Düzce University Journal of Science & Technology

Research Article

Implementation of Decision Support System with Data Mining Methods in the Quality Control Process of the Automotive Sector

Hikmet CANLI [a], Sinan TOKLU [a,*]

[a] *Department of Computer Engineering, Engineering Faculty, Düzce University, Düzce, TURKEY*
** Sorumlu yazarın e-posta adresi: sinan.toklu@duzce.edu.tr*

ABSTRACT

Today, the automotive sector is the "key" sector for developed and even developing countries. A strong automotive sector is striking as one of the common features of industrialized countries. Production in this sector consists of many processes. One of the most important of these processes is quality control. The measurement data in this area is very large and as the volume of data increases, the rate that people understand is reduced. Variations are the enemy of quality. There are many variations in the area of quality control. In this study, a decision support system is applied in the quality control process with classification algorithms which are data mining methods. C4.5, Naive Bayes, SMO and Random Forest algorithms are run on data set collected from production. These algorithms are used to measure the quality and accuracy of the product without completing the operations during production. Algorithms have been cost-reduced by determining that the product is faulty before operations are completed. The algorithm C4.5 has been the best performing algorithm. In addition, these algorithms make quality analysis very fast and easy. Thanks to this work, the cost of labor and materials has been reduced in the production company.

Keywords: *Data mining, Data visualization, Decision support, Production, Quality control*

Otomotiv Sektörünün Kalite Kontrol Sürecinde Veri Madenciliği Yöntemleri ile Karar Destek Sistemi Uygulaması

ÖZET

Günümüzde otomotiv sektörü, gelişmiş ve hatta gelişmekte olan ülkeler için “anahtar” sektör rolündedir. Güçlü bir otomotiv sektörü, sanayileşmiş ülkelerin ortak özelliklerinden biri olarak gözümüze çarpmaktadır. Bu sektörde üretim birçok süreçten oluşmaktadır. Bu süreçlerin en önemli olanlarından biri de kalite kontroldür. Bu alanda ölçüm verileri çok fazladır ve verilerin hacmi arttıkça insanların anladığı oran azalmaktadır. Varyasyonlar kalitenin düşmanıdır ve her şeyde varyasyon bulunmaktadır. Bu çalışmada veri madenciliği yöntemlerinden olan sınıflandırma algoritmaları ile kalite kontrol sürecinde bir karar destek sistemi uygulaması yapılmıştır. C4.5, Naive Bayes, SMO ve Random Forest algoritmaları, üretimden toplanan veri seti üzerinde çalıştırılmaktadır. Bu algoritmalar, üretim sırasında işlemler tamamlanmadan ürünün kalitesini ve doğruluğunu ölçmek için kullanılır. Algoritmalar, işlem tamamlanmadan önce ürünün arızalı olduğunu belirleyerek maliyet düşürülmektedir.

Geliş: 28/05/2018, Düzeltme: 18/06/2018, Kabul: 18/09/2018

Algoritma C4.5 en iyi performans gösteren algoritma olmuştur. Ek olarak, bu algoritmalar kalite analizini çok hızlı ve kolay hale getirmektedir. Bu çalışma sayesinde, firmalarda işçilik ve malzeme maliyeti azaltılmıştır.

Anahtar Kelimeler: *Kalite Kontrol, Karar Destek, Üretim, Veri Görselleştirme, Veri Madenciliği*

I. INTRODUCTION

The purpose of quality control is to develop, implement and efficiently execute plans and programs that would ensure the production of products that would meet both, the consumers' requests and the general objectives of the enterprise at the most economical level possible. [1]. A product is undergone many processes during the production phase. Each of these processes has specific control mechanisms. The control mechanisms contain nominal measurement values for the related process of the part to be produced. A control mechanism may require hundreds of measurement values for a single process. Because of the large number of products in mass production enterprises in the automotive sector, the measured values constitute very large data sets. As the datasets grow, it becomes difficult to analyze them and it is time consuming. Furthermore, analyzes made through basic statistical and reasoning methods give us the results after production has occurred and therefore, no specific rules and predictions can be generated. Quality control is one of the application areas of data mining. By means of the quality control methods applied on the data obtained from the database, it is investigated whether the quality level meets the desired standards. If the quality level does not meet the desired standards, various measures are taken to bring the quality to the desired level. Data mining is the process of storing large amounts of information and organizing and sorting useful information. Another definition; Discovery of information in databases is often referred to as Data Mining, which aims to discover useful information from a large volume of data collections [2]. Data mining has a widespread use today. Today, all businesses want to predict the behavior of their customers. Data mining is a technique that can be used for this purpose. It is used in fields such as banking, marketing, insurance, telecommunication, stock exchange, medicine, industry, science and engineering [3]. Data Mining is a process that explores patterns and relationships in data with the use of many analysis tools and uses them to make valid estimates [4].

