Machine Learning Approach for Predicting Production Delays: A Quarry Company case study

Rathimala kannan (rathimala.kannan@mmu.edu.my)
Multimedia University
Haq'ul Aqif bin Abdul Halim
PETROPRO (Malaysia) Sdn Bhd
Kannan Ramakrishnan
Multimedia University
Shahrinaz Ismail
Albukhary International University
Dedy Rahman Wijaya
Telkom University

Research Article

Keywords: Machine Learning, production delay, prediction models, Quarry Industry

Posted Date: February 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1317868/v1

Abstract

Predictive maintenance employing machine learning techniques and big data analytics is a benefit to the industrial business in the Industry 4.0 era. Companies, on the other hand, have difficulties as they move from reactive to predictive manufacturing processes. The purpose of this paper is to demonstrate how data analytics and machine learning approaches may be utilized to predict production delays in a quarry firm as a case study. The dataset contains production records for six months, with a total of 20 columns for each production record for two machines. Cross Industry Standard Process for Data Mining approach is followed to build the machine learning models. Four predictive models were created using machine learning algorithms such as Decision Tree, Neural Network, Random Forest, and Nave Bayes. The results show that Multilayer Perceptron Neural Network outperforms other techniques and accurately predicts production delays with a sensitivity score of 0.979. The quarry company's improved decision-making reducing potential production line delays demonstrates the value of this study.

Introduction

The mining and quarrying industry is regarded as a potentially substantial contributor to Malaysia's economy [1]. Perceiving knowledge and learning from data is a major difficulty in industrial organizations, especially those in the quarry and mining industries. Real-time data analytics faces numerous challenges in real-world settings, while a significant amount of legacy, enterprise, and operational data stays untapped [2]. In this research, the case study company has been operating for more than 40 years, providing a firm foundation of quality stones and rocks to all construction works in Malaysia, ranging from road stones, housing constructions, bituminous products, railways, and airport runways. The case company's bestselling products are high quality aggregate as well as premix products. The success of the company is largely related to their professional expertise and the ability to provide rapid and efficient services to their customers, especially in providing quality granite products at a very competitive price. Consequently, the company is recognised as a progressive and viable business entity, contributing effectively towards the nation's economic development in general and in the states of Negeri Sembilan, Selangor, and the Federal Capital, Kuala Lumpur in particular. In this company, prediction of potential production delays including identification of the causative factors is very important so that it can immediately mitigate and improve company performance.

The goal of this research is to discover potential delays in the quarry company's production so that they can increase operational efficiency by lowering the causal and important elements that affect production time and output. A predictive model was developed to assist and aid the company identify the potential and causation of the delay, therefore offering data-driven decision making in decreasing the prospective delay, based on a research opportunity in the area of machine leaning for prediction of production efficiency. The dataset consists of six (6) months period of production records, which include a total of 20 columns for each production record for two machines, namely Machine 1 (C1008) and Machine 2 (C125). Cross Industry Standard Process for Data Mining (CRISP-DM) approach is followed to build the machine learning models. Four predictive models were built by applying machine learning techniques i.e.

Decision Tree, Neural Network, Random Forest and Naïve Bayes. The results of the potential production line delay provide insight into operational efficiency.

The rest of this paper is organized as follows. The related works section contains the works of literature related to technological trend for business processes and operations improvement including applications of machine learning techniques for prediction tasks. The research methodology section discusses the CRISP-DM approach to solve this problem. The findings are explained in the Result and discussion section. Finally, we draw the conclusion in the conclusion section.

Related Works

“Data mining and analytics have played an important role in knowledge discovery and decision making in the process industry over the past several decades” [3]. Machine learning serves as a computational engine to data mining and analytics, in which it is used for information extraction, data pattern recognition and predictions. Machine learning techniques have been successfully reported for prediction such as rainfall amount [4], poverty level prediction [5–7], income of campus alumni [8], and COVID-19 related cases [9,10], etc. Predictive modelling approaches in business process management provide a way to streamline operational business processes [11]. Process mining can discover the process workflows in the company, activity actions, and the mechanism of machines [12], as well as allowing the identification or diagnosis of fact-based problems [13]. Process mining explores the discrepancy between data of events, i.e. observed behavior, and models of processes to detect anomalies, compliance checks, predicts delays, facilitates decision-making, and suggest process redesigns [12]. Nevertheless, machine learning algorithms could be adopted into the process mining techniques, in producing predictive analysis and models.

