PressAcademia
www.pressacademia.org/conferences
GBRC
ISBN 978-605-06192-0-1
PressAcademia Procedia YEAR 2020 VOLUME 1
Global Business Research Congress, June 4-5, 2020

PLACE AND SOLUTION PROPOSALS OF DATA MINING IN PRODUCTION PLANNING AND CONTROL PROCESSES: A BUSINESS APPLICATION

DOI: 10.17261/Pressacademia.2020.1265
PAP- V.11-2020(37)-p.189-193

Ezgi Demir[1], Sait Erdal Dincer[2]
[1]Piri Reis University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Istanbul, Turkey.
edemir@pirireis.edu.tr, ORCID: 0000-0002-7335-5094
[2] Marmara University, Faculty of Economics, Department of Econometrics, Istanbul, Turkey.
edincer@marmara.edu.tr, ORCID: 0000-0002-8310-1418

To cite this document
Demir, E., Dincer, S.E., (2020). Place and solution proposals of data mining in production planning and control processes: a business application. PressAcademia Procedia (PAP), V.11, p.189-193
Permanent link to this document: http://doi.org/10.17261/Pressacademia.2020.1265

ABSTRACT

Purpose- In this study, textile goods, of manufactured by a textile company, whether being returned because of defects or not has been investigated by data mining and machine learning techniques. Main purpose of the study is to determine which of the products passing through 15 different production lines during the manufacturing process being defective and faulty at the last stage.
Methodology- In this study, there are 250 different variables and 72959 lines of data on the production line. In order to perform a data mining process, it is firstly necessary to understand the data and determine the process. For this, CRISP-DM algorithm has been used. Modelling and classification algorithms are applied to estimate the production of faulty goods. In the model, a supervised learning model based machine learning methods have been used. The dimensions, loops and some statistical features of the data have been examined, and then it has been studied in the Python programming language. The feasibility of model and success rates have been evaluated with findings.
Findings- The results of the model show that logistic regression and k-nearest neighbour algorithms give above %90 percentage confusion rate. It has been said that with this model is succesful for predicting defective and faulty product in manufacturing line.
Conclusion- It has been tried to predict whether there will be faulty products that reduce quality. With this study, it has been aimed to give a signal to the production line in advance.

Keywords: Data mining, machine learning, production planning, fault detection.
JEL Codes: C44, C45, L23

1. INTRODUCTION

Companies would like to maximize their profits in an increasingly competitive environment, shorten the processing time and increase the functionality of the system by making less errors in the process. Performing error detection in the intensive manufacturing process creates confusion in the system where multiple variables are taken into account, which parameter(s) have more affects during the manufacturing process. In this study, the process data has been examined about the manufacturing process of the consumer goods in the last one year by a company being one of the Turkey's leading textile manufacturing.

Quality has an important place in the manufacturing sector, hence quality is the main focus of workforce efficiency and customer satisfaction. In addition, the process quality of the goods is also important at the point of purchase and distribution of the sellers. Performing error detection in the intensive manufacturing process creates confusion in the system where multiple variables are taken into account, which parameter(s) have more affects during the manufacturing process. It is difficult to estimate the total quality because of large and complex data obtained during production. In addition, statistical methods should be employed to find which variables cause a decrease in quality in which range. However, statistical methods are complicated and use a lot of time. Excess data reduces the functionality of statistical methods. In addition, while statistical methods can evaluate the quality after the production is finished, then it cannot make a forward estimate and classify. Main purpose of the study is to determine which of the products passing through 15 different production lines during the manufacturing process being defective and faulty at the last stage. It has been tried to predict whether there will be faulty products that reduce quality. With this study, it has been aimed to give a signal to the production line in advance. For this purpose, in this study it can be tried to solve a problem in the literature in second section. Then, it has been given the methodology and data in third section. Afterwards, it has been finished with conclusion and references.

