Volume 6, Issue 6, June 2017, ISSN: 2278 – 1323

A Novel Integrated Random Forest and Gradient Boosting Machine Technique for Software Reuse Analytic

Pawandeep Kaur[1], Dr. Pankaj Deep Kaur[2]

[1]Research Scholar, Department of Computer Engineering & Technology, Guru Nanak Dev University Regional Campus ,Jalandhar, Punjab, India-143001

[2]Assistant Professor, Department of Computer Engineering & Technology, Guru Nanak Dev University Regional Campus ,Jalandhar, Punjab, India-143001

Abstract— **. Cleaner production (CP) nowadays is considered as the important mean for production companies to get the sustainable production. However, there are various parameters of the CP like recycling, production cost, reuse, energy consumption and minimization of waste material, which help for the successful implementation of CP process for manufacturing and maintenance processes (MMP). Big data based analytics for product lifecycle (BDA-PL) architecture is one of the various architectures which helps to get the better CP and product lifecycle management(PLM) decisions based on the large amount of heterogeneous big data. Software reuse is process of producing new products or software from the present software by making some changes. This paper presents an integrated Random forest and Gradient Boosting machine (GBM) technique for software reuse analytics which improves the matrices like accuracy, error rate, Mean Absolute Error (MAE) and Relative Absolute Error (RAE).The result shows that there is 20% increase in performance of proposed algorithm with respect to existing algorithm.**

keywordss— **Cleaner production, Big data analytics, BDA-PL architecture, PLM, Software reuse, Random forest, Gradient Boosting Machine, Accuracy, Error rate,**

I. INTRODUCTION

Big data analytics (BDA) is described as the process of collecting, organizing and analyzing of huge sets of data to find the patterns and other meaningful information. The most important characteristics of big data have volume, velocity and varieties including two additional characteristics are veracity and value.

Nowadays, manufacturing enterprises prefers to manufacture environmental-friendly products to avoid the pollution threats. So, to achieve the sustainable production, Cleaner Production (CP) is used for manufacturing products in the manufacturing industries. Cleaner production can provide various benefits such as economic, environmental and the social benefits. There are lot of issues related to the implementation of the CP program which includes the lack of information about the clean technologies, insufficient information, less supply of equipment, poor communication systems, available procedures, lack of availability and accessibility for the useful information relevant to product and lack of skills. CP and product lifecycle management (PLM) both strategies are used to improve the sustainable competitive advantage of the enterprises. The main three things that allow the CP strategy to be implemented successfully are capture lifecycle data, discover knowledge from raw data and share knowledge among all lifecycle stages. Manufacturing and maintenance process (MMP) mainly includes Research and Development and Manufacture (RDM) and Operation and Maintenance (OM), respectively. The various parameters that are related for the successful implementation of CP process in MMP are as:

(i)Accuracy: Itis the rate of correct predictions made by the classifier over a given data set. The value of accuracy should be high. Accuracy metric can be calculated as:

$$Accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN}\right)$$

Where, TP= True positives, TN= True negatives, FP= False positives and FN= False negatives.

(ii) Error rate: Itis the rate of incorrect predictions made by the classifier over a given data set. It needs to be minimized. Mathematically,

$$Error\ rate = 1 - \left(\frac{TP + TN}{TP + TN + FP + FN}\right)$$

(iii) Root Mean Squared Error (RMSE): It measures the mean magnitude of the errors by considering their directions and useful for avoiding the large errors. RMSE can be calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

(iv) Mean Absolute Error (MAE): It measures the mean magnitude of the errors by without considering their directions and it can be calculated as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Predicted_i - Actual_i|$$

(v) Relative Absolute Error (RAE) : it uses the total absolute error and normalizes it by dividing by the total absolute error of the simple predictor.Mathematically,

$$RAE = \frac{\sum_{i=1}^{N}|predicted_i - Actual_i|}{\sum_{i=1}^{N}|\theta_i - Actual_i|}$$

Where, ɵ = mean value of the actual values.

(vi) Root Relative Square Error (RRSE) : It uses the total squared error and normalizes it by dividing by the total squared error of the simple predictor. It needs to be minimized and can be calculated as:

$$RRSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{\sum_{i=1}^{N}(\theta_i - Actual_i)^2}}$$

Where, ɵ = mean value of the actual values.