Industry revolution 4.0 emphasizes the use of technology to improve production operations in the manufacturing industry. This drive attracted many academic researchers and experts to focus on applications of machine learning in production operations for fault diagnosis and machine maintenance [14]. Various types of machine learning techniques have been used in prediction of production delay by the existing literature and this study have used four machine learning techniques; Decision Tree, Neural Network, Random Forest and Naïve Bayes which performed better than other supervised learning algorithms. Decision tree and random forest algorithms are commonly used in fault diagnosis and are considered as classification techniques. While decision tree algorithm builds one optimal decision tree model for predicting the target, random forest algorithm builds a number of decision trees and the final prediction is based on the voting of outcome from each decision tree [15]. Depending on the dataset and variables used, the performance of these two techniques vary where decision tree outperforms random forest and the vice versa [16,17]. Artificial neural network (ANN) technique, on the other hand, is popularly known for their noise tolerance and is capable of diagnosing a predetermined fault type. ANN has been used for fault detection in die-casting industry [18], to predict faults in a blade pitch system [13] and many more in manufacturing industry. Naïve Bayes algorithm is one of the popular machine learning technique used in predictive models because of its efficiency and it can perform well with a small training

dataset [19]. In this paper, these four machine learning techniques have been applied to build predictive models to determine the production delays in the case study company.

Research Methodology

This study adopts CRISP-DM approach to analyze the problem and apply data analytics using machine learning techniques to build predictive model that could be implemented to improve operational efficiency of the production line [20]. Fig 1 shows the phases in CRISP-DM, followed by the details of the research process undertaken in this research.

Business Understanding

The CRISP-DM starts with the Business Understanding phase, which consists of identifying business goal and data mining goal. As mentioned in above, the dataset used for this research came from case study Company that has been operating for more than 40 years. As part of their initiative to improve their overall performance especially in the production and operation sector, the case company is executing a project that requires them to explore on data analytics and machine learning. The source of delay, which can affect the ultimate output and production in days, is a common issue that occurs throughout the manufacturing process. As a result, the business goal is to identify significant delay reasons, which will aid in making manufacturing more efficient and removing the potential and cause of delays.

The purpose of data mining is to use the details of production operations to predict the cause of the delay. To achieve the data mining purpose, four machine learning algorithms are used to develop prediction models: Naive Bayes, Decision Tree, Neural Network, and Random Forest. Stratified sampling is utilized in dealing with imbalanced data. To evaluate the performance of the various predictive models, standard metrics such as sensitivity, precision and accuracy are used. KNIME analytics platform, open source data science software was used to carryout data mining process.

Data Understanding

The target dataset was obtained from the Operation Department of the case company containing 180 rows and 24 columns such as Job Start, Job End, Total Operation Time, Operation Start, Maintenance Plan, Maintenance Unplanned, Insurance Briefing, Full Stockpile, Blasting, Pump Cleaning, Out of Stone, Rain, Stone Stocked, Late Lorry, Quarry-Top Full Water, Road Expansion Quarry-Top, Real-Time Operation Hour, Lorry Trip, Total Output and Total Tonnes per Hour. The outcome variable is column “Delay”, which has True / False values indicating whether the production has occurred delay or not in the current production period. Fig 2 below shows the screenshot of the dataset for a few samples before data pre-processing.

Data Preparation

Primarily in Data Preparation phase, this research explored the dataset to see whether the input dataset is standardized and any missing values are observed. In the preliminary process, we observed that the data

for each month had different format and was not standardized. Therefore, major preparations were made to standardize all parameters in the data for each month. On the other hand, the dataset has many missing values that were represented by "-" in most delay predictors, in which later was changed to "0" to signify that there is not delay value within the predictor. Besides that, the dataset is segregated into Machine 1
and Machine 2 to different spreadsheets, which later was restructured with added columns labelled "Machine" and "Delay". Furthermore, the row where the Date falls on a holiday and no production was produced is removed as there is no data input in the dataset. As a result, there are 151 rows and 23 columns of data merged for both machines into one spreadsheet.

Most of the pre-processing work was mainly on formatting the parameters, in which most data input was not in the same category. Prior discussion suggested that the predictor operation time data input was labelled in the unit of hours. However, the input was not standardized as some data was in time format and few in number format. Besides that, the time format is changed to 24 hours formatting. The column "operation" was removed as it was observed redundant and overlapping the category of real-time operation. Therefore, it was not used in the prediction model since most of the data under the "operation" column is empty. Apart from that, all the variables were combined into one spreadsheet to ensure that it is readable by the software.

Observing the distribution of number of delays, it is found that 19 occurrences of the production days are delayed due to Maintenance Unplanned and 67 occurrences are delayed caused by Late Lorry. Hence the delay was labeled as two categories where "True" means the production are delayed and "False" means that there is no delay in the production. All the columns were normalized using min-max normalization method as part of the pre-processing to apply neural network technique.

Modelling

In this step, various machine learning techniques were used to develop predictive models. Neural Network, Naive Bayes, Random Forest and Decision Tree are used in predicting the cause of delays in the production days. Naïve Bayes models can produce robust predictions if the predictors have small correlations, even with a simple architecture [3]. Decision Trees are easy to interpret and are capable of giving insights about the important features. Random Forest is an improved version of decision tree, which can produce really good and robust predictions [15]. Artificial Neural Network (ANN) allow complex nonlinear relationships between the target variable and its predictors [13].

