type is=>[no].
   - Then if "NA2SO4 quantity in brine" is in range =3 and "average withdrawal brine" =2 and average P–1207 A=4, the predicted type is=>[B].

The other derived rules in Figures 6 and 7 can be explained in a similar manner.

## 4. CONCLUSION

In this research paper data mining techniques like association rules algorithm (Apriori) and CHAID analysis tree have been applied on real manufactory datasets by using (SPSS modeler tool) these techniques help in discovering the hidden patterns and relations among data. We used these relations and patterns to build a predictive model which predicts the type of stopping/breakdown. by knowing the potential breakdown, the factory owners will be able to manage the risks, and this will lead to avoid the losses caused by these breakdowns which were about 1193 tons in 2017. The results also show that "NA2SO4 quantity in brine" is the most important and the most strongly related to the target which was (stopping depending on the type) and the predictive model contains six levels of the five depths, a total of 32 nodes. The results also show that data mining techniques and knowledge discovery in databases are very important for discovering the hidden knowledge among the huge amount of data in any field and all companies and factories should apply it to benefit from their recorded data to help them in their decisions now and in the future.

## ACKNOWLEDGEMENTS

We would like to acknowledge MEAPAL Data Science @ CORALIA https://www.corelia.ai for their wonderful collaboration and support and for all the opportunities was given to further research. Also, we would like to thank Emisal factory for providing us with all data we needed for our research.

## REFERENCES


[1] A. U. M. Abourayya, "Artificial neural networks implementation to predict the solution of nonlinear ODE system for the application to turbulent combustion modeling," M.S. thesis, Department of Technical Mathematics, Vienna University of Technology, Vienna, Austria, 2021.

[2] A. Guha and N. Veeranjaneyulu, "Prediction of bankruptcy using big data analytic based on fuzzy C-means algorithm," *IAES International Journal of Artificial Intelligence*, vol. 8, no. 2, pp. 168–174, Jun. 2019, doi: 10.11591/ijai.v8.i2.pp168-174.

[3] M. R. Kaseb, M. H. Khafagy, I. A. Ali, and E. S. M. Saad, "An improved technique for increasing availability in Big Data replication," *Future Generation Computer Systems*, vol. 91, pp. 493–505, Feb. 2019, doi: 10.1016/j.future.2018.08.015.

[4] T. Nkonyana, Y. Sun, B. Twala, and E. Dogo, "Performance evaluation of data mining techniques in steel manufacturing industry," *Procedia Manufacturing*, vol. 35, pp. 623–628, 2019, doi: 10.1016/j.promfg.2019.06.004.

[5] Nurhachita and E. S. Negara, "A comparison between deep learning, naïve bayes and random forest for the application of data mining on the admission of new students," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 2, pp. 324–331, Jun. 2021, doi: 10.11591/ijai.v10.i2.pp324-331.

[6] S. Singhal and P. Tanwar, "A prediction model for benefitting e-commerce through usage of regional data: a new framework," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 4, pp. 1009–1018, Dec. 2021, doi: 10.11591/IJAI.V10.I4.PP1009-1018.

[7] R. Rosen, G. von Wichert, G. Lo, and K. D. Bettenhausen, "About the importance of autonomy and digital twins for the future of manufacturing," *IFAC-PapersOnLine*, vol. 48, no. 3, pp. 567–572, 2015, doi: 10.1016/j.ifacol.2015.06.141.

[8] Y. Liu, X. Wang, X. Gong, and H. Mu, "Mechanical equipment fault detection applying data mining technology," *Journal of Physics: Conference Series*, vol. 1684, no. 1, p. 012024, Nov. 2020, doi: 10.1088/1742-6596/1684/1/012024.

[9] H. Yan, N. Yang, Y. Peng, and Y. Ren, "Data mining in the construction industry: present status, opportunities, and future trends," *Automation in Construction*, vol. 119, p. 103331, Nov. 2020, doi: 10.1016/j.autcon.2020.103331.