The goal of this study is to make the quality control process better understood faster by using appropriate data mining methods. In addition, instead of examining the entire process, samples in a quality that would represent the entire process are collected at specific time intervals. The goal is to discover the rules and relationships hidden within these data that can be used to make predictions about the future. Thus, the classification model that best estimates the faulty parts in the quality control processes of the automotive sector through the data mining methods based on the classification of data sets would be determined.

II. RELATED WORK

The use of data mining in quality control ensures quick and easy access to the database, thus saving time and money. If we look at the main data mining studies conducted in quality control processes through a literature search: A study was conducted by Deng and Wang based on the time series data mining methodology, proposing a new general analysis framework for water-quality time series data.

This study consists of two parts: application components and the common tasks of time series data mining within the data on water quality. In the first part, it is suggested to break down the time series into two-dimensional normal clouds and to calculate the similarities in granular level. In the second part, studies are conducted on similarity search to the water-quality time series sampledata through similarity matrix, anomaly detection and model finding. They analysed a case study made on weekly time-series data of Dissolved Oxygen collected from five monitoring stations in China's upper reaches of the Yangtze River. Experimental results showed that the proposed framework of analysis is a feasible and effective method for discovering hidden and valuable information from the historical time series of water quality data [5].

Baykasoğlu explained the application of data mining he conducted in cement sector. Strength of compression is the most important cement property, which is the main parameter for quality control. The standard "28 day compression strength test" is commonly used to determine compression strength. This test involves the determination of compressive strength experimentally by keeping each batch of samples taken during the cement production process for 28 days. Nevertheless, it is a long time for the industry to wait 28 days for the experimental results of cement mortar resistance. For this reason, a faster determination method of compressive strength is a necessity for the cement industry and deserves the attention of researchers. In order to estimate the compressive strength of Portland composite cement in the study, a new data mining method known as genetic equation programming and artificial neural networks, known as data mining and regression analysis, were used. The performances of these methods were compared. It was observed that methods based on artificial intelligence yielded better results. In particular, gene equation programming yielded better results than other methods [6].

Glawar et al. describe, in a study conducted, the application of data mining to quality focused maintenance planning supported by. According to the study; proper maintenance precautions taken at the right time are the important factors in ensuring the availability of the plant, product quality and process efficiency in modern manufacturing systems. The maintenance strategies established are often insufficient to combine these strongly related aspects. These strategies cannot make predictions in a holistic way, thus, unnecessary high maintenance works lead to loss of time and failure in quality and accessibility. To achieve a holistic and predictive approach in maintenance planning, a method for compiling and correlating various data consistent with the "cause-and-effect" consistency is proposed. By breaking down the production facilities at the component level, a basis is established for linking condition monitoring data, data wear, quality and production data using data mining methods. This framework provides for the determination of critical maintenance conditions, prediction of error moments and quality deviations [7].

Harding et al. conducted a detailed research on data mining in manufacturing. This research made analyses in data mining, production engineering applications, especially production processes, operations, failure detection, maintenance, decision support and product quality improvement. In their research, instead of discussing the field of data mining in general, they tried to show that data mining was related to the manufacturing industry. Numerous applications in data mining were examined in this research. In recent years, there has been a significant increase in the number of publications in some manufacturing fields such as failure detection, quality improvement, production systems and engineering design. Other fields are relatively less important. It was also suggested that a more general process for data cleaning was required to enable data mining to progress in manufacturing industry [8].

In a study by Kamal, data mining approaches applied to improve the quality control process in production were described. As the volume of the data increases, inevitably the rate that humans can understand shrinks. Diversity and probability is the enemy of quality. Product quality should be the focal point for any operation. By using appropriate data mining tools and the concept of statistical reasoning, managers and employees provided a better understanding of the processes. Kamal explained that besides data mining, the SPC had also played an important role in understanding the variations in quality. As a result, data mining concepts and techniques in quality control processes provide an overview of SPC design and performance to seek and improve patterns in the data [9].

In a study by Khan et al., methods of statistical data mining are described for efficient quality control in production. The widespread use of machines, flexible / reconfigurable production and the transition to a fully automated factory require the intelligent use of information recorded during the production process. Modern production processes, at different stages of the process, such as sensor measurements, machine readings, etc. generate terabytes of information with applications and the main contribution of this large data set is the different quality control processes. In the study, a method for obtaining valuable information from manufacturing data is provided. Based on the proposed method, the Statistical Genetic Algorithm is based on a comparison of the probabilities and the prolongation of probability principles as a performance function. It is concluded in the study that the QC7 (UT) resources can be improved by about 98%, allowing an error rate of around 0.0095. However, according to industrial quality standards, this figure should not exceed 0,00025. They also considered different methods to meet these quality standards [10].