Stratified sampling method was used with k-fold cross validation to handle imbalance dataset. Ten numbers of validations are set for training and testing of data. Besides that, cross validation helps us to evaluate the quality of the model, facilitating us in selecting the model that will perform the best on unseen data and help avoid overfitting and under fitting of the dataset. Lastly, for the performance evaluation precision, accuracy and sensitivity are chosen to determine which model of the machine learning would give the best results. Fig 3 shows the overall KNIME workflow in predicting the production

delay within a manufacturing company. The overall workflow consists of three major parts, which is the descriptive analysis, the unsupervised clustering and supervised learning classification.

It is noticed that the dataset has no obvious segregation of groups, therefore, clustering is required to cluster all data sets in which is deemed fit. For this study, K-mean clustering technique was applied to cluster our dataset [3]. Based on the Silhouette coefficient score, the optimal number for k = 3 was selected. The three clusters are labelled as Low Performance, Medium Performance and High Performance. From the clustering process, the data is segmented into high performance and low performance production which can be viewed through its productivity. Through observation, it is found that the high performance operation has a higher number of delays compared to the low performance operation, whereby high performance has 67.5% of the occurrence delayed production whereas the low performance only has 50% of the occurrence labelled as delayed production. Therefore, the company would have to prepare themselves should they receive a job that requires a high number of productions.

Once pre-processing is completed for each prediction technique, four predictive models have been built, which consist of Decision Tree, Neural Network, Random Forest, Naive Bayes. As the data set is small, k-fold cross validation method of sampling the data is used, in which the number of validation is set to ten (10). K-fold cross validation allows the machine learning process to increase the accuracy of the prediction by learning the concepts from all type of data. After the Machine Learning process is complete, the performance of the predictive models is calculated.

Evaluation

In this phase of CRISP-DM, all the machine learning models will be evaluated and compared to select the best model to predict potential delay in the case company production operations. Most commonly used performance evaluation metrics such as accuracy, sensitivity and precision are calculated and compared for all the four machine learning models.

Deployment

The best model is recommended for deployment with the insights found from the dataset for data-driven decision making after analyzing the performance of multiple machine learning models using standard metrics; accuracy, sensitivity and precision.

Results And Discussion

Descriptive Analytics

The dataset consists of 63% of delayed operations and 37% of regular operations. It is observed that Machine 1 was underutilized and Machine 2 was over utilized as shown by Figure 4. Thus, there are more delay occurred in Machine 2 operation line compared to Machine 1 operation line. In addition to that, the other descriptive analysis conducted is the box plot analysis where the outcome shows the average of total ton per hour is 187.7 whereby the company has to achieve more than the average

number as this value is considered as a benchmark value for the company to sustain a good productivity.

Besides that, the correlation analysis depicted in Fig 5 shows that the main factor causing the delay is late lorry, which has a positive correlation to the delay. On the other hand, maintenance unplanned has a significant negative correlation to the real-time operation. These two factors are significant enough for the company to investigate the cause and predict future occurrence as well as prepare a mitigation plan in order to reduce the number of occurrences. Fig 6 boxplot illustrates basic statistics and outlier of the production dataset.

Table 1 depicts the summarized results of the performance evaluation of all machine learning models. Using Decision Tree model, it is found that the overall accuracy is 0.927. This is because the true samples (i.e. production) are much higher compared to false samples. Neural Network is then applied and the overall accuracy shows 0.98, which is the highest among other models applied in this project, where Random Forest gives overall accuracy of 0.954 and Naive Bayes gives 0.384 as the two lowest values out of all model. However, as the goal is to predict the all possible production delay, sensitivity score is the best to compare the model performance. The Neural Network delivers the greatest performance in predicting operating delay in the production line, based on sensitivity score (0.979).

Table 1. Performance evaluation of Machine Learning models

Machine Learning Technique	Accuracy	Delay status	Sensitivity	precision
Decision Tree	0.927	True	0.926	0.956
		False	0.93	0.883
Neural Network – Multilayer perceptron	**0.98**	True	**0.979**	0.989
		False	0.982	0.966
Random Forest	0.954	True	0.936	0.989
		False	0.982	0.903
Naïve Bayes	0.384	True	0.011	1
		False	1	0.38

It is suggested that Neural Network performs greater among the four models. The production delay can be predicted through Neural Network as it can accurately predict the outcome. Based on the findings, the following recommendations are proposed to the company:

To avoid overusing one machine and causing production delays, the company should use both machines equally and efficiently. The important factors that contribute to low Real Time Operation and Delay include unplanned maintenance and late lorries. The best model to deploy for predicting production

delays is the neural network, which has the best sensitivity and accuracy. For model enhancement, more features such as the number of workers, downtime, and other relevant data should be incorporated to the study.