2. LITERATURE REVIEW

In the literature, in the latest studies in terms of national and international articles on data mining in the production sector are listed as follows. Customer product ratings were made using SVM (support vector machines) and latent class algorithms in the cinema sector. The

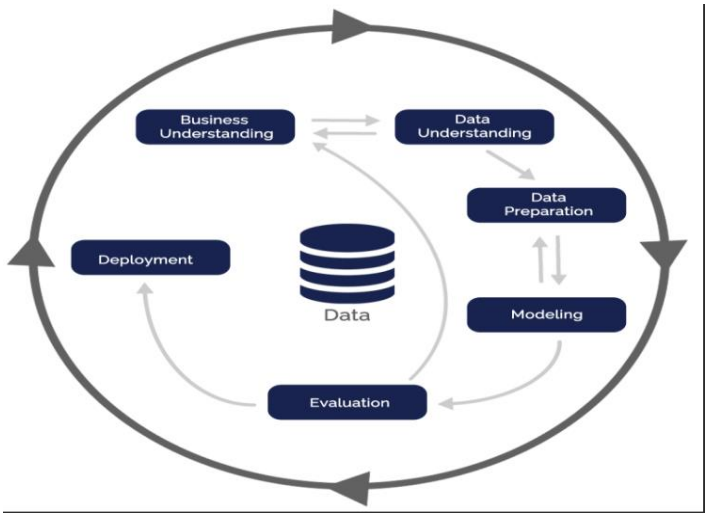
study was modeled with statistical methods, t-test and f-test statistics. In another study, 11.320 production data are grouped with data mining for optimal quality process based on raw material, sensitivity criteria, production network and production class characteristics in order to increase the quality of the production with 1 month data in a glasses manufacturing company. CHAID, ID3, C5.0 algorithms were used with decision trees classification methods. In another study in data mining, an optimism test was carried out in the product process. The optimal production process was determined by data mining according to the processes that 3 product groups pass through 5 stations according to the number of machines in the stations.In another study, in the spray production process, it was analyzed according to the data parameters of the production process according to certain parameters. In this study, classification algorithms and decision trees were used. In another study, classification has been made with data mining considering the capacities of old - new, 2 machines that do the same job in an enterprise. In this study, classification algorithms, decision trees and Naive Bayes algorithm were used. For another study, delay estimation was made with data mining techniques from the existing delays in the production process. In this study, classification algorithms and linear regression technique were used. For another study, 3 different automation systems compared to data mining techniques in a production firm. Of these automations, the first automation system is periodic operation, the second is automatic operation, and the third is temporary operation. Data mining was carried out according to serial number, production line, machine id, production start time and end time, and faulty product parameters. It was analyzed according to the CHRIS algorithm from data mining techniques and classification algorithms group. In another study, production cycle estimation was made in a production firm during the production process. In accordance with the network models, one of the operational research techniques, the earliest start, start-up and late completion times were estimated. In this study, prediction was made by statistical methods. Classification was made in SPSS Clementine program by statistical methods in order to improve product quality in a production enterprise, based on variables such as date, design, color, department, id, employee, quality, weight, height, size, cause of error, type of error, customer, barcode number, barcode date. In another study, in a company in the car manufacturing industry, the production time was analyzed by the rules of association. In this study, time series analysis was done by using Bayesian and markov chains. In a latest study in 2017, Standard classification and regression trees (CART) , Boosting tree (Boost), Random forest (RF) , Multivariate adaptive regression splines (MARS), Support vector machine (SVM) , K-nearest neighbor (KNN) ,Multiple regression (MR) , Neural networks (Neural Networks) methods were addressed to investigate which method gives the best results in prediction.

3. DATA AND METHODOLOGY

3.1. Methodology

In this study, textile goods, of manufactured by a textile company, whether being returned because of defects or not has been investigated by data mining and machine learning techniques. It has been used CRISP-DM methology. The CRISP-DM methodology provides a structured approach to planning a data mining project. CRISP-DM methodology can be seen in figure 1. It is a robust and well-proven methodology. It has a powerful practicality, flexibility and usefulness when using analytics to solve business issues. In this study, In this study, there are 250 different variables and 72959 lines of data on the production line. In order to perform a data mining process, it is firstly necessary to understand the data and determine the process. For this, CRISP-DM algorithm has been used. Modelling and classification algorithms are applied to estimate the production of faulty goods. In the model, a supervised learning model based machine learning methods have been used. The dimensions, loops and some statistical features of the data have been examined, and then it has been studied in the Python programming language. The feasibility of model and success rates have been evaluated with findings. First of all, it has been determined the firms' needings. Faulty products cause loss of productivity in the production process. Loss of productivity have caused extra cost to the business. At the same time, faulty products must be used in the company in another way. In addition, some products can be returned from suppliers according to the fault conditions after they are released. In this case, it has resulted in a second logistics cost and time loss. Since the firm is a mass production company, it should be treated quickly. At this point, a system that can detect faulty products is needed before the product is created in the production process. This study consists of the introduction of this system.