In a study by Chen et al., use of data mining for quality control design in the manufacturing industry was investigated. The study compares the accuracy of two data mining classification analyzes (Decision Trees and Bayesian Algorithm) used to explore the underlying causes of inconsistency in the manufacturing process of semiconductor plants. Four characteristics were examined in the study; Human, Machine, Material and Management. The aim is to identify the machine that will give the best result in the shortest time. According to the results obtained, the decision tree algorithm is more effective and suitable than the Bayes algorithm for analyzing the quality problems in the semiconductor packaging industry [11].

In the study conducted by Ferreiro, data mining was investigated in the quality control process for detecting burrs in drilling in the aviation industry. A simple model for experimental design and data mining techniques, especially for selection of variables and machine learning algorithms, to detect burrs during drilling was developed. The model was based on the internal signal and certain parameters of the process conditions, making it easier to apply and no external sensors were used. A monitoring system was established to detect online when burrs form during drilling. Secondly, it was defined in which parameters of the drilling process the burr was formed. Regarding the results, almost all advanced models provide a higher accuracy than the current mathematical model. Furthermore, for the final model based on the Naive Bayes algorithm, the accuracy is 95% and the standard deviation equals to 0, which means that it is a very stable model. At this point, it is explained, in most of the cases where the model is poorly predicted, the burr is very close to the overhead limits (127 lm), and which wind rate it makes harder to determine. Another thing to consider is, whether the drilled hole has burrs or not, the estimated error does not have the same significance. By developing a model for the detection of these cases, the number of inspections of the burrs is significantly reduced [12].

The overall objective in these studies is to reduce the quality control process in time, to facilitate the determination of quality degrading situations and the ability to make decisions. These surveys clearly

show that data mining has a very important place in the quality control process. It is also verified by this study that classification algorithms are used for the quality control field, which is the general opinion. As a result of these studies, it was decided to use algorithms based on data mining techniques in the works to be conducted. In this study, studies were made on C4.5, Random Forest, SMO and Bayes algorithms.

III. The Proposed Methods

The studied data set, are the measurements taken from a bench in the assembly line of an automotive company in Turkey. The product made on the bench is a front door hinge of a truck. The dataset consists of 17046 records and 19 variables. The names, explanations and data types of these 19 variables are shown in Table 1.

Table 1. *Variables of production data set, display formats and types.*

Attribute	Explanation	Type
aci	Angle of rotation of the hinge	Numeric
pim_kuvveti	Force applied when pin is pushed	Numeric
pul_kuvveti	Force applied when stamp is pushed	Numeric
sıklık	Hinge Frequency	Numeric
r_bosluk	Radial gap	Numeric
e_bosluk	Axial gap	Numeric
mesafe	Distance	Numeric
agirlik	Semi product weight	Numeric
kodu	Product code	Text
adi	Product name	Text
operkodu	Operation Code	Numeric
operadi	Operation name	Text
p_kodu	Employee code	Text
sirano	Operation sequence no	Numeric
tarih	Operation date	Date
saat	Operation hour	Date
t_kodu	bench code	Text
sicaklik	Ambient temperature	Numeric
sonuc	Product Okey / Reject status	Binary

The main critical inspections for this hinge are; axial gap, radial gap, hinge density, angle, distance, washer pressing force, pin pressing force. The accuracy of the data set was tested 100%.

The data were recorded with the help of the servo-motor, PLC and sensors on the bench and recorded in the database. Preprocessing was performed on the recorded data and the data set was formed. Various graphic are used to analyse the changes in the numerical variables in the data set. The

histogram graphic are also one of these. Histograms of the characteristics of angle, density, pinning and stamping are shown in Figure 1.