(vii) Kappa statistics: it simply compares the accuracy of the system to the accuracy of random system and it can be calculates as:

$$Kappa = \frac{Total\ accuracy - Random\ accuracy}{1 - Random\ accuracy}$$

Rest of the organization of paper is as follow: Section II discusses BDA-PL architecture and PLM. Section III defines the concept of software reuse analytics using hybrid technique Random forest and GBM. Section IV gives a brief summary of related works. Gaps in literature survey are presented in Section V. Section VI and VII evaluate the methodology and results obtained by experiments. Section VIII concludes the paper.

II. BACKGROUND

This section gives brief introduction about BDA-PL architecture and PLM for the maintenance and manufacturing process of complex products. These both approaches help to get the better CP decisions based on large amount of big data.

A. BDA-PL architecture

An overall architecture of big data – based analytics for product lifecycle (BDA-PL) is proposed [23], for making the better CP and PLM decisions based on the large amount of the multi-source and real-time lifecycle big data**.** This architecture benefited manufactures, environment, customers and all the stages of PLM, and effectively supported the implementation of CP. An overall architecture of the BDA-PL is proposed in Fig. 1.

1. Application services of PLM
2. Acquisition and integration of big data
3. Big data processing and storage
4. Big data mining and knowledge discovery in databases (KDD)

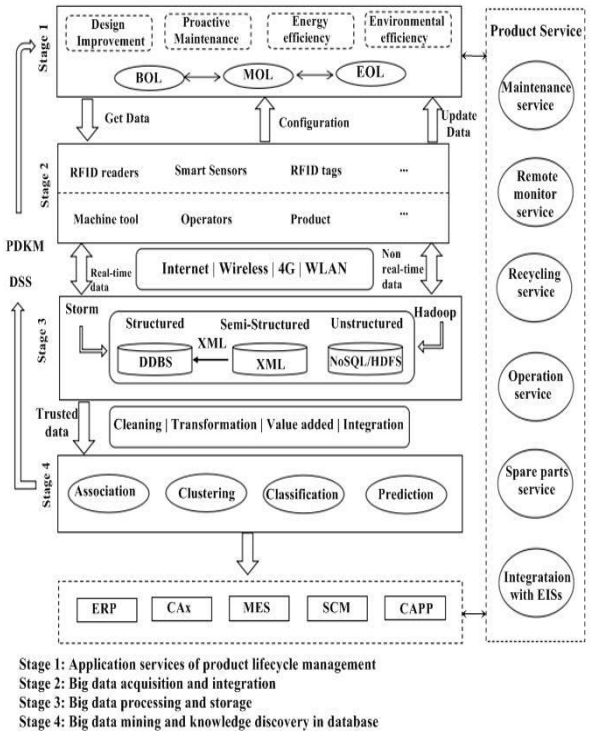


Fig.1. Overall architecture of big data-based analytics for product lifecycle.

B. Product lifecycle management (PLM)

Product lifecycle management (PLM) introduced for handling the useful information rigorous process consisting of item design, product manufacturing, and product in use and product recycling. These all phases can be divided into three stages: beginning of life (BOL), middle of life (MOL), and end of life (EOL).

In **BOL** phase, the most two important steps are: marketing evaluation and the merchandise design. The most essential task in marketing evaluation is meeting the demands of the customers. In the phase of merchandise design, the data used may be tracked from the explanation of needs to the particular product information and ultimately to the comprehensive design specifications. In **MOL** phase of PLM, as merchandises and services have endured in the ultimate form, problems related the service have become the significant and must be paid high concentration. In **EOL** phase volume conclusion must be taken which matter the EOL item disposable and recycle.

III. SOFTWARE REUSE ANALYTICS USING INTEGRATED RANDOM FOREST AND NEURAL NETWORK

Software reuse is a process of developing the software from the existing software by performing some changes in the existing ones. It helps to increase the productivity and also saves time and production cost. Random forest combines the decisions of individual trees and helps to improve the accuracy. It uses the bagging for randomness. Mathematically, as

$$bagging = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

x’ = predictions for unseen samples.

b = number of tress i.e. b= 1,2,3……. B.

f_b = Train a decision tree f_b on X_b, Y_b.

The major drawback of random forest is its speed. So, to avoid the issue of speed GBM is combined with it to get the better results for software reuse analytics. GBM is a technique used for regression and classification techniques for the better prediction results and also helps in reducing errors by decreasing bias.