Conclusion

The goal of this research was to use data analytics and machine learning approaches to create prediction models that would help a quarry company's production line run more efficiently. The results were created in the form of descriptive analysis, clustering, and predictive models using four machine learning techniques: Decision Tree, Neural Network, Random Forest, and Naive Bayes. In a nutshell, this research compares the efficiency of the production line that goes through two main machines, Machine 1 and Machine 2, to identify the potential delay in the production. Taking into account various factors made available in the dataset, it is found that Neural Network, gives the best performance of machine learning model in predicting operation delay, with a high score of 0.98 accuracy and sensitivity score of 0.979. Thus, Neural Network prediction model is recommended for this case study, in which a quarry company could use to provide further decision-making analysis and to strategize an improvement plan to reduce potential delays in production line.

Abbreviations

CRISP-DM: Cross Industry Standard Process for Data Mining; ANN: Artificial Neural Network

Declarations

Acknowledgment

Not Applicable

Authors’ contributions

Rathimala Kannan Conceptualization, Writing-original draft preparation. Haq'ul Aqif bin Abdul Halim Data curation. Kannan Ramakrishnan Writing-Review & Editing. Shahrinaz Ismail Writing-Review & Editing. Dedy Rahman Wijaya Validation.

Funding

No funding received.

Availability of data and materials

The data was collected from the case company and is not available to the general public. The authors' data are, however, available upon reasonable request and with the permission of the case study company.

Ethics approval and consent to participate

This article does not contain any studies with human participants or animals performed by any authors.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

1. ‘New economic powerhouse’ for Malaysia | The Star. 2019.
2. Lepenioti K, Pertselakis M, Bousdekis A, Louca A, Lampathaki F, Apostolou D, et al. Machine Learning for Predictive and Prescriptive Analytics of Operational Data in Smart Manufacturing. Lecture Notes in Business Information Processing. Springer; 2020. p. 5–16.
3. Dogan A, Birant D. Machine learning and data mining in manufacturing. Expert Systems with Applications. Elsevier Ltd; 2021. p. 114060.
4. Liyew CM, Melese HA. Machine learning techniques to predict daily rainfall amount. Journal of Big Data [Internet]. Springer International Publishing; 2021;8. Available from: https://doi.org/10.1186/s40537-021-00545-4
5. Wijaya DR, Paramita NLPSP, Uluwiyah A, Rheza M, Zahara A, Puspita DR. Estimating city-level poverty rate based on e-commerce data with machine learning. Electronic Commerce Research. Springer; 2020;
6. Pangestu A, Wijaya DR, Hernawati E, Hidayat W. Wrapper Feature Selection for Poverty Level Prediction Based on E-Commerce Dataset. 2020 International Conference on Data Science and Its Applications, ICoDSA 2020. Bandung: IEEE; 2020.
7. Aulia TF, Wijaya DR, Hernawati E, Hidayat W. Poverty Level Prediction Based on E-Commerce Data Using K-Nearest Neighbor and Information-Theoretical-Based Feature Selection. 2020 3rd International Conference on Information and Communications Technology, ICOIACT 2020. 2020;28–33.
8. Gomez-Cravioto DA, Diaz-Ramos RE, Hernandez-Gress N, Preciado JL, Ceballos HG. Supervised machine learning predictive analytics for alumni income. Journal of Big Data [Internet]. Springer International Publishing; 2022;9. Available from: https://doi.org/10.1186/s40537-022-00559-6
9. Budiharto W. Data science approach to stock prices forecasting in Indonesia during Covid-19 using Long Short-Term Memory (LSTM). Journal of Big Data [Internet]. Springer International Publishing;

2021;8. Available from: https://doi.org/10.1186/s40537-021-00430-0

10. Hssayeni MD, Chala A, Dev R, Xu L, Shaw J, Furht B, et al. The forecast of COVID-19 spread risk at the county level. Journal of Big Data [Internet]. Springer International Publishing; 2021;8:1–16. Available from: https://doi.org/10.1186/s40537-021-00491-1
11. Breuker D, Matzner M, Delfmann P, Becker J. Comprehensible Predictive Models for Business Processes. Management Information Systems Quarterly. 2016;40.
12. Faizan M, Zuhairi MF, Ismail SB, Ahmed R. Challenges and use cases of process discovery in process mining. International Journal of Advanced Trends in Computer Science and Engineering. World Academy of Research in Science and Engineering; 2020;9:5164–71.
13. Cho S, Choi M, Gao Z, Moan T. Fault detection and diagnosis of a blade pitch system in a floating wind turbine based on Kalman filters and artificial neural networks. Renewable Energy. Elsevier Ltd; 2021;169:1–13.
14. Dalzochio J, Kunst R, Pignaton E, Binotto A, Sanyal S, Favilla J, et al. Machine learning and reasoning for predictive maintenance in Industry 4.0: Current status and challenges. Computers in Industry. Elsevier B.V.; 2020. p. 103298.
15. Gong S, Wu X, Zhang Z. Research on Fault Diagnosis Method of Photovoltaic Array Based on Random Forest Algorithm. Chinese Control Conference, CCC. IEEE; 2020;2020-July:4249–54.
16. Tsai MF, Chu YC, Li MH, Chen LW. Smart machinery monitoring system with reduced information transmission and fault prediction methods using industrial internet of things. Mathematics. MDPI AG; 2021;9:1–14.
17. Zhang C, Hu C, Xie S, Cao S. Research on the application of Decision Tree and Random Forest Algorithm in the main transformer fault evaluation. Journal of Physics: Conference Series. IOP Publishing; 2021;1732:012086.
18. Lee J, Lee YC, Kim JT. Migration from the traditional to the smart factory in the die-casting industry: Novel process data acquisition and fault detection based on artificial neural network. Journal of Materials Processing Technology. Elsevier Ltd; 2021;290:116972.
19. Truong D. Using causal machine learning for predicting the risk of flight delays in air transportation. Journal of Air Transport Management. Elsevier Ltd; 2021;91:101993.
20. Schröer C, Kruse F, Gómez JM. A Systematic Literature Review on Applying CRISP-DM Process Model. Procedia Computer Science. Elsevier; 2021;181:526–34.