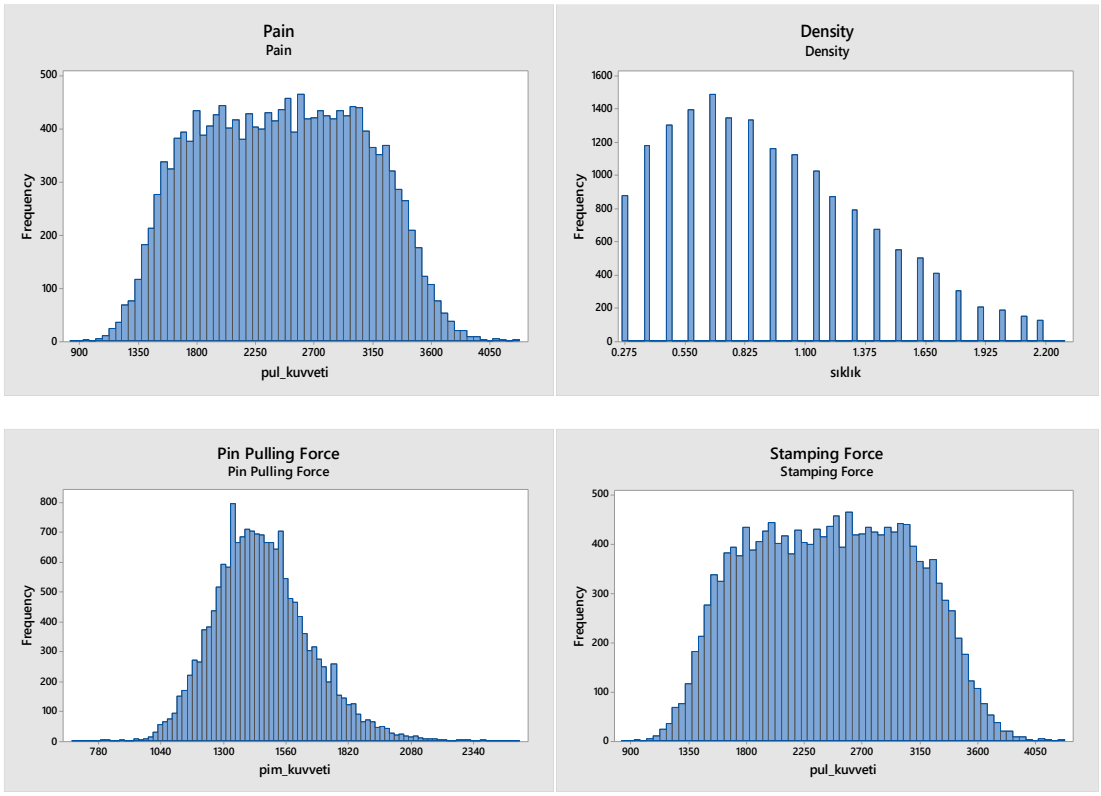

Figure 1. *Angle, Density, Pin strokes and Stamp strokes histogram.*

The distribution graphic helps us find the possible associations between values and outliers in data sets. A distribution chart is a great way to visualize the correlation of two or more calculations at the same time. Particularly suitable for comparing the range and distribution of numerical data groups. Angle - Density distribution graphic is given in Figure 2. When we examine the graphic, it is seen that there is an inversely proportional relationship between angle and density. It is observed that as the angle increases, the density decreases.

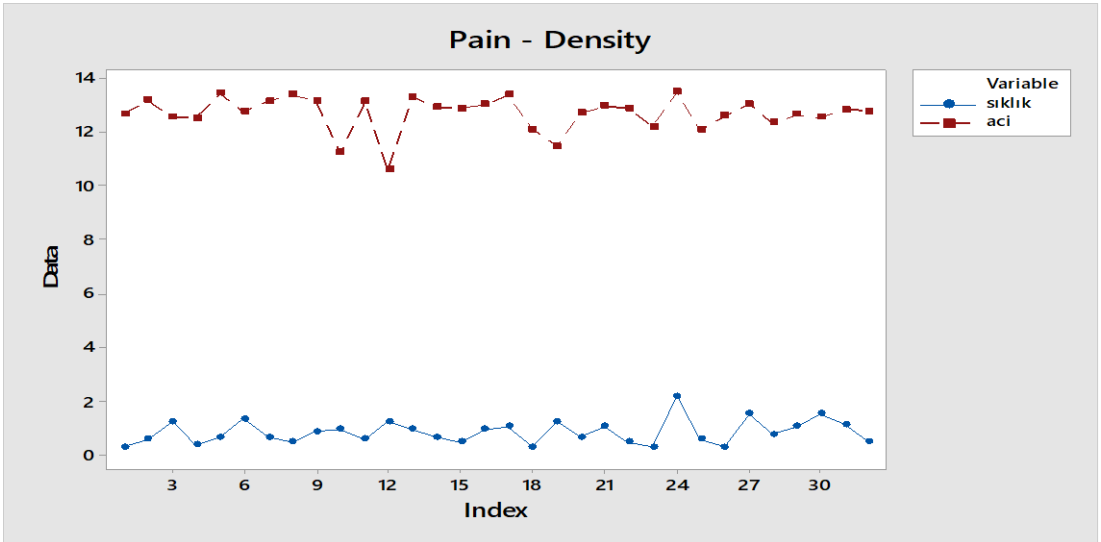

Figure 2. *Angle - Density Distribution Graphic*

Based on this data set, classification models were developed to predict whether automobile parts in the automotive industry would be faulty or not. These classification models are C4.5, SMO, Random Forest and Naive Bayes algorithms.