IV. RELATED WORK

This section summarizes work related to big data analytics, product lifecycle and cleaner production (CP) for the manufacturing and maintenance process (MMP) of products and some parameters essential for the successful implementation of CP as shown in Table 1. Galletti et al. (2013) [11] focused on Big Data Analytics (BDA) and introduced how BDA may be observed and act as a driver for the industries’ competitive advantage. Dai et al. (2011) [10] described a approach in which big data applied in the cloud, for building a simple, lightweight and the extremely scalable performance analyzer used for the dataflow-based evaluation. Rabl et al. (2012) [23] discussed about the analysis of challenges of big data for company application efficiency management and based on this experience and lessons learned from the investigation, big data applications in enterprise could be promoted. . To reduce the use of water, handle pollution at low cost, proper use of resources & reduce waste material, Kupusovic et al. (2005) [13] proposed a project for slaughterhouse industry. Wang et al. (2010) [26] presented a warehouse design management system for the tobacco industry, which is based on the RFID technology. By using the RFID technology, the system produced a digital warehouse to get the maximum capacity, automatic storage, better visualization and high accuracy of inventory control. Vinodh et al. (2011) [25] reported the utilization of the fuzzy association rules mining method that allowed the developers to take the efficient decisions by using the various attributes like quality, cost, pro-activity, robustness, innovativeness and flexibility for evaluating agility in the supply chains and also indicated that there is no need of any constraints for decisions for the processing of evaluation of agility . CP has several social, economic and environmental benefits. So, for the successful implementation by

overcoming barriers of it, Silva et al. (2013) [24] proposed a CP technique in which the various Quality Tools (QTs) are integrated. To obtain the emission reduction and energy conversation for a medium-scale ceramic tile plant, Huang et al. (2013) [14] presented an extensive application of cleaner production. Corominas et al. (2013) [9] proposed a tool i.e. Life Cycle Assessment (LCA) to enhance the efficiency of wastewater treatment plants by choosing the best strategy. Cheung et al. (2015) [7] investigated the disposal costs by using original Equipment Manufacturer (OEM) and then on the basis of this cost, the decision will be taken whether the EOL parts to be recycled or destroyed. Zhang et al. (2016) [28] proposed an overall architecture i.e. Big data-based analytics for product lifecycle (BDA-PL), for making better product lifecycle management (PLM) and cleaner production (CP) decisions based on data.

V. GAPS IN LITERATURE SURVEY

Zhang et al. (28) focused on to make better CP and PLM decisions based on the big amount of real-time data and multi-source heterogeneous big data. They proposed an overall architecture of big data based analytics for product lifecycle (BDA-PL):

1. This architecture simply integrated big data analytics and service- driven patterns that helped to avoid the problems of CP such as lack of complete data and valuable information.

2. The accessibility and availability of data and knowledge related to merchandise were also occurred in this architecture.
3. They also focus on MMP of product lifecycle and for implementing big data analytics, various key technologies were developed.

The result showed that this architecture benefited to consumers, developers, and environment and also helped a lot for the successful implementation of CP. But still there are few areas where improvement is possible.

Some of the issues not covered in the study are as follow:

i. The review has shown that the not much work is done for software reuse analytics.
ii. Computational speed is still found to be challenging issue in big data analytics.
iii. The use of data preprocessing is also ignored by existing researchers.

To handle above stated issues a new technique will be proposed by use in future work, which will use unsupervised filtering and Random forest based machine learning technique for software reuse analytics to enhance the performance.

