**VTT Technical Research Centre of Finland**

# Open Source Analytics Solutions for Maintenance

Jantunen, Erkki; Campos, Jaime; Sharma, Pankaj; McKay, Mark

*Published in:*
Proceedings of the 5th International Conference on Control, Decision and Information Technologies

*DOI:*
10.1109/CoDIT.2018.8394819

Published: 13/04/2018

*Document Version*
Early version, also known as pre-print

# Open Source Analytics Solutions for Maintenance

E. Jantunen[1], J. Campos[2], P. Sharma[3], M. McKay[4]
[1], [4]Technical Research Centre of Finland Ltd. P.O. Box 1000, FI-02044 VTT, Finland
[2] Department of Informatics, Linnaeus University, Sweden
[3] Indian Institute of Technology, Delhi, India

***Abstract*— Abstract—The current paper reviews existent data mining and big data analytics open source solutions. In the area of industrial maintenance engineering, the algorithms, which are part of these solutions, have started to be studied and introduced into the domain. In addition, the interest in big data and analytics have increased in several areas because of the increased amount of data produced as well as a remarkable speed attained and its variation, i.e. the so-called 3 V's (Volume, Velocity, and Variety). The companies and organizations have seen the need to optimize their decision-making processes with the support of data mining and big data analytics. The development of this kind of solutions might be a long process and for some companies something that is not within their reach for many reasons. It is, therefore, important to understand the characteristics of the open source solutions. Consequently, the authors use a framework to organize their findings. Thus, the framework used is called the knowledge discovery in databases (KDD) process for extracting useful knowledge from volumes of data. The authors suggest a modified KDD framework to be able to understand if the respective data mining/big data solutions are adequate and suitable to use in the domain of industrial maintenance engineering.**

***Keywords****—data mining; big data; open source; maintenance; CBM*

## I. INTRODUCTION

This In the area of Condition Monitoring (CM) and Maintenance, there have been many efforts made by academia and industry, and various e-maintenance platform systems have been developed as a result [1; 2]. In this process, various ICTs have been applied in CM and Maintenance such as, for instance, those, which use artificial intelligence [3; 4; 5; 6]. The interest of a sub-discipline of artificial intelligence, in this case, in the data mining technologies has grown again, and this is due to the growth of big data and its 3 Vs, i.e. volume, velocity and variety of data. It is well known that data mining and big data analytics are significant enablers for the finding of hidden patterns and enhancing decision-making as well as supporting knowledge creation within an organization. It is, therefore, crucial to evaluate various open source solutions and their usability in the do-main of interest. The reason is that the development of such systems or ICT systems, in general, is costly and time-consuming, which in itself demonstrates that these are variables that might increase more than expected. For small and medium-sized companies these factors might impede the possibilities to acquire this kind of data mining as well as the technical aspects. Consequently, companies that are not able to keep up with the pace of the latest ICT developments may lose their competitive ad-vantage and the ability to compete under the same conditions as their competitors, leading to shrinking market share and profitability [7]. It is, therefore, crucial to understand the potential use of the data mining as well as big data analytics solutions systems for the domain of interest. Thus, there is a need to understand the data mining tools that are adequate to use in a big data context. In addition, some researchers say that the big data is some kind of revolution, though, to be part of this revolution it is important to comprehend how to make use of the huge amount of data to aid the knowledge discovery process as well as decision making [8]. Several frameworks have been used to highlight different aspects of data mining. For instance, Fayyad et al. [9] presented one of the first and most used frameworks, namely, the knowledge discovery in databases (KDD), which gives a general idea of the knowledge discovery activities and puts them together in a process flow. Other frameworks that have been suggested are the ones of Yao et al. [10], Geist [11], Chapman et al. [12], Kleinberg et al. [13]. In this work, the authors use the KDD as a point of reference to evaluate the data mining and big data analytics systems, since it highlights the different processes for data mining as well as the technical aspects. In addition, there are other works that have reviewed or discussed some of the open source solutions, such as Parikh and Tirkha [14]; Feng Chen, et al. [15]. However, the current paper highlights aspects that are relevant for the domain of interest, i.e. the industrial maintenance engineering and asset management. The authors provide the aspects of the domain of interest in section 2. In section 3 the framework is presented and in section 4 the findings are high-lighted while considering different aspects of the modified KDD framework. Finally, the discussion section is given.