These algorithms were applied to the generated data set and a decisilon was treid to be made which model gave better results. During this decision-making process, the results were evaluated in two different methods, using hold-out and cross-validation and performance evaluation analysis methods. For the first method, hold-out, the test and training data set with 40% -60%, 25% -75%, 20% -80% discrimination rates were evaluated. 5-fold and 10-fold cross-validation was performed with cross validation as the second evaluation method. In such doing, the accuracy of the result we had found using two different analysis methods was proven. This modeling process was accomplished by means of Weka, R and Minitab programs. Weka is a modular program that holds almost all of the data mining algorithms and provides the use of these algorithms, and possesses similar features such as data visualization, data analysis, and business intelligence applications. In this study, the application of algorithms on the data set, the creation of rules and decision trees were performed on this program. R is a statistical and analysis program that statisticians and mathematicians prefer intensively. Analysis on the dataset used in the study was carried out with R packages. Another program used in this study is Minitab. Minitab is also a statistical program like R. A data visualization application was also utilised so that the quality of Minitab could be better interpreted.

C4.5 is a decision tree algorithm. It is an upper level of the ID3 algorithm. There are some deficiencies in the ID3 algorithm, which are solved by the C4.5 algorithm. The biggest difference between C4.5 and ID3 is the normalization. In addition, pruning is performed different to that of ID3 decision tree. C4.5 is carried out in two process steps. The first one is the process of building the tree and the other is the pruning process [13]. The Naive Bayes Classifier is named after the English mathematician Thomas Bayes who lived in the 17th century. Naive Bayes classifier is a probabilistic classifier based on Bayes theorem with independent assumptions. Despite the lean design and seemingly simplified assumptions, the Naive Bayes classifier gives much better results than expected in real world situations [14]. Squential Minimal Optimisation (SMO); SMO is essentially an algorithm that uses support vectors. It applies the SMO Algorithm to train support vector classifiers using multiterminated kernel. This application globally replaces all lost values and converts nominal attributes to binary ones. It also normalizes all attributes to predefined values [15]. In the RF method, the decision trees that make up the decision forest are constructed from different samples selected from the original data set by the bootstrap method. In each decision tree, few variables randomly selected from all variables in the data set are used. Every tree votes for a class, and the forest classifier makes a final guess for a class by collecting the votes given by all the trees.

A. MODEL APPLICATIONS

Holt-out and cross-validation performance evaluation methods were used when creating models. In the hold-out method, 60% -40%, 75% -25%, 80% -20% of the training and test data set are sequentially differentiated. In k-fold cross validation, 5-fold cross validation and 10-fold cross validation were used.

Table 2. *Modelling Summary*

Target Variable	Result (E/H-Yes/No)
Performance Evaluation and Model Selection Method	5-fold cross-validation and 10-fold cross-validation. %60-%40, %75-%25, %80-%20 hold-out.
Transactions with WEKA	The data set. Uploading program. Application of models.
Transactions with R	Performing data analysis.
Transactions with Minitab	Drawing of histograms and graphics.