Figures

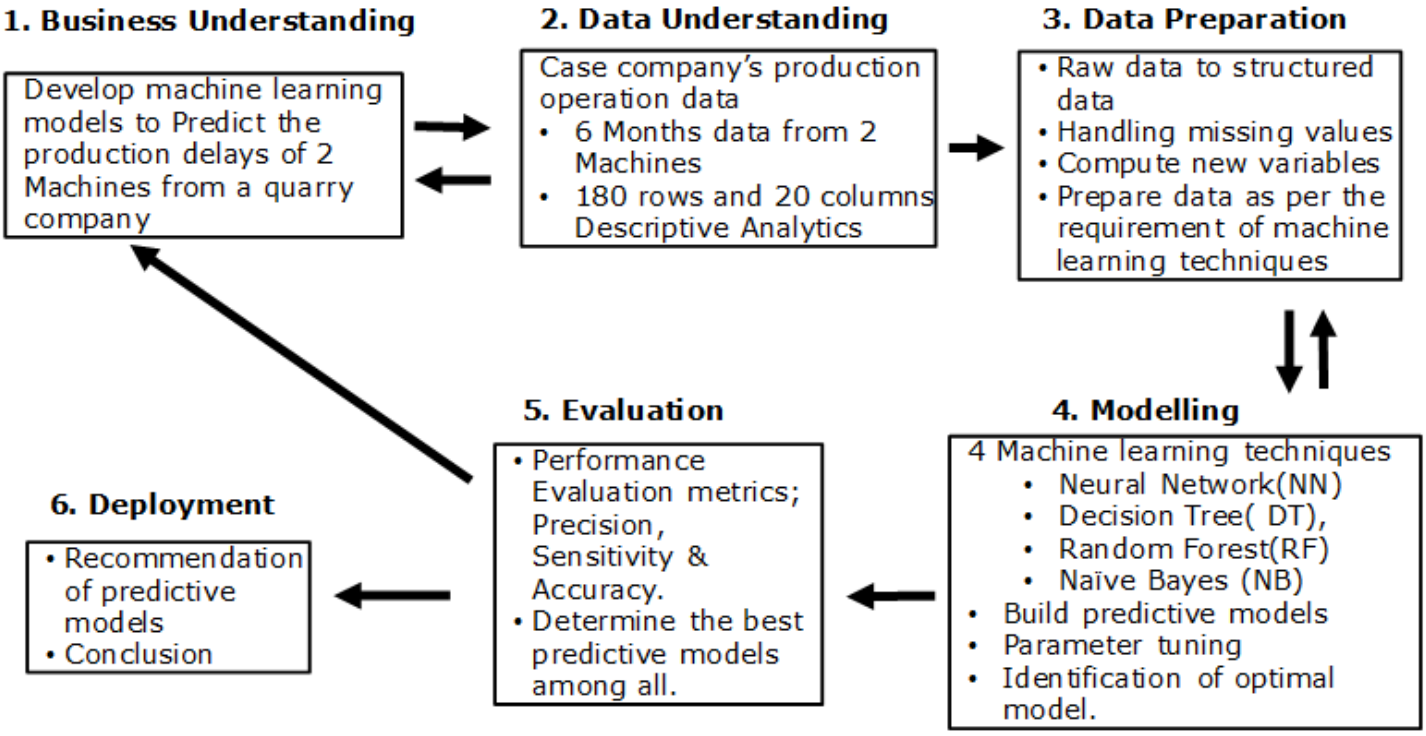

Figure 1

Research Methodology based on CRISP-DM

Figure 2

Preview of the Production Dataset – the raw data

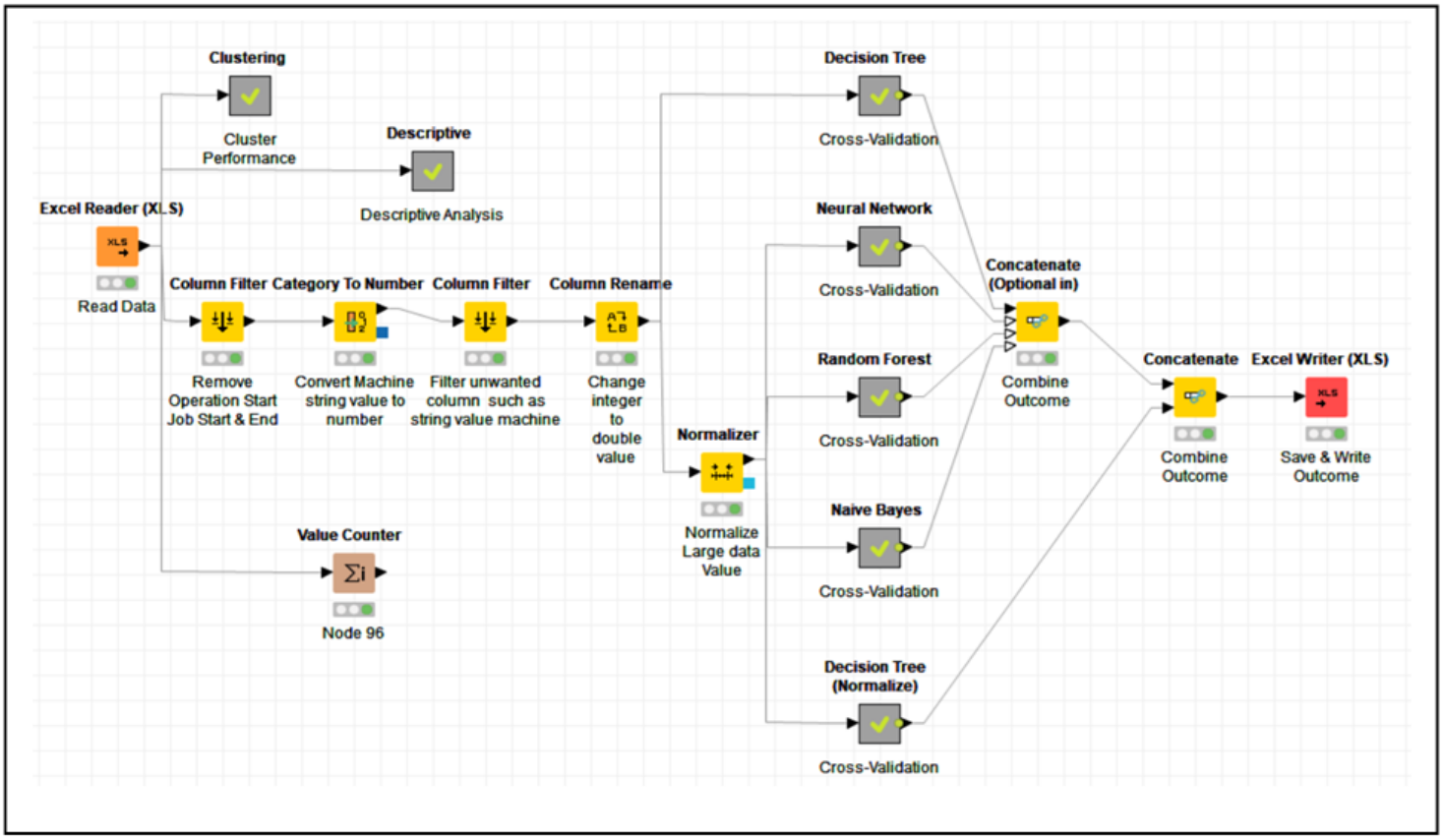