## II. THE DOMAIN OF INTEREST

One of the main targets of condition-based maintenance (CBM) is to enable the utilization of the full potential of the monitored asset, such as a bearing, i.e. to exploit the entire asset lifespan [16]. The other important purpose of CBM is the reduction of the risk of personal injury due to sudden machine failures [17]. An effective CBM system comprises managerial support at all levels, appropriate data and analysis methods as well as an experienced maintenance staff [16]. CBM strategies include two main types, namely model-based and data-based methods [18]. Systems commonly exhibit various faults prior to actual degradation and failure. One of the most common fault

indicators is vibration. The first step in any CBM process is to acquire data regarding the asset via, for example, vibration sensors. Next, the collected data needs to be manipulated in order for information to be extracted from it. Depending on whether the CBM method is model-based, different diagnostics and prognostics methods are applied to the manipulated data in order to detect faults and predict failure respectively [18]. Once the data has been collected, it is treated differently depending on whether it is continuous, discrete or multi-dimensional. With continuous waveform data, signal processing is utilized in order to extract useful features for fault diagnostics and prognostics [18]. One of the most established feature extraction methods for continuous data is the envelope analysis. This method helps to display faults and the frequencies that might occur by helping to pinpoint the source [17]. Another common feature extraction method is the wavelet analysis. This method produces a graphical outcome that, unlike the envelope analysis, also retains time domain information [19]. However, both of these methods have their discrepancies and new approaches are constantly being developed [17]. A some-what hybrid approach, for example, is the method called expert system in which acquired data is combined with experience-based knowledge of typical, safe or otherwise important limits or trends [17]. Asset management with maintenance management being its integral part can be summarized as the balancing of cost, risk, and performance in order to achieve an optimal outcome. This is challenging due to the fact that events in the industry, as else-where, are somewhat random. More specifically, in the field of maintenance the challenge is to minimize time, cost and risk while maximizing quality, service level, and output. Regardless of the challenges related to asset management in CBM, its benefits are clear and include better use of assets, improved operations and improved communication [20].

## III. The Framework

The framework used to order the findings is briefly presented in this section. However, first, it is important to define the concept of data mining. There are many existing definitions of data mining, such as "Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories or data streams." [21]. Regarding the KDD process framework, it highlights the importance of the data mining process. Figure 1 highlights the steps part of the framework, which are the following: data selection, pre-processing, transformation, data mining, in-terpretation/evaluation and discovery or presentation of knowledge. The basic problem addressed by the KDD process is one of low-level mapping data into other forms that might be more compact, more abstract, or more useful. At the core of the pro-cess is the application of specific data-mining methods for pat-tern discovery and extraction [9]. The KDD process starts with data selection. It is an important step because the selection of relevant and correct data alone can lead to accurate analysis.

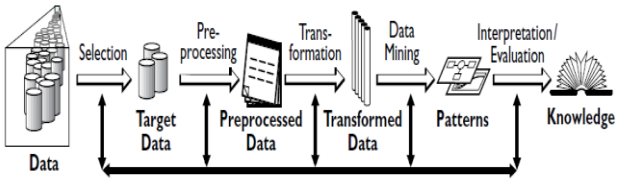


Figure 1. The knowledge discovery in databases (KDD) process, Fayyad et al. [9].