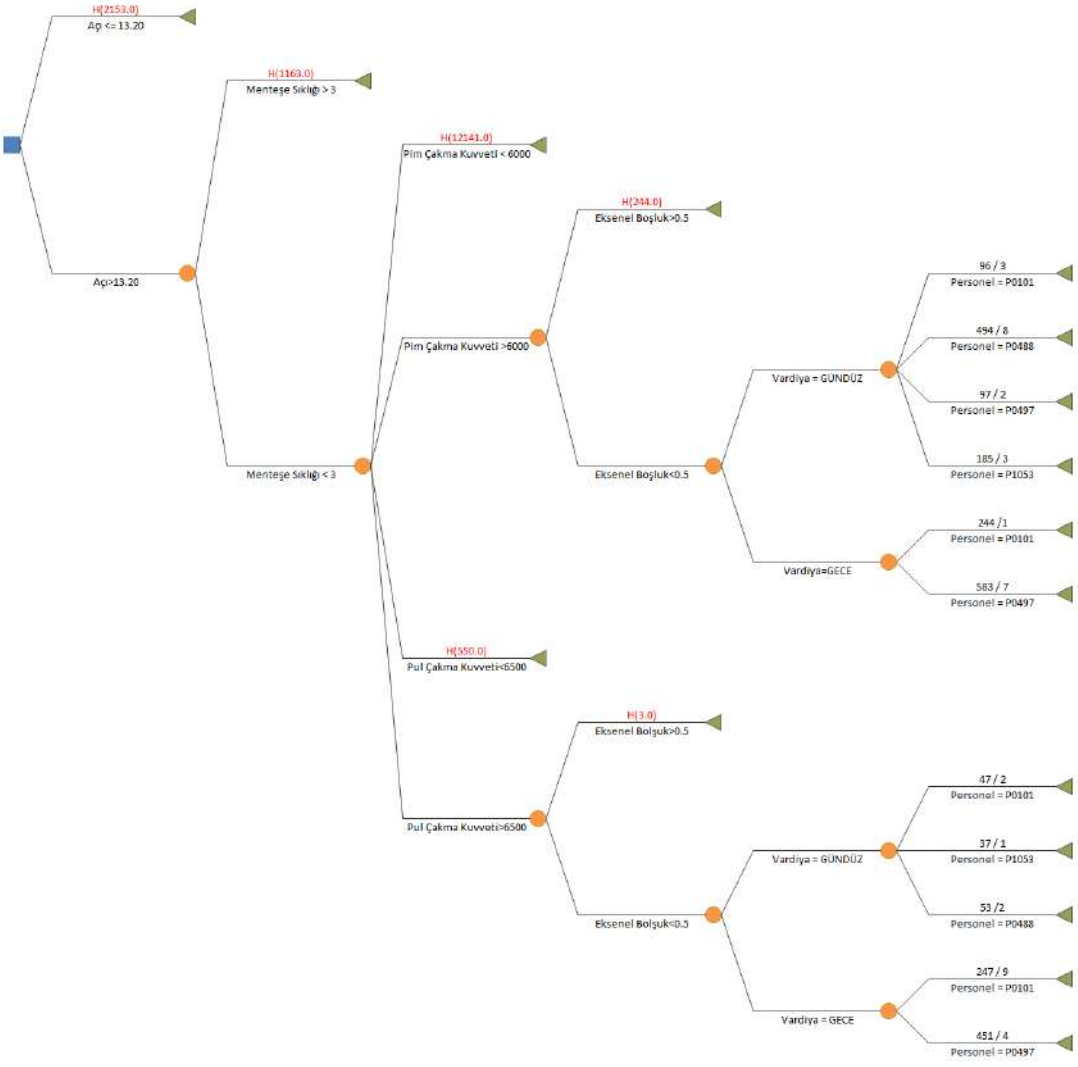

Figure 3. *C4.5 Algorithm Decision Tree*

As shown in Figure 3, after applying the C4.5 algorithm to the dataset, a decision tree was formed according to the rules. If we happen to explain these rules;

Rule 1: If Angle is equal to or smaller than "13.20", the part is defective. Result is "0".

Rule 2: If Angle is greater than "13.20" AND Hinge Frequency is greater than "3", Result is "0".

Rule 3: If Angle is great "13.20" AND Hinge Frequency is greater than "3" AND Pin Pressing Force is Smaller than "6000"; Result is "0".

Rule 4: If Angle is greater than "13.20" AND Hinge Frequency is greater than "3" AND Washer Pressing Force is smaller than "6500"; Result is "0".

Rule 5: If Angle is greater than "13.20" AND Hinge Frequency is greater than "3" and Pin Pressing Force is greater than "6000" AND Axial gap is greater than "0.5"; Result is "0".

Rule 6: If Angle is greater than "13.20" AND Hinge Frequency is greater than "3" AND Washer Pressing Force is greater than "6500" AND Axial Gap is greater than "0.5"; Result is "0".

Rule 7: If Angle is greater than "13.20" AND Hinge Frequency is greater than "3" AND Pin Pressing Force is greater than "6000" AND Axial gap is smaller than "0.5" AND Shift equals "DAY"; it contains accuracy rate of P0101 = (96 / 3), P0488 = (494 / 8), P0497= (97 / 2), P01053 = (185 / 3).

Rule 8: If the Angle is greater than "13.20" AND Hinge Frequency is greater than "3" AND Pin Pressing Force is greater than "6000" AND Axial gap is smaller than "0.5" AND Shift equals "NIGHT"; it contains accuracy rate of P0101 = (244 / 1), P0497 = (583 / 7).

Rule 9: If Angle is greater than "13.20" AND Density is greater than "3" AND Washer Pressing Force is greater than "6500" AND Axial gap is smaller than "0.5" AND Shift equals "DAY"; it contains accuracy rate of P0101 = (47 / 2), P01053 = (37 / 1), P0488 = (53 / 2).