[10] A. K. Tiwari, G. Ramakrishna, L. K. Sharma, and S. K. Kashyap, "Academic performance prediction algorithm based on fuzzy data mining," *IAES International Journal of Artificial Intelligence*, vol. 8, no. 1, pp. 26–32, Mar. 2019, doi: 10.11591/ijai.v8.i1.pp26-32.

[11] T. Bhandare, P. Pawar, R. Patil, A. Sawant, and K. Khot, "CSEIT1722253 | Project management using data mining," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2017 IJSRCSEIT*, vol. 2, no. 2, pp. 806–808, 2017.

[12] A. Massaro, V. Vitti, A. Galiano, and A. Morelli, "Business intelligence improved by data mining algorithms and big data systems: an overview of different tools applied in industrial research," *Computer Science and Information Technology*, vol. 7, no. 1, pp. 1–21, Jan. 2019, doi: 10.13189/csit.2019.070101.

[13] H. Jantan, A. Razak, and Z. Ali, "Data mining classification techniques for human talent forecasting," in *Knowledge-Oriented Applications in Data Mining*, K. Funatsu, Ed. London: InTechOpen, 2011.

[14] P. Vazan, D. Janikova, P. Tanuska, M. Kebisek, and Z. Cervenanska, "Using data mining methods for manufacturing process control," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 6178–6183, Jul. 2017, doi: 10.1016/j.ifacol.2017.08.986.

[15] S. Chen, X. Li, R. Liu, and S. Zeng, "Extension data mining method for improving product manufacturing quality," *Procedia Computer Science*, vol. 162, pp. 146–155, 2019, doi: 10.1016/j.procs.2019.11.270.

[16] C. Reuter, F. Brambring, J. Weirich, and A. Kleines, "Improving data consistency in production control by adaptation of data mining algorithms," *Procedia CIRP*, vol. 56, pp. 545–550, 2016, doi: 10.1016/j.procir.2016.10.107.

[17] M. Khakifirooz, C. F. Chien, and Y. J. Chen, "Bayesian inference for mining semiconductor manufacturing big data for yield enhancement and smart production to empower industry 4.0," *Applied Soft Computing Journal*, vol. 68, pp. 990–999, Jul. 2018, doi: 10.1016/j.asoc.2017.11.034.

[18] H. K. Lin, C. H. Hsieh, N. C. Wei, and Y. C. Peng, "Association rules mining in R for product performance management in industry 4.0," *Procedia CIRP*, vol. 83, pp. 699–704, 2019, doi: 10.1016/j.procir.2019.04.099.

[19] S. Munirathinam and B. Ramadoss, "Predictive models for equipment fault detection in the semiconductor manufacturing process," *International Journal of Engineering and Technology*, vol. 8, no. 4, pp. 273–285, Apr. 2016, doi: 10.7763/ijet.2016.v8.898.

[20] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data mining and analytics in the process industry: the role of machine learning," *IEEE Access*, vol. 5, pp. 20590–20616, 2017, doi: 10.1109/ACCESS.2017.2756872.

[21] Y. Cheng, K. Chen, H. Sun, Y. Zhang, and F. Tao, "Data and knowledge mining with big data towards smart production," *Journal of Industrial Information Integration*, vol. 9, pp. 1–13, Mar. 2018, doi: 10.1016/j.jii.2017.08.001.

[22] Y. Yun, D. Ma, and M. Yang, "Human–computer interaction-based decision support system with applications in data mining," *Future Generation Computer Systems*, vol. 114, pp. 285–289, Jan. 2021, doi: 10.1016/j.future.2020.07.048.

[23] N. Singh, N. Garg, and V. Mittal, "Big Data – insights, motivation and challenges," *International Journal of Scientific & Engineering Research*, vol. 4, no. 12, pp. 2172–2176, 2013.

[24] K. Kambatla, G. Kollias, V. Kumar, and A. Grama, "Trends in Big Data analytics," *Journal of Parallel and Distributed Computing*, vol. 74, no. 7, pp. 2561–2573, Jul. 2014, doi: 10.1016/j.jpdc.2014.01.003.