Data collection and gathering actions are often loosely controlled and result in out of range values, impossible data combinations, missing values, etc. Any analysis of such data can lead to incorrect analysis and, therefore, catastrophic decision making. Data pre-processing is a necessary step to rid the data of such inaccuracies/inadequacies. The process consists of such steps as data cleaning and data integration. Real-world databases are generally incomplete, noisy and inconsistent. Data cleaning routines are employed to cleanse the data by filling in missing values, smoothing noisy data, removing the outliers and sorting out the inconsistencies. Data cleaning activity is a mix of human and machine actions to manually/semi-automatically/automatically clean the data. A number of techniques like binning, clustering and regression are commonly used. Data pre-processing may involve data integration step. A number of data files in varying formats are combined to integrate the data. This process requires metadata - the data about the data which helps in correct integration of the schema. Data transformation is the next step involved in KDD process. This step converts the clean data into formats that are suitable for the mining operation. This may include data normalization, smoothing, aggregation and/or generalization. The next step in the KDD process is data mining, i.e. possibilities that the specific open source tools offer. In other words, what kind of data mining algorithms are available. The interpretation/evaluation aspects of the tools imply what kind of statistical options are available to evaluate the results and hidden patterns provided by the algorithm/s. It involves extraction of previously unknown patterns and knowledge from large amounts of data. The key properties of data mining are automatic discovery of patterns, prediction of likely outcomes, the creation of actionable information and focus on large datasets and databases. Prediction of likely out-comes and description of data regarding human-interpretable patters are the two key goals of data mining. These goals are achieved through the application of one or more of the following methods: anomaly/change/deviation detection, i.e. the identification of unusual data records that might be interesting or data errors that require further investigation. It also involves discovering the most significant changes in the data from previously measured or normative values [22]. Dependency modelling searches for relationships between variables. Dependency models describe the association between variables at two levels; structural and quantitative. Structural dependency describes which variables are dependent on each other. Quantitative dependency is described as the strength of the association on a numeric scale. Clustering is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data

[23]. The categories can be mutually exclusive and exhaustive or consist of a richer representation, such as hierarchical or overlap-ping categories [9]. Classification is the task of categorizing the data into one of the previously defined classes. Regression is aimed at deriving a function that maps a data item to a real-valued prediction variable. Summarization is used to provide a more compact representation of the data set, including visualization and report generation. Finally, the knowledge presentation, i.e. visualization and knowledge representation techniques. In addition to the aspects highlighted by the KDD framework, it is important to understand if the specific open source tool is adequate to use for big data analytics purposes. The authors, therefore, add to the KDD framework other aspects in order to understand if the software is suitable for purposes of big data analytics. These aspects include, for instance, whether it supports the 3V's (Volume, Variety, and Velocity aspects), whether it is possible to use the soft-ware in a big data analytics ecosystem, such as the Hadoop or Spark, and what kind of visualization aspects it has, etc. Open Source Solutions

## IV. The Open Source Solutions

The different open source solutions in this section are highlighted against the modified KDD framework presented in section 3. The open source tools reviewed are Weka, Rapidminer, Orange, NLTK, KNIME, KEEL and Tanagra. The Weka workbench is a data mining tool developed by Waikato University in New Zeeland by professor Witten and his team (cs.waikato.ac.nz/ml/weka). It is issued based on the GNU General Public License. The Weka tool has possibilities to connect to databases with JDBC and URL connection and to work with ARFF, and it has possibilities to convert files to ARFF. It contains converters for several file formats, such as spreadsheet files with extension .csv, C4.5's native file format with extensions .names and .data, and XML-based ARFF format files with extension .xrff, as well as ASCII Matlab files with ex-tension .m etc. The Weka workbench provides several alternatives for pre-processing and transformation of data, such as filters that are crucial in many cases since, for instance, some techniques such as association rule mining are only possible to perform with categorical data. Thus, it entails performing dis-cretization on numeric or continuous attributes. Other filters that Weka provides are for supervised and unsupervised attributes as well as for instances of a dataset. For example, a filter that assumes instances form time-series data and replaces attribute values in the current instance with the difference between the current value and the equivalent attribute value of some previous (or future) instance. This is convenient when the time-shifted value is unknown or the instance can be dropped, or missing values are used. As a result, Weka has a rich set of filters for data pre-processing. Data mining and interpretation/evaluation aspects of Weka are comprehensive.