Rule 10: If Angle is greater than "13.20" AND Hinge Frequency is greater than "3" AND Washer Pressing Force is greater than "6500" AND Axial gap is smaller than "0.5" AND Shift equals "NIGHT"; it contains accuracy rate of P0101 = (247 / 9), P0497 = (541 / 4).

B. COMPARISON OF CROSS VALIDATION PERFORMANCE OF MODELS

In cross-validation data mining studies, it is used to test the success of the applied method. The dataset is divided into training and test clusters. For the application, C4.5, Naïve Bayes, SMO and Random Forest models were selected. In the application, the data set was subjected to 5-fold and 10-fold cross-validation performance testing. In Table 3, the results obtained by applying 5-fold cross-validation and 10-fold cross-validation to the dataset are shown. The "Accuracy" and "Runtime" results are important in comparing models. According to these results, the C4.5 model is the model that provides the best performance in terms of accuracy. The accuracy of 5-fold cross-validation is 0.8576. When 10-fold cross-validation was applied to this model, there was a change of 0.0005 in accuracy. This change is too small. This shows us that the C4.5 gives a very stable result. The C4.5 model is the best performing model with a "Runtime" rating of 0.042. The closest result to the C4.5 model is Random Forest. The accuracy of this model is very high. However, the attribution of "Runtime" is very long. Because of this, it is behind C4.5. Naïve Bayes is in third place, the accuracy rate is 0.79, and the time performance of this model is also good. The SMO model did not perform well. It has an accuracy rate of about 0.69 and a very high Runtime value. Compared to other models, it is far behind. In addition, this accuracy is not enough.

Table 3. *Results of analyses made in 5-fold and 10-fold cross-validation performance*

	5-fold cross-validation				10-fold cross-validation			
FEATURE	C4.5	N.Bayes	SMO	Rand.F	C4.5	N.Bayes	SMO	Rand.F
Accuracy	0.8576	0.7957	0.6938	0.8546	0.8581	0.7989	0.6938	0.8564
Error	0.1424	0.2043	0.3062	0.1454	0.1419	0.2011	0.3062	0.1436
Precision	0.871	0.805	0.738	0.862	0.873	0.808	0.738	0.862
Recall	0.858	0.796	0.694	0.855	0.858	0.799	0.694	0.855
F-measure	0.848	0.778	0.604	0.847	0.848	0.782	0.604	0.847
Roc Area	0.828	0.761	0.549	0.891	0.827	0.762	0.549	0.891
Run time	0.42	0.8	50.31	10.73	5.59	0.08	56.34	12.05

The precision is the ratio of the number of pertinent items found over the total number of items found

$$Precession = \frac{|\{Relevant\ items\} \cap \{Retrived\ items\}|}{|\{Retrieved\ items\}|} \quad (1)$$

The recall is the ratio of the number of pertinent items found over the total number of relevant items.

$$Recall = \frac{|\{Relevant\ items\} \cap \{Retrived\ items\}|}{|\{Relevant\ items\}|} \quad (2)$$

F-measure (or F1 score) is the harmonic mean of precision and recall.

$$F = \frac{2(precision\ x\ recall)}{precesion + recall} \quad (3)$$

C. HOLD-OUT PERFORMANCE COMPARISON OF MODELS

When generated models to the data set, they were also compared in Hold-out method as well. The results of the hold-out performance analysis is given in Table 4. The model was applied to the data set at the rate of 40% - 60%, 25% - 75%, 20-80% (test - training) respectively. The C4.5 model is also the best end result model for hold-out performance comparison. This shows the correctness of our performance tests. The "Runtime" values of all models gave better results because the hold out test performed less analysis and processing than cross verification. C4.5, Naïve Bayes and Random Forest ended up in approximately the same accuracy rate when compared to cross-validation in terms of "accuracy". However, the SMO model yielded a better result than that of the cross-validation performance test. Nevertheless, the SMO model remains below the standards in both methods. C4.5 was the model that gave the best result in both methods.