[25] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, "Knowledge discovery in databases: an overview," *AI Magazine*, vol. 13, no. 3, 1992.

[26] N. P. More, V. B. Nikam, and B. Banerjee, "Novel approach of association rule mining for tree canopy assessment," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 3, pp. 771–779, Sep. 2021, doi: 10.11591/ijai.v10.i3.pp771-779.

[27] K. Lai and N. Cerpa, "Support vs confidence in association rule algorithms," in *Proceedings of the OPTIMA Conference, Curicó*, 2001, pp. 1-14.

[28] H. Waguih, "A data mining approach for the detection of denial of service attack," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 2, no. 2, Jun. 2013, doi: 10.11591/ij-ai.v2i2.1937.

[29] M. Milanović and M. Stamenković, "CHAID decision tree: methodological frame and application," *Economic Themes*, vol. 54, no. 4, pp. 563–586, Dec. 2016, doi: 10.1515/ethemes-2016-0029.

[30] Y. Guo, N. Wang, Z. Y. Xu, and K. Wu, "The internet of things-based decision support system for information processing in intelligent manufacturing using data mining technology," *Mechanical Systems and Signal Processing*, vol. 142, p. 106630, Aug. 2020, doi: 10.1016/j.ymssp.2020.106630.


## BIOGRAPHIES OF AUTHORS

**Amani Gomaa Shaaban** Received her bachelor's degree in "information systems" from Faculty of Computers and Information, Fayoum University, 2016. She received her master's degree in information systems department, Faculty of Computers and Information, Fayoum University, Egypt, 2022. she worked as a teaching assistant at "the higher institute of computer science and business administration" and "the higher institute for commercial science". Also, she is working as an online instructor since 2016. Research interested in data analytics, data mining, data science and information systems fields. She can be contacted at email: amanygom3ayahoo.com.

**Mohamed Helmy Khafagy** is the head of the Big Data research group at Fayoum University. Mohamed received his Ph.D. in computer science in 2009. He also works at Oracle Egypt as a consultant. Mohamed worked as project manager of many projects, and he is the manager of the National Electronic Exam center in the Supreme Council of Universities. Mohamed worked as a postdoc in the DIMA group in Technique University Berlin in 2012. Mohamed established the first big data Research group in Fayoum University in 2013. He has many publications in Big Data, Cloud computing, and database. Mohamed has been in the final round of the Social Entrepreneurship Track in the 9th MIT Arab Startup Competition. Currently, He is the Dean of Faculty of Computers and Information at Fayoum University. He can be contacted at email: mhk00@fayoum.edu.eg.

**Dr. Mohamed Abbas Elmasry** is an associate professor specializing in management information systems. He is the head of the Business Administration Program at the Egyptian National E-Learning University (EELU). He is responsible for the operational management and implementation of the academic aspects of the programme. Furthermore, he is a member of scholarly societies including the American Association for Electronic Commerce (AAE), USA, Institute of E-Commerce Consultants (Certified E-Commerce Consultant™), USA, Egyptian Association for Educational Technology (EAET) and Egyptian Curriculum Association and Teaching Methods (ECCI). He can be contacted at email: malmasry@eelu.edu.eg.

**Assoc. Prof. Dr. Heba El-Beih** is associate professor in artificial intelligence. She is a staff member at the Computer Sciences Department, Faculty of Computers and Information, Al-menoufia University, Egypt from November 2012 till now. She got her Ph.D. in 2012 from the Faculty of Engineering, Ulm University, Germany. Her research focuses on artificial intelligence, block chain, and Big Data. She can be contacted at email: heba_elbeh@yahoo.com.

**Assoc. Prof. Dr. Mohamed Hasan Ibrahim** working as data scientist with deep and diverse domain experience. Mohamed is working as Vise Dean of higher Studies and Research, Faculty of Computers and Information Systems. Mohamed has deep knowledge in data science and data analysis tools such as Power Bi, Google Data studies, and IBM SPSS Modeler. Mohamed working as Lab Manager at www.corelia.ai. He can be contacted at email: mhi11@fayoum.edu.eg and mib@corelia.ai.