The ***Weka*** workbench has several classification, clustering and association algorithms embedded in it. It provides possibilities to use training dataset, sup-plied dataset, cross-validation and percentage split of the data. In addition, it has alternatives even to visualize the results of different data mining algorithms used. The workbench even provides an experiment environment, where several data mining algorithms can be tested on the same dataset at the same time which is a crucial and beneficial option that the analytics tools offer. Furthermore, it is possible to use Weka for big data purposes because the main problem lies in training models from large datasets and not prediction aspects for large datasets (cs.waikato.ac.nz/ml/weka). In addition, it is also specified that Weka is able to make predictions in real time in demanding real-world applications. Using the Weka explorer graphical user inter-face to train models from large datasets is, however, not recommended, since, for instance, the explorer loads the entire dataset into the main memory in a computer. It is, instead, recommended to employ command-line interface (CLI) for the interaction with Weka. It is also possible to use a library that gives access to MOA data stream application having advanced algorithms developed to be used for large datasets or data streams [24], (https://moa.cms.waikato.ac.nz). The latest version of Weka has a package for distributed data mining, which provides a base "map" and "reduce" task that is not dependent on any specific platform. In addition, the last version provides both Hadoop and Spark specific wrappers. Furthermore, it is even possible to use Weka together with Spark (markahall.blogspot.se). However, internally, distributed Weka Spark handles only CSV files now, but in the future it will support other data formats, such as Parquet, Avro and sequence files. Finally, the tool is developed in Java and it is possible to work with Weka embedded into Net-Beans or Eclipse which provides possibilities to change and optimize the code. ***RapidMiner*** is a data science software platform that is developed by the company with the same name. It provides possibilities to perform tasks, such as data preparation, machine learning, deep learning and predictive analytics (rapidminer.com). It provides the use of several data formats, such as .cvs, excel URL, SPSS, C4.5, ARFF, XML, etc. and the ability to connect to data-bases and make SQL queries, i.e. it provides support for 40 file formats and commercial databases. However, the open source version of the RapidMiner is limited to 10,000 rows of data. Data pre-processing and transformation is made possible by an opera-tor called Rename that can be used to rename one or more at-tributes or to rename multiple attributes of a dataset. It is a comprehensive part of Rapidminer. Other parts of the pre-processing and transformation are selection, value modification, data cleansing, filtering, sorting, rotation and aggregation etc. The Modelling part, i.e. data mining and interpretation/evaluation aspects of Rapidminer, provides several kinds of supervised and unsupervised algorithms, such as classification, regression, attribute weighting, clustering, association, correlation and dependency computation, as well as computation and model application. In addition, as far as evaluation aspects are concerned, it contains validation, performance measurements, significance and visual evaluation, i.e. visualization. Furthermore, it is also possible to use all the algorithms that are part of other tools with a so-called extension, i.e. Weka, Series extension (containing Wavelet and Fourier Transformation, as well as methods for extracting features from value series). RapidMiner provides possibilities to use Hadoop by allowing users to do an advanced distributed analysis of data

on Hadoop and its ecosystem (rapidminer.com). The version that is recommended to use that has no data constraints among other and that is completely open source is the alternative version 6.5. Rapidminer also provides the possibility to choose a paid version of its software, which provides the users and enterprise with the possibility to work with more data and achieve faster performance. ***Orange*** is the open-source software toolkit that provides possibilities to use machine learning, data mining and visualization (orange.biolab.si). In addition, it is based on the GNU General Public License. As far as data formats are concerned, it can read files in native tab-delimited format; it can also load data from standard spreadsheet file types, such as CSV and Excel as well as create tables from an ODBC connection. There are several options for pre-processing and transformation, such as impute discretization, normalization, randomization, etc. Data mining and interpretation/evaluation aspects of Orange are wide-ranging as far as data mining is concerned, it contains several classifications, regression and clustering algorithms. The user interface, as well as its graphics, is advanced and well developed. It is easy to understand and change parameters for further details when needed, etc. In addition, it has alternative options to connect to Spark data, as well as to utilize Python's packages, such as Pan-das, which contain data mining and machine learning algorithms. To summarise, Orange is a software tool developed in Python language, which is a convenient tool, since it has a big community of developers that are constantly working on the improvement of packages and other features of the programming language and its environment. (docs.orange.biolab.si/3/data-mining-library/tutorial/).