Figure 1: CRISP-DM Methodology for Data Mining



3.2. Data

As a starting point, it has been tried to develop an innovative, cheaper analysis method for the business. The data has been firstly sampled as a time series. After the completion of required analyses is held, machine learning algorithms have been employed to classify the data. In this study supervised machine algorithms have been used for classification. After the data acquisition part, classification algorithms have been tried for the dimensions. The first part of the analyses is to classify the production's situation at the last stage. For this purpose, logistics regression and k-nearest neighborhood algorithms have been used and their results have been compared. Logistic regression method, which is one of the nonlinear regression methods designed for binary dependent variable, is also included in the literature as logit regression. In a logistic regression model, if the dependent variable is expressed in two categories, it is defined as the "Binary Logistic Regression Model". If it is expressed with three or more categories, it is named as "Multiple Logistic Regression Model". In this study, there are two situations according to whether the products are defective and whether the products are returned or not. For this reason, binary logistic regression algorithm has been used. Binary logistic regression algorithm can be defined as follows. In logistic regression method, a logistic function has been used for switching between 0 and 1, such that in Eq. 1

$$\eta(x) = \frac{1}{1+e^{-x}} \quad (1)$$

of x is an n^th order polynomial. And X can be stated as in Eq. 2

$$x = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \quad (2)$$

The data are labeled as "faulty" or "not", "returned" or "not" in the python program. For this reason, since the results have been coded as "0" and "1" in the model, they have been made compatible with the logistic regression function. Also, the logistic regression algorithm can be shown in figure 2. In machine learning, the comfusion matrix is looked at in order to understand the accuracy of the model established. And while doing machine learning, the learning rates have been generally chosen as 67%. This means that the machine have reserved 67% of the data for learning in each time, while 33% of it hve allocated for testing. Studies conducted in recent years show that 80% learning rate have given better results. For this reason, both learning rates, 67% and then 80%, have been tried. In this study, it has also been seen that the prediction results have given better results as the learning rate increases. Later, another predition algorithm, k-nearest neighborhood algorithm, was used. Mathematical distance algorithms have been used when making K-nearest neighbor algorithms in machine learning. The most known distance algorithm is the Euclidean distance algorithm based on the "Euclidean distance formula". The most important thing in machine learning to consider when making classification is that the model can be built with the data. In the model to be installed, the features must be well distinguished. And with the applied model, the correct results can be achieved. The K-nearest neighbor algorithm provides a prediction of which group the data shown as a question mark will be in the figure 3. In KNN method, the test data are placed in cartesian coordinates. And then, odd numbered nearest neighbors are selected. In KNN method, the test data are placed in cartesian coordinates. And then, odd numbered nearest neighbors are selected. Also, Manhattan distance algorithm has been used for k-nearest neighbourhood for classfying the product.

Figure 2: The Concept of Logistic Regression Algorithm

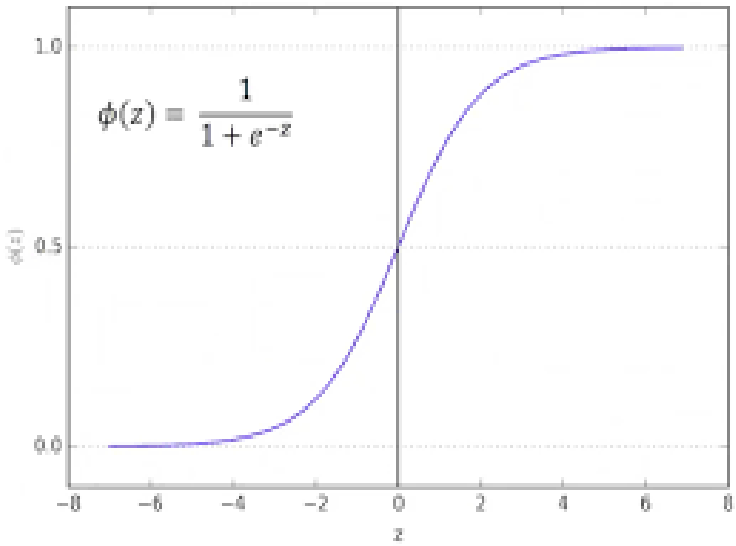
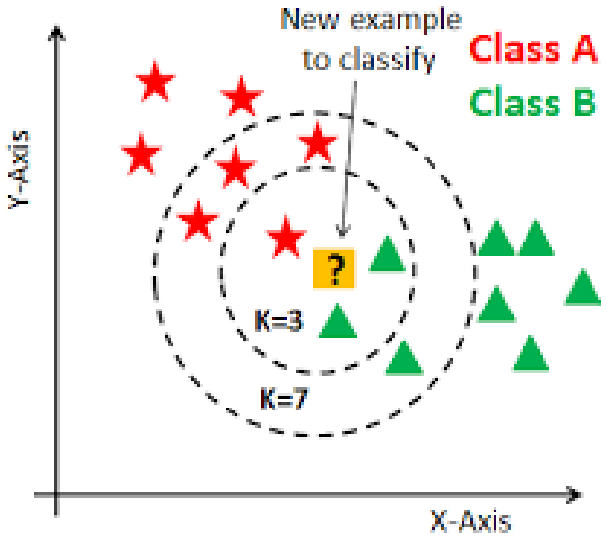