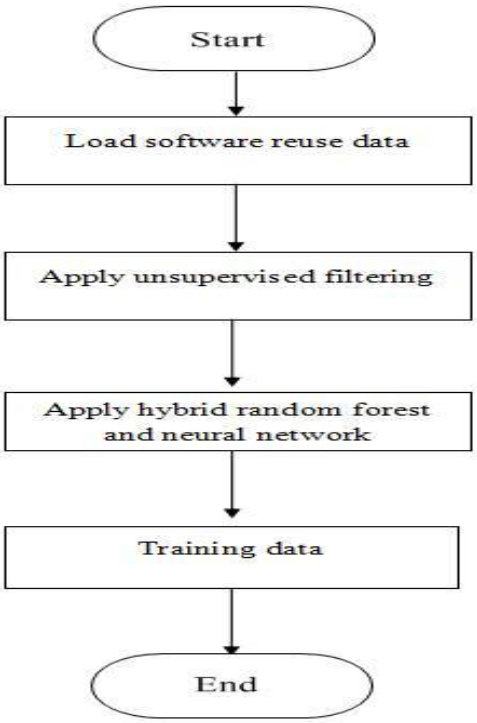
Reference	Title	Objective	Parameters			
			Material recycling	Production cost	Energy consumption	Minimize waste generation
Kupusovic et al. [13]	Cleaner production measures in small-scale Slaughterhouse industry e case study in Bosnia and Herzegovina	To minimize use of water, handle pollution at low cost, efficient use of resources & reduction of waste material at source.	✓	Low	Less	✓
Giannetti et al.[12]	Cleaner production practices in a medium size gold-plated jewelry company in Brazil: when little changes make the difference	To reduce waste material and pollution, a waste minimization project is used which reduces 86% waste and 36% electricity consumption.	✓	Low	Less	✓
Calia et al. [5]	The impact of Six Sigma in the performance of a Pollution Prevention program.	To reduce pollution and production cost, they focus on to implement expressive six sigma programs.		Low	Less	✓
Huang et al. [14]	Application of cleaner production as an important sustainable strategy in the ceramic tile plant e a case study in Guangzhou, China.	To reduce energy consumption by 4.3% and water by 22.33%, use a cleaner production application in tile plant.	✓	Low	Less	✓
Corominas et al. [9]	Including Life Cycle Assessment for decision-making in controlling wastewater nutrient removal systems.	Life Cycle Assessment (LCA) cost effective tool is used to handle the wastewater removal systems.		Low	Less	✓
Silva et al. [24]	Quality Tools Applied to Cleaner Production Programs: A First Approach Towards a New Methodology.	A systematic integration of QTs is proposed, to implement and overcome the barriers of CP process ,	✓	Low		✓
Cheung et al. [7]	Towards cleaner production: a roadmap for predicting product end-of-life costs at early design concept.	To predict disposal cost by OEM, which will help to get a solution whether the EOL part to be recycled or destroyed.	✓	Low		✓
Zhang et al. [28]	A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products.	An architecture BDA-PL is proposed, to make the better PLM and CP decisions.	✓	Low	Less	✓

Table 1: Comparison table gives summary of various parameters essential for CP.

VI. METHODOLOGY

To validate the performance using R tool, the methodology has presented below in the form of flow chart. The methodology contains two parts i.e. for training data and for testing data. The steps of flow chart for training data are given as follow:

a.) For training data

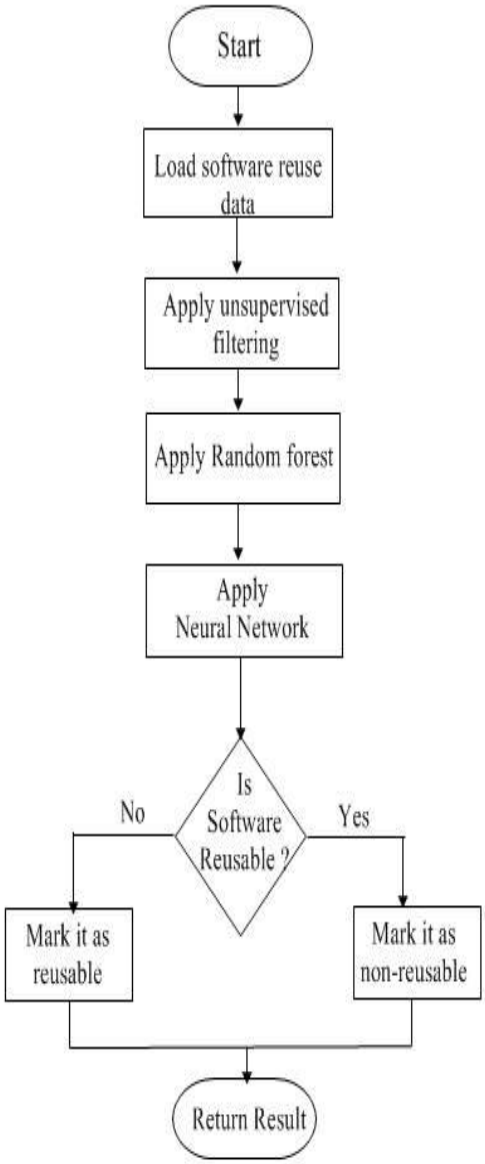


Step 1: load software reuse data from the data set for training the data.
Step 2: Apply unsupervised filtering on data for unknown clusters.
Step 3: Apply hybrid random forest and neural network for accuracy and performance, respectively.
Step 4: So, finally get the training data with potentially predictive relationships among the data.

b.) For testing data

Step 1: Load software reuse data for testing the data.