The increasing pervasiveness of mobile devices has resulted in a renewed need to process natural language to allow computers to manipulate them for performing the intended tasks. Big data analytics of unstructured data, which is available in abundance, requires quick processing. Sentiment mining of tweets and blogs, videos and text analytics, etc., are all based on natural language processing (NLP). Language processing has become an important facet of modern-day analytics in a multilingual information society. ***Natural Language Toolkit (NLTK)*** is a free, open source and community-driven platform for building Python pro-grams to work with human language data. It is a suite of libraries and programs for symbolic and statistical natural language processing (NLP). NLTK includes extensive software, data, and documentation, all freely downloadable from http://nltk.org/. It is a simple, intuitive framework along with substantial building blocks. It has an extendable structure into which new modules can be easily added. The modularity of the toolkit allows the users to use any module without having to understand the rest of the toolkit. There are some drawbacks of NLTK. It is merely a toolkit. not a system. It cannot be considered complete in any sense, as it will continue to be improved and evolve alongside with natural language processing. The toolkit is also not optimised for high runtime performance. ***Konstanz Information Miner (KNIME)*** is an open source data analytics, reporting and integration platform. Its development began in 2004 at the University of Konstanz in a team led by Michael Berthold. It is written in Java and is based on Eclipse. This modular environment enables easy integration of new algorithms, data manipulation and visualisation methods as models [25]. The extendibility of the platform allows the use of plugins to provide additional functionality. Pre-written modules are available as part of the platform that can manage numerous data needs like integration, transformation, data mining and text analytics. KNIME is more user-friendly as it allows to visually create data pipelines, execute analysis steps, and view the results later. It has a graphical user interface that allows assembly of nodes for data pre-processing (Extraction, Transformation, Loading), for modelling and data analysis and visualization without, or with only minimal, programming. It can process large volumes of data and is ideal for industrial asset maintenance scenarios. It has the ability to add plugins that allow the integration of methods for text and Image mining, as well as time series analysis. It can be integrated with other open-source platforms, such as Weka or R, etc. It supports data I/O in ARFF, as well as in other data formats, such as XML. It also allows other languages like Python, Perl and Java etc. to be run. It has over 1000 modules that allow a host of data analytics. It is possible to con-duct mathematical and statistical analysis on the data and the platform has advanced predictive and machine learning algorithms. KNIME provides big data extensions for integration with the Apache Hadoop and Spark as well as with the KNIME Analytics Platform (www.knime.com). In addition, KNIME provides a server extension that provides power and flexibility to its plat-form. The KNIME server is a paid version and is available in three versions with different characteristics when it comes to possibilities of collaboration, deployment, and management. ***Knowledge Extraction based on Evolutionary Learning (KEEL)*** is a software tool that facilitates the analysis of the behaviour of evolutionary learning in the different areas of learning and pre-processing tasks, making the management of these techniques easy for the user [26]. It is an open source Java software tool that is used for knowledge data discovery tasks. The data management module of the software facilitates the creation of new datasets, converting imported data to KEEL format, and vice-versa. The software also provides easy to use features for editing and partitioning the data, and preparing data for analysis by conducting cleaning, transformation and reduction. It contains a Knowledge Extraction Algorithms Library (http://www.keel.es/algorithms.php) that uses the techniques like Evolutionary rule learning models, Fuzzy systems, Evolutionary neural networks, Genetic programming, Subgroup discovery and Data reduction (instance and feature selection and discretisation). It provides a simple Graphical User Interface (GUI) based on a data flow to design experiments with different datasets and computational intelligence algorithms (paying special attention to evolutionary algorithms) to assess the behaviour of the algorithms. ***Tanagra*** is a free suite of machine learning software. Ricco Rakotomalala at the Lumière University, France developed it in 2003. It covers several ML schemes, data preparation, and ex-perimental analysis. It is a user-friendly data mining software which allows users to design a GUI and carry out analysis of real or synthetic data. Tanagra supports several standard data