Figure 3

The overall KNIME workflow for the prediction of production delay analysis

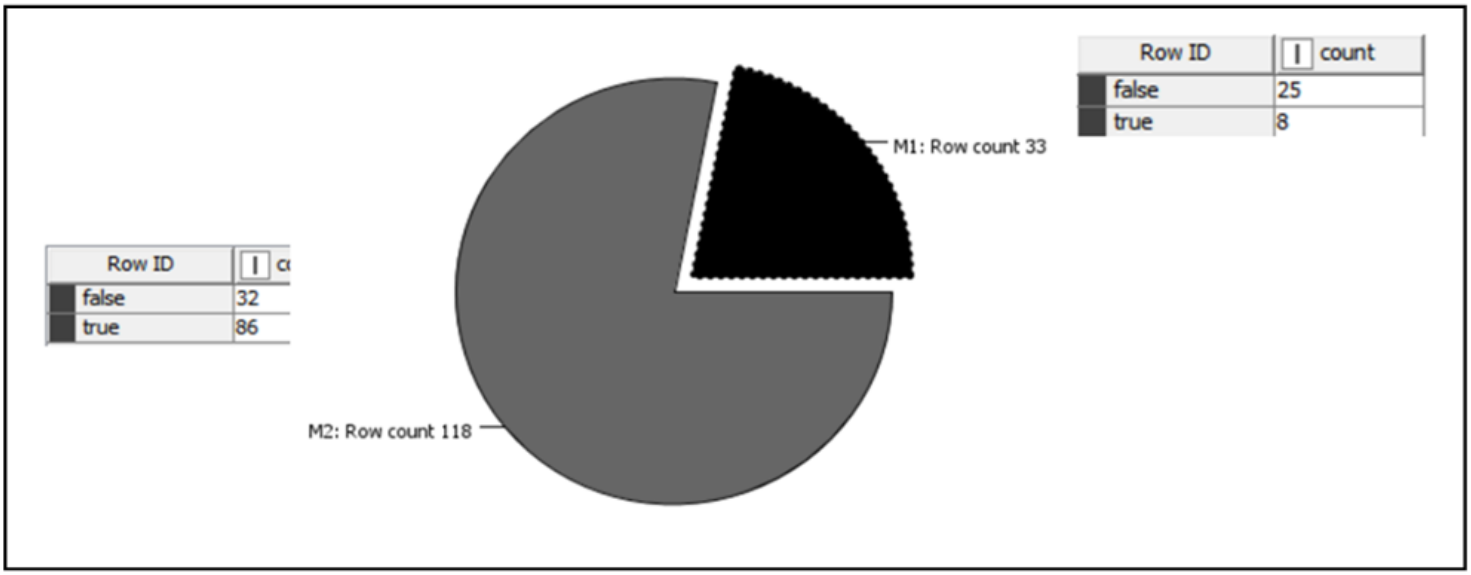

Figure 4

Pie chart on the machine operation occurrences.