Step 2: Same as training data, here also apply unsupervised filtering for unknown data.



Step 3: Apply random forest for constructing trees and returns the final prediction as output with accurate results.

Step 4: Apply neural network method on the output data of random forest for high performance.

VII. RESULTS AND DISCUSSIONS

Based on the methodology as described in the previous section, the extensive simulations as been performed and the results were obtained. The analysis of proposed algorithm is done on the basis of seven parameters. These are: Accuracy, Error rate, Kappa statistics, RAE, RMSE, MAE and RRSE. This section mainly deals with the results and discusses the observations.

For the comparisons of the existing and proposed algorithms, J48 classifier is used.

else

mark it as non- reusable and return the results.

Step 5 : If the software is reusable then mark it as reusable and

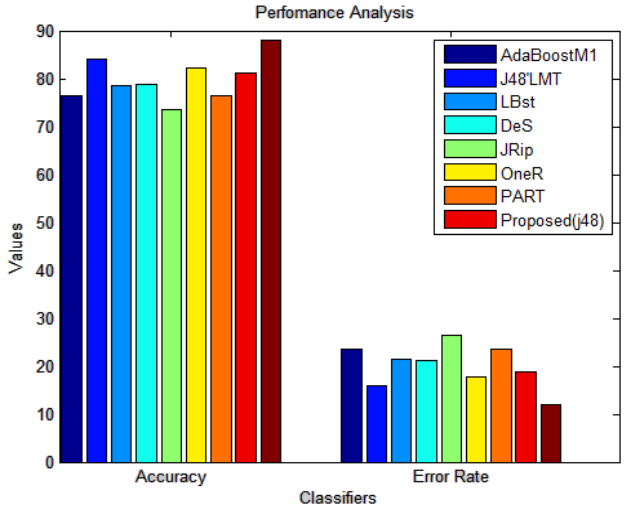
return the results.

Table 2: Comparison of various existing classifiers on the basis of some parameters with the proposed algorithm

.

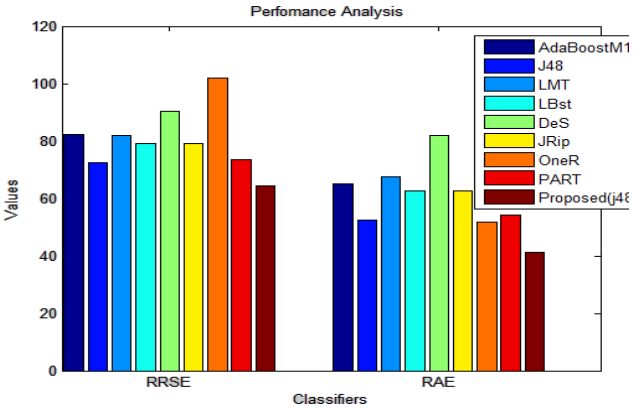
	AdaBoostM1	J48	LMT	LogitBoost	Decision Stump	JRip	OneR	PART	Proposed
Accuracy	76.5625	84.1146	78.5156	78.9063	73.5677	82.2917	76.4323	81.25	87.9182
Error rate	23.4375	15.8854	21.4844	21.0938	26.4323	17.7083	23.5677	18.75	12.0818
Kappa statistics	0.4723	0.6319	0.5026	0.5067	0.4257	0.5999	0.4484	0.6184	0.7346
Mean Absolute error	0.2956	0.2383	0.3069	0.2853	0.3719	0.2841	0.2357	0.2466	0.1877
Root mean squared error	0.3922	0.3452	0.3908	0.3777	0.4312	0.3769	0.4855	0.3512	0.3064
Relative absolute error	65.0313	52.4339	67.523	62.7697	81.8217	62.5074	51.8551	54.2692	41.2768
Root relative squared error	82.2787	72.4207	81.9891	79.251	90.4671	79.0719	101.8515	73.6772	64.2588