Figure 3: The concept of K-nearest Neighbourhood Algorithm



4. CONCLUSION

According to this study, it has been conducted a new approach to classify the products as faulty goods and refund goods. To perform the analyses, 250 variables have been collected and filtered. Afterwards, these time series have been examined using machine learning methods. The solution of machine learning prediction algorithms are very successful. It has been shown that in table 1. It has been tried to develop an innovative, cheaper analysis method for the business. This study's scope is detecting the faulty and normal goods in production line before the production process is completed. Thanks to machine learning algorithms, it can be predicted successfully. The data consists of the last 1-year production data of the enterprise. The data has been firstly sampled as a time series. It has been shown that logistic regression algorithm and KNN algorithm are succesful in classifying the products. And machine learning has given better results when the learning rate has increased. For further studies, the other machine learning algorithms can also be tried for checking the consistency. For further studies, boosting algorithms and simulation will also be tried for checking the consistency.

Table 1: The Results of the Machine Learning Algorithm

MACHINE LEARNING ALGORITHMS	LOGISTIC REGRESSION %67 LEARNING RATE	LOGISTIC REGRESSION %80 LEARNING RATE	KNN %67 LEARNING RATE (EUCLID DISTANCE)	KNN %80 LEARNING RATE (EUCLID DISTANCE)	KNN %67 LEARNING RATE (MANHATTAN DISTANCE)	KNN %80 LEARNING RATE (MANHATTAN DISTANCE)
Refund Detection	0,98	0,99	0,98	0,99	0,99	Nearly %100
Faulty Detection	0,97	0,98	0,96	0,97	0,97	0,98

REFERENCES

Bernstein, J. H. (2011). The Data-Information-Knowledge-Wisdom-Hierarchy and its Antithesis. NASKO, 68-75.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). Discovering data mining from concept to implemantation. New Jersey: Prentice Hall.

Chakrabarti, S. (2002). Mining the web: statistical analysis of hypertex and semi-structured data. Morgan Kaufmann.

Dasu, T. & Johnson, T. (2003). Exploratory data mining and data cleaning. John Wiley & Sons.

Frawley, J. W., Piatetsky-shapir, G. & Matheus, C. (1992). Knowledge discovery in databases: an overview. ai Magazine, 57-70.

Gürsakal, N. (2001). Sosyal bilimlerde araştırma yöntemleri. Bursa: VİPAŞ.

Han, J. & Kamber, M. (2006). Data mining: concepts and techniques. USA: Morgan Kaufmann Publishers, Elsevier.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction, 2nd ed., Springer-Verlag.

Jacobs, P. (1999). Data mining: what general managers need to know. Harvard Management Update, 8-10.

Kittler, R. & Wang, W. (1999). The emerging role for data mining. Solid State Technology, 45-58.

Liu, B. (2006) .Web data mining, Springer.

Özekes, S. & Çamurcu, Y. (2003). Veri madenciğinde karar ağaçları yöntemi uygulaması. Bilgi Teknolojileri Kongresi II. Denizli: Pamukkale Üniversitesi.

Özkan, Y. (2016). Veri madenciliği yöntemleri. İstanbul: Papatya Bilim.

Sağın, A. (2018). Veri madenciliği algoritmalari ile birliktelik kurallarinin belirlenmesi: perakende sektöründe bir uygulama. İstanbul, İstanbul Ticaret Üniversitesi.

Silahtaroğlu, G. (2016). Veri madenciliği: kavram ve algoritmaları. İstanbul: Papatya Bilim.

Witten, I.H. & Frank, E. (2005) Data mining: practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, 2nd ed.