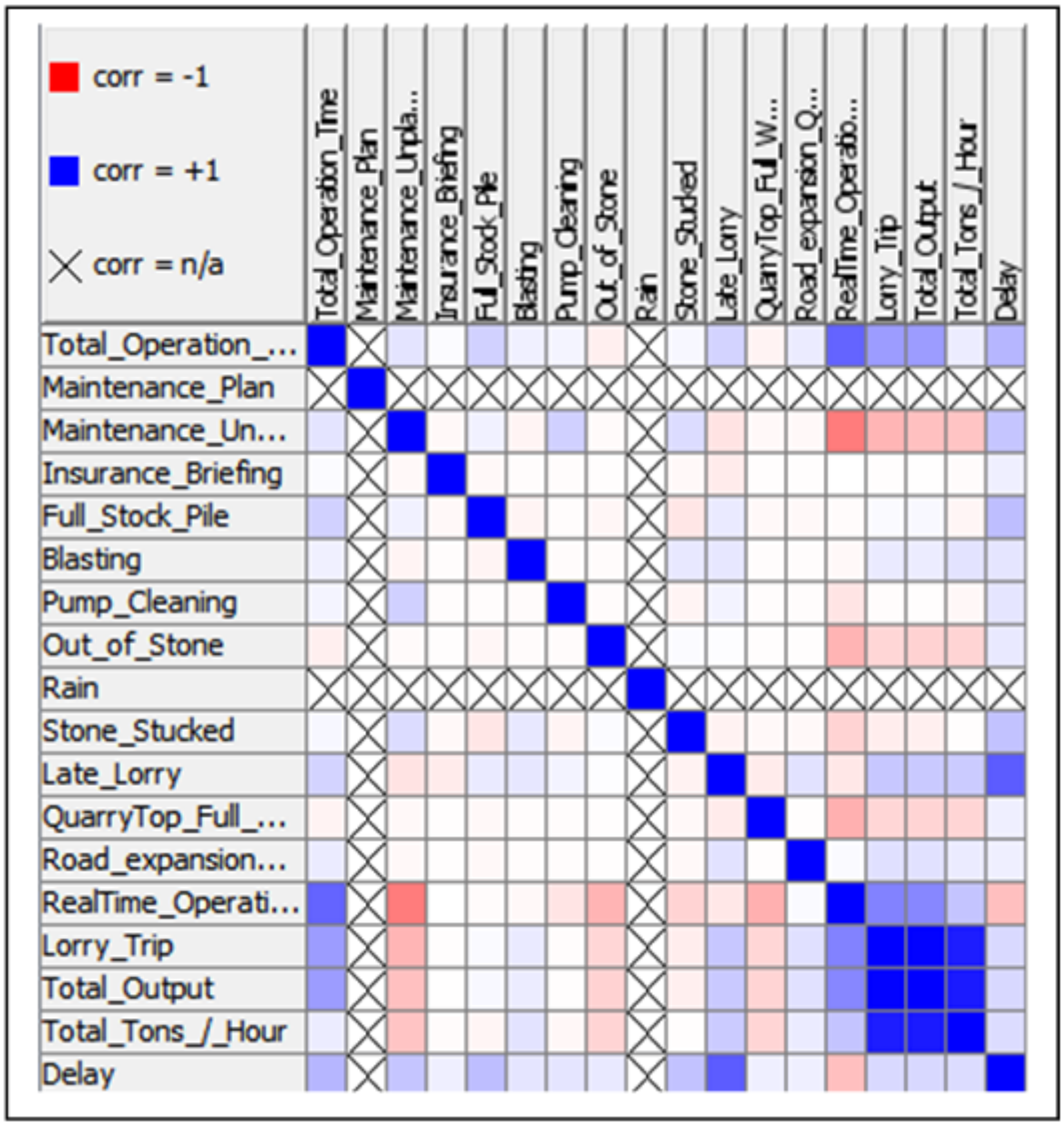

Figure 5

Correlation coefficients of variables.