Table 4. *Results of analysis made in Hold-Out performance*

	Accuracy			Error		
Test - Train	40-60	25-75	20-80	40-60	25-75	20-80
C4.5	0.8580	0.8556	0.8524	0.142	0.1444	0.1476
N.BAYES	0.7855	0.7793	0.7688	0.2145	0.2207	0.2312
SMO	0.7224	0.7169	0.7160	0.2776	0.2831	0.284
RAND. F	0.8580	0.8549	0.8503	0.142	0.1451	0.1497

	Runtime		
Test - Train	40-60	25-75	20-80
C4.5	0.39	0.08	0.38
N.BAYES	1.11	0.2	0.48
SMO	0.66	0.76	0.11
RAND. F	1.38	1.56	0.28

IV. Discussion And Conclusion

In the proposed method, C4.5, Random Forest, Sequential Minimal Optimization (SMO) and Naïve Bayes algorithms were used. Models were evaluated in hold-out and cross-validation performance methods. In the hold-out method, the dataset was again the best performing model with a precision of C4.5 to about 86%, with 20% -80%, 25% -75%, 40% -60% dedicated to the test and training data set respectively. Random Forest was the second, Naive Bayes was the third, and SMO was the fourth best performing algorithm. C4.5 and Random Forest models showed no significant change in performance when cross-validation results were compared to the hold out results. While the Naive Bayes model gave worse performance in the hold out method, the SMO model performed better.

Models were tested on data from production collected for 6 months. The current situation was identified. The process was under control and improved. Based on the evidence, not on memorization, it was decided on the basis of monitoring and measurement. The results of continuous improvement were revealed.

As a result, not all known methods in the literature were applied to this area before. In this area, actual production data was accumulated for 6 months and an original database was created. Important decision support and data mining methods existed in the literature on this database and the methods giving the most accurate result were applied successively and a success rate of 90% was obtained. In this respect, errors that can occur in this field of the industry, ie in real production, were reduced to a minimum level. Since the data in the generated database is generated from new and real-time systems in the system, the error rates reduced to the minimum level can be obtained more precisely thanks to the new databases to be created later and by combining different applications. If the applied methods are known methods, the error rate is reduced to the minimum level by applying different methods in different stages and choosing the methods giving the most accurate result at each stage and combining these methods in succession.

V. REFERENCES

[1] M. Yılmaz, "Applicability of the evolution of quality management systems and of total quality management in the banknote printing general directorate", Master thesis, Department of Business Administration, Gazi University, Ankara, Turkey, 2003.

[2] A. S. Koyuncugil, and N. Özgülbaş, "Data Mining: Use and applications in medical and health services", Journal of Information Technologies, vol. 2, no. 2, pp. 22-32, 2009.

[3] M. S. Başarslan, "Customer loss analysis in telecommunication sector", Master thesis, Department of Computer Engineering, Düzce University, Düzce, Turkey, 2017.

[4] (Anonymous). (2017, 20 September). Introduction to Data Mining and Knowledge Discovery [Online]. Access: http://www.twocrows.com/intro-dm.pdf.

[5] W. Deng, and G. Weng "A novel water quality data analysis framework based on time-series data mining," *Journal of Environmental Management*, vol. 196, pp. 365–375, 2017.

[6] A. Baykasoğlu," Data mining and an application in the cement sector", *Akademik Bilişim*, pp. 1-14, 2005.

[7] G. Robert, "A Holistic Approach for quality oriented mainctance planning supported by data mining methods.", *Procedia CIRP*, vol. 57, pp. 259-264, 2016.

[8] J. A. Harding, M. Shahbaz, Srinivas and A. Kusiak "Data Mining in Manufacturing: A Review", *Journal of Manufacturing Science and Engineering*, vol. 128, no. 4, pp. 969–976, 2005.

[9] A. M. M. Kamal, "A Data Mining Approach for Improving Manufacturing Processes Quality Control", Next Generation Information Technology, Gyeongju, South Korea, 2011.

[10] A.R. Khan, H. Schiøler, and T. Knudsen "Statistical data mining for efficient quality control in manufacturing", Emerging Technologies & Factory Automation (ETFA), Luxembourg, Luxembourg, 2015.

[11] R. S. Chen, Y.C. Chen and C.C. Chen, "Using data mining technology to deign a quality control system for manufacturing industry", Advances in Communications, Computers, Systems, Circuits and Devices, Puerto De La Cruz, Tenerife, 2015.

[12] S. Ferreiro, B. Sierra, I. Irigoien and E. Gorritxateg, "Data mining for quality control: Burr detection in the drilling process.", *Computers & Industrial Engineering*, vol. 60, no. 4, pp. 801-810, 2011.

[13] A. Gümüşçü, R. Taşaltın, İ. B. Aydilek, "C4.5 pruning with genetic algorithm in decision trees", *Dicle University Graduate School of Natural and Applied Sciences*, pp. 77-80, 2016.

[14] E. Uzun. (2014, 3 September). Naive bayes classifier [Online].Access:https://www.e-adys.com/makine_ogrenme si/naive-bayes-classifier/.

[15] E. Ardıl, “Software Error Estimation with Flexible Calculation Approach”, Master thesis, Department of Computer Engineering, University of Trakya, Tekirdağ, Turkey, 2009.