Various classifiers are used for the comparisons such as LMT, JRip, PART, DecisionStump and J48. Among all the classifiers, J48 has the highest performance than all other classifiers. So, Table 2 shows the comparison of both existing J48 algorithm and the proposed J48 algorithm. The results of comparisons are showed in the bar graphs with respective parameters as follow:



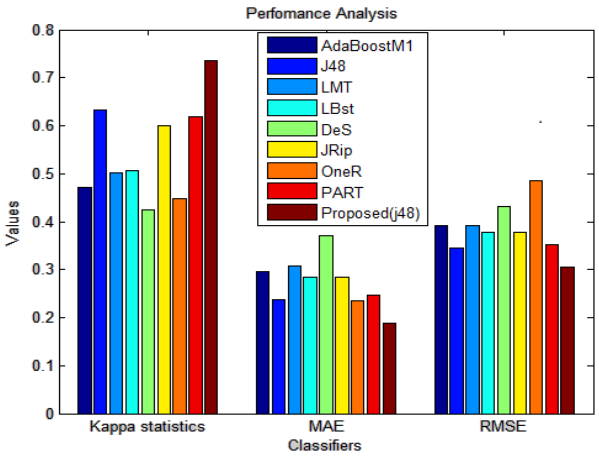
a.) Accuracy : the graph shows that there is 4.5% increase in accuracy of proposed algorithm than the existing J48 claasifier which shows better performance as the accuracy is increased.

b.) Error rate :23 % decrease of error rate using proposed technique shows high performace of classifier.



c.) RRSE : this graph shows that there is 11 % decrease in RRSE by using the proposed algorithm.

d.) RAE : less the error rate, more accurate the algorithm. Here,21 % decrease of RAE by using the proposed algorithm.



e.) Kappa Statistics : the value of proposed J48 is increased by 16% by using the proposed algorithm.

f.) MAE : graph shows there is decrease in MAE by 21% , which will help to get the more accurate results.

g.) RMSE :11 % decrease by using proposed algorithm.

VIII. Conclusion

As, the Cleaner Production (CP) strategy was facing barriers, such as the lack of complete data and valuable knowledge that can be employed to provide better support on decision-making of coordination and optimization on the product lifecycle management (PLM) and the whole CP process. So, for software reuse analytics, integrated Random forest and GBM technique is used for accuracy and as well as for speed, respectively. This paper presents an algorithm which improves the performance by increasing accuracy and decreasing error rate and also improves the values of some other parameters. As, also computational speed was still found to be challenging issue in big data analytics, the proposed algorithm helps to increase the computational speed.

References

[1] Auschitzky, E., Hammer, M., Rajagopaul, A., 2014. How big data can improve manufacturing. McKinsey Glob. Inst. Available at: http://www.mckinsey.com/insights/operations/how_big_data_can_improve_manufacturing (accessed 25.05. 2015).

[2] Ball, P.D., Evans, S., Levers, A., Ellison, D., 2009. Zero carbon manufacturing facility -towards integrating material,energy, and waste processflows. Proc. Inst. Mech. Eng. Part B: J. Eng. Manuf. 223 (9), 1085-1096.

[3] Bennane , A., Yacout,S., 2012. LAD-CBM; new data processing tool for diagnosis and prognosis in condition-based maintenance. J. Intell. Manuf. 23(2), 265-275.

[4] Beuren, F. H., Ferreira, M. G. G., Miguel, P. A. C. 2013. Product-service systems: a literature review on integratedproducts and services. J. Clean. Prod. 47, 222-231.

[5] Callia, R.C., Guerrini, F.M., de Castro, M., 2009. The impact of Six Sigma in the performance of a pollutionprevention program. J. Clean. Prod.17 (15),1303-1310

[6] Chen, Y. S., Cheng, C. H., Lai, C. J., 2012. Extracting performance rules of suppliers in the manufacturing industry: An empirical study. J. Intell. Manuf. 23(5), 2037-2045.

[7] Cheung, W.M., Marsh, R., Griffin, P.W., Newnes, L.B., Mileham, A.R., Lanham, J.D., 2015. Towards cleaner production: a roadmap for predicting product end-of-life costs at early design concept. J. Clean. Prod. 87, 431-441.

[8] Choudhary, A. K., Harding, J. A., Tiwari, M. K., 2009. Data mining in manufacturing: a review based on the kind of knowledge. J. Intell.Manuf. 20(5), 501-521.