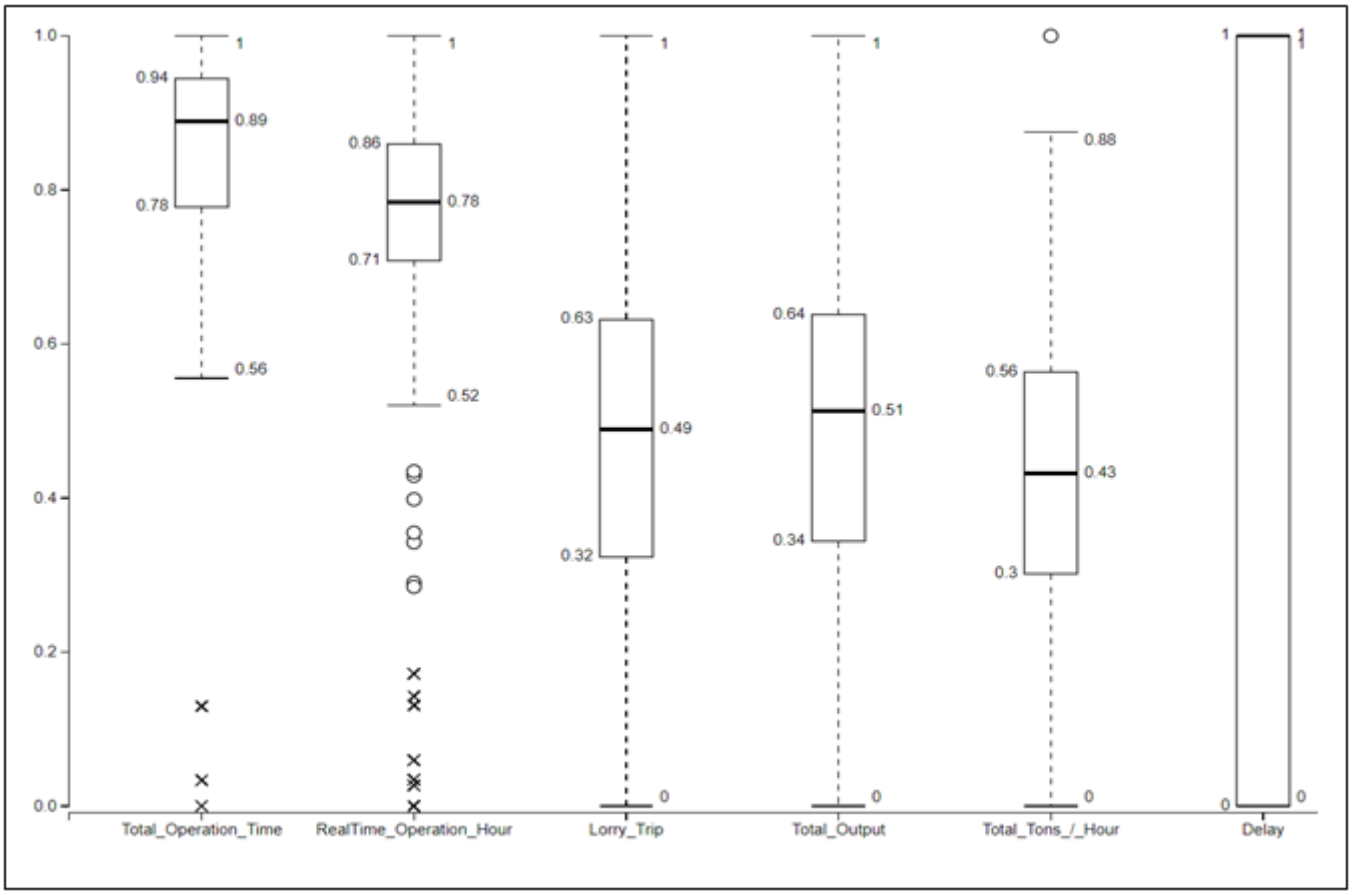

Figure 6

Boxplot illustrating basic statistics and outlier of the production dataset