mining tasks, such as Visualization, Descriptive statistics, Instance selection, feature selection, feature construction, regression, factor analysis, clustering, classification and association rule learning [27]. It supports statistical approaches (e.g. parametric and non-parametric statistical tests), multivariate analysis methods (e.g., factor analysis, correspondence analysis, cluster analysis, regression) and machine learning techniques (e.g., neural network, decision trees). Tanagra allows the user to design the data mining process in the form of a diagram. Machine learning techniques are represented as nodes in the diagram and the connections between the nodes show the transfer of data that takes place. The output of the data mining process is in HTML format allowing easy export and viewing. It is also possible to copy the result tables to a spreadsheet.

There are additionally both the R and Python programming languages that are extensively used for data mining and data science. R is a free software mainly used for statistical computing and graphics. It is, however, also possible to use it for data mining and data science purposes (www.r-project.org). It contains several packages, such as rminer, DMwR, in addition to several visualisation packages. R was developed mainly by C, C++, Fortran and contains an interface developed with the R script language. Python is a general-purpose programming language that has become very popular for conducting data mining and data science tasks (www.python.org/). It has a number of packages/libraries to use for data mining and big data analytics, such as Pandas, which is a library for data manipulation and analysis. The machine-learning packages are SciKit-Learn, Keras, Theano, and Tensor-Flow. For visualisation, Python contains Matplotlib, Seaborn, Bokeh, and Plotly that provide possibilities too, for instance, draw line and scatter plots, pie charts etc.

## V. Discussion

Data analytics is slowly becoming the backbone of successful businesses. It enables companies to learn from their past mistakes in order to perform better in the future. Small companies, which have production machinery, often shy away from analysing big data because of the large investments that these solutions require. Open source data mining software solutions present a viable and economical alternative for data analysis to such small businesses. Furthermore, none of the reviewed open source tools has been implemented and used in a big data ecosystem for purposes of big data analytics in the domain of interest. The current work provides an insight into the open source solutions that could be suitable for these purposes. The factors of the 3Vs, which are more relevant for the area of industrial maintenance, are “velocity” and “volume” of the data. However, if we talk about asset management, then the “variety” aspects of the data could also come into the picture. The use of, for instance, unsupervised and supervised approaches, that is, clustering and classification, are important to comprehend, i.e. to be able to under-stand if the open source tools are adequate to use in the domain of interest. For instance, the wavelet characteristic, such as its multi-resolution property, permits the use of cluster algorithms, especially the WaveCluster algorithm, since it could efficiently detect arbitrary shape clusters on various scales with different degrees of correctness [28]. The WaveCluster is part of one of the R packages. Consequently, it is part of several open source solutions reviewed, such as Rapidminer and KNIME, since it is possible to add R package extensions to these tools. In the case of WEKA, it is possible to add R packages too. In addition, WEKA also contains a filter for wavelet transformation. The Wavelets may be highly suitable for classification as well. For instance, classification algorithms can be applied to the wavelet domain data. The authors discuss several other approaches, such as linear regression and neural network [28]. All these methods and approaches can be found in the open source tools. In addition, for all flow of data, such as sensor data, time data, signal analysis, diagnosis, it is possible to further investigate the different output with data mining/big data analytics like clustering and/or classification to find possible hidden patterns. Other works that combine maintenance approaches and data mining can be found in [29; 30; 31; 32]. However, if the aim is to use the open source tools for big data analytics, then there is a need to under-stand whether they are able to perform job tasks that are part of the big data approach, i.e. that consider the aspects of the 3 Vs. Rapidminer and KNIME have paid versions that fulfil these requirements. Other tools, such as WEKA and Orange, which are free to use, are also available to be implemented in the domain of big data analytics. R and Python are also convenient for analytics. Python being a general-purpose programming language is additionally suitable for developing algorithms and/or the software part of a data mining or big data analytics system. WEKA, Orange and Tanagra’s main drawback is the data file formats to be used in their workbench, i.e. it enforces a limitation of alternatives. Thus, many of the tools reviewed do not support, for in-stance, all the data file format sources. However, by being open source, they provide the possibility to update the code to change those constraints. In conclusion, it is possible to use several of the open source solutions for purposes of data mining and big data analytics. The main drawback is the different alternatives available when it comes to the data formats these tools can access in their environment. It is, therefore, crucial to understand the data aspects and how they will be dealt with. Furthermore, the speed of data is also an important aspect to consider/take into account, as well as the volume, in order to be able to use them optimally. It is, therefore, of great significance to investigate the matter further and implement and test these tools in the domain of interest in order to understand their suitability.