[9] Corominas, L., Larsen, H.F., Flores-Alsina, X., Vanrolleghem, P.A., 2013. Including life cycle assessment for decision-making in controlling wastewater nutrient removal systems. J. Environ. Manag. 128, 759-67.

[10] Dai, J.Q., Huang, J., Huang, S.S., Huang, B., Liu, Y., 2011. Hitune: data flow-based performance analysis for big decision-making in controlling wastewater nutrient removal systems. J. Environ. Manag. 128, 759-67.

[11] Galletti, A., Papadimitriou, D.C., 2013. How big data analytics are perceived as a driver for competitive advantage: a qualitative study on food retailers. Master thesis, 1-58.

[12] Giannetti, B.F., Bonilla, S.H., Silva, I.R., Almeida, C.M.V.B., 2008. Cleaner production practices in a medium sizegold-plated jewelry company in Brazil: when little changes make the difference. J. Clean. Prod. 16, 1106-1117.

[13] Hadaya, P., Marchildon, P., 2012. Understanding product lifecycle management and supporting systems. Ind. Manage. Data. Syst. 112(4), 559-583

[14] Huang, Y., Luo, J., Xia, B., 2013. Application of cleaner production as an important sustainable strategy in the ceramic tile plant e a case study in Guangzhou, China. J. Clean. Prod. 43, 113-121.

[15] Jun, H. B., Shin, J. H., Kim, Y. S., Kiritsis, D., Xirouchakis, P., 2009. A framework for RFID applications in product lifecycle management.Int. J. Comp. Integ. M. 22(7), 595-615.

[16] Kiritsis, D., Bufardi, A., Xirouchakis, P., 2003. Research issues on product lifecycle management and information tracking using smart embedded systems. Adv. Eng. Inform. 17(3), 189-202.

[17] Kupusovic, T., Midzic, S., Silajdzic, I., Bjelavac, J., 2005. Cleaner production measures in small-scale slaughterhouse industry e case study in Bosnia and Herzegovina. J. Clean. Prod. 15, 378-383.

[18] Laney, D., 2001. 3D data management: controlling data volume, velocity and variety. META Group Research Note. Available http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-DataManagement-Controlling-Data-Volume-Velocity-and-Variety .pdf

[19] Li, J.R., Tao, F., Cheng, Y., Zhao, L.J., 2015. Big Data in product lifecycle management. Int. J. Adv. Manuf. Tech. 84(1-4), 667-684.

[20] Metan, G., Sabuncuoglu, I., Pierreval, H., 2010. Real time Selection of Scheduling Rules and Knowledge Extraction Via Dynamically Controlled Data Mining. Int. J. Prod. Res. 48(23): 6909-6938.

[21] Murillo-Luna, J.L., Garces-Ayerbe, C., Rivera-Torres, P., 2011. Barriers to the adoption of proactive environmentalstrategies. J. Clean. Prod. 19 (13), 1417-1425.

[22] Ngai, E. W., Xiu, L., Chau, D. C., 2009. Application of data mining techniques in customer relationship management: A literature review and classification. Expert. Syst. Appl. 36(2), 2592-2602.

[23] Rabl, T., Gómez-Villamor, S., Sadoghi, M., Muntés-Mulero, V., Jacobsen, H.-A.,Mankovskii, S., 2012. Solving big data challenges for enterprise applicationperformance management. Proc. VLDB Endow. 5 (12), 1724-1735.

[24] Silva, D.A., Delai, I., Castro, M.A., Ometto, A.R., 2013. Quality Tools Applied to Cleaner Production Programs: A First Approach Towards a New Methodology. J. Clean. Prod. 47, 174-187.

[25] Vinodh, S., Prakash, N.H., Selvan, K.E., 2011. Evaluation of agility in supply chains using fuzzy association rules mining. Int. J. Prod. Res. 49(22): 6651-6661

[26] Wang, H. W., Chen, S., Xie, Y., 2010. An RFID-based digital ware-house management system in the tobacco industry: A case study. Int. J. Prod. Res. 48(9), 2513-2548.

[27] Wei, F. F., 2013. ECL Hadoop: "Big Data" processing based on Hadoop strategy in effective e-commerce logistics. Comput Eng Sci 35(10):65–71.

[28] Yingfeng Zhang, Shan Reh, Yang Liu, Shubin Si., 2016. A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products. J. Clean. Prod. JCLP 7965.