## ACKNOWLEDGMENTS

The research has been conducted as a part of MANTIS Cyber Physical System based Proactive Collaborative Maintenance project. The project has received funding from the Electronic Component Systems for European Leadership Joint Undertaking under grant agreement No 662189. This Joint Undertaking receives support from the European Union’s Horizon 2020 research and the national funding organisation Finnish Funding Agency for Innovation Tekes.

## REFERENCES


[1] A. Muller, A.C., Marquez, and B, Iung, "On the concept of e-maintenance: Review and current research," Reliability Engineering & System Safety, vol 93, no8, pp. 1165-1187, 2008.

[2] J. Campos, "Current and prospective information and communication technologies for the e-maintenance applications." J of Qual in Maintenance Eng 20, pp. 233–248, 2014. https://doi.org/10.1108/JQME-05-2014-0029

[3] T, Marwala, "Condition monitoring using computational intelligence methods. 2012, Springer-Verlag. doi:10.1007/978-1-4471-2380-4

[4] P.M, Patel, and J.M, Prajapati, "A review on artificial intelligent system for bearing condition monitoring," International Journal of Engineering Science and Technology, Vol. 3, pp. 1520–1525, 2011.

[5] J, Campos, "Development in the application of ICT in condition monitoring and maintenance," Computers in Industry, vol. 60, issue 1, pp. 1–20, 2009. https://doi.org/10.1016/j.compind.2008.09.007

[6] J, Campos and O, Prakash, "Information and communication technologies in condition monitoring and maintenance," IFAC Proceedings Volumes, vol. 39, no. 3, Jan. 2006, pp. 3–8. Science Direct, doi:10.3182/20060517-3-FR-2903.00003.

[7] Turban, E., Sharda, R., & Denlen, D. (2011). Decision support and business intelligence systems (9th ed.). Upper Saddle River, NJ: Pearson Prentice Hall. 2011.

[8] Mayer-Schönberger, V., & Cukier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work, and Think. American Journal of Epidemiology, 179(9), 1143-144.

[9] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27-34.

[10] Yao, Y. Y., Zhong, N., & Zhao, Y. (2004). A three-layered conceptual framework of data mining. In Workshop Proceedings Foundations of Data Mining ICDM (pp. 205-212).

[11] Geist, I. (2002, March). A framework for data mining and KDD. In Proceedings of the 2002 ACM symposium on Applied computing (pp. 508-513). ACM.

[12] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

[13] Kleinberg, J., Papadimitriou, C., & Raghavan, P. (1998). A microeconomic view of data mining. Data mining and knowledge discovery, 2(4), 311-324.

[14] D, Parikh, and P, Tirkha, "Data Mining & Data Stream Mining - Open Source Tools, " International Journal of Innovative Research in Science, Engineering and Technology vol. 2, pp. 5234–5239, 2013.

[15] F, Chen, P, Deng, J, Wan, D, Zhang, A. V, Vasilakos, and X, Rong, "Data Mining for the Internet of Things: Literature Review and Challenges," International Journal of Distributed Sensor Networks, vol. 11, no. 8, p. 431047, Aug. 2015.

[16] M, Bengtsson, Condition Based Maintenance Systems – an Investigation of Technical Constituents and Organizational Aspects. Dissertation. Mälardalen University, Department of Innovation, Design, and Product Development. Mälardalen, ISSN number: 1651-9256, 2004.

[17] S, Orhan, N, Aktu¨rk, and V, Celik, "Vibration monitoring for defect diagnosis of rolling element bearings as a predictive maintenance tool: Comprehensive case studies,"Journal of Independent Nondestructive Testing and Evaluation. vol. 39, pp. 293-298, 2006, DOI: 10.1016/j.ndteint.2005.08.008

[18] A. K. S , Jardine, and D, Lin, and D, Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance" Journal of Mechanical Systems and Signal Processing. vol. 20, pp.1483–1510, 2006, DOI:10.1016/j.ymssp.2005.09.012

[19] P. W, Tse, and Y. H, Peng, and R, Yam, "Wavelet Analysis and Enevlope Detection for Rolling Element Bearing Fault Diagnostics - Their Effectiveness and Flexibilities. Journal of Vibration and Acoustics. vol. 123, pp. 303-310, 2001, DOI:10.1115/1.1379745

[20] J, Cambell, and A, Jardine, and J, McGlynn, "Asset Management Excellence". FL, USA: Taylor and Francis Group, 2016, ISBN-13:978-0-8493-0324-1 E-book

[21] J, Han, J, Pei, and M, Kamber, "Data mining: concepts and techniques," Elsevier, 2011.

[22] D.J, Berndt, J, Clifford, "Finding patterns in time series: a dynamic programming approach," Advances in Knowledge Discovery and Data Mining, pp. 229–248, 1996.

[23] A, Jain, and R, Dubes, "Algorithms for Clustering Data," Prentice-Hall, Englewood Cliffs, NJ. 1988.

[24] A, Bifet, G, Holmes, R, Kirkby, and B, Pfahringer, "MOA: Massive Online Analysis," Journal of Machine Learning Research vol. 11, pp. 1601-1604, 2010.

[25] M.R , Berthold, N, Cebron, F, Dill, G, Di Fatta, T.R, Gabriel, F, Georg, T, Meinl, and P, Ohl, "KNIME: The Konstanz Information Miner", In: Proceedings of the 4th annual industrial simulation conference, Workshop on multi-agent systems and simulations, Palermo. 2006.

[26] J, Alcala-Fdez, L, Sanchez, S, Garcia, M.J, del Jesus, S, Ventura, J.M Garrell, J, Otero, C, Romero, J, Bacardit, V.M Rivas, J.C Fernández, and F, Herrera, "KEEL: a software tool to assess evolutionary algorithms for data mining problems", Soft Computing, vol. 13, issue 3, pp. 307-318, 2009.

[27] R, Rakotomalala, "TANAGRA: un logiciel gratuit pour l'enseignement et la recherché", In: Proceedings of the 5th Journées d'Extraction et Gestion des Connaissances, pp. 697–702, 2005.

[28] T, Li, S, Ma, and M, Ogihara, "Wavelet Methods in Data Mining," in: Maimon, O., Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook. Springer US, Boston, MA, pp. 603–626. 2005. https://doi.org/10.1007/0-387-25465-X_27

[29] Pena, M., Alvarez, X., Jadán, D., Lucero, P., Barragán, M., Guamán, R., Sánchez, V. and Cerrada, M., ANOVA and cluster distance based contributions for feature empirical analysis to fault diagnosis in rotating machinery. SDPC 2017, At China, vol. 1, 2017.

[30] F. Pacheco, J. Valente-de-Oliveira, R.-V. S´anchez, M. Cerrada, D. Cabrera, C. Li, G. Zurita, and M. Art´es, "A statistical comparison of neuroclassifiers and feature selection methods for gearbox fault diagnosis under realistic conditions," Neurocomputing, vol. 194, pp. 192 – 206, 2016.

[31] J. Zhang, W. Ma, L. Ma, and Z. He, "Fault diagnosis model based on fuzzy support vector machine combined with weighted fuzzy clustering," Transactions of Tianjin University, vol. 19, no. 3, pp. 174–181, 2013.

[32] M. Cerrada, G. Zurita, D. Cabrera, R.-V. S´anchez, M. Art´es, and C. Li, "Fault diagnosis in spur gears based on genetic algorithm and random forest," Mechanical Systems and Signal Processing, vol. 70-71, pp. 87–103, 2016.