International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:11, No:7, 2017

Knowledge Discovery and Data Mining Techniques in Textile Industry

Filiz Ersoz, Taner Ersoz, Erkin Guler

Abstract**—**This paper addresses the issues and technique for textile industry using data mining techniques. Data mining has been applied to the stitching of garments products that were obtained from a textile company. Data mining techniques were applied to the data obtained from the CHAID algorithm, CART algorithm, Regression Analysis and, Artificial Neural Networks. Classification technique based analyses were used while data mining and decision model about the production per person and variables affecting about production were found by this method. In the study, the results show that as the daily working time increases, the production per person also decreases. In addition, the relationship between total daily working and production per person shows a negative result and the production per person show the highest and negative relationship.

Keywords**—**Data mining, textile production, decision trees, classification.

I. Introduction

DATA mining under advanced data analysis systems; Clustering, classification, association analysis and OLAP (Online Analytical Processing). Uses techniques such as. The advanced applications of data mining are visual data mining, web mining and text mining [1] and the thing which is the removal of potentially useless information that is unclear, unclear, unknown beforehand. Data mining is an interdisciplinary study that uses statistics, database technology, machine learning, artificial intelligence and of visualizing [2]. This, too, Clustering, data summarization, analysis of changes, determination of deviations. On the other hand data mining is the semi-automatic discovery of patterns, relations, changes, irregularities, rules, and statistically significant structures in data. Basically, data mining is concerned with the use of patterns or arrangement between data sets, analysis of data, and software techniques. The computer is responsible for determining the relationships, rules, and properties between the data. The goal is to be able to detect previously unrecognized data patterns.

The main reason for the great interest that data mining has made in the information industry in recent years is the availability of large amounts of data and the need to translate this data into useful information. The information gained can be used in a range of applications ranging from business management, production control and market analysis to engineering design and information discovery [3].

Filiz Ersoz is with the Department of Industrial Engineering at Karabük University, Turkey (phone: 370-433-20-21, e-mail: fersoz@karabuk.edu.tr).

Taner Ersoz and Erkin Guler are with the Department of Actuarial and Risk Management, Karabük University, Turkey (e-mail: tanerersoz@karabuk.edu.tr, erkinguler@icloud.com).

It may be possible to see data mining as a set of statistical methods. However, data mining differs from traditional statistics in several ways. The goal in data mining is to extract qualitative models that can be easily translated into logical rules or visual presentations. In this context, data mining is human-centered and sometimes human, computer interface is merged.

II. The Methodology of Data MInIng

Cross Industry Standard Processing for Data Mining (CRISP-DM); SPSS Inc. It is an open data mining methodology developed by OHRA, NCR, Daimler Chrysler and Teradata. The CRISP-DM methodology is supported by two hundred data mining organizations. The aim of methodology; Data mining projects use a certain method regardless of the program used. A data mining project based on the CRISP-DM methodology;

- Understanding Business Needs
- Understanding of Data
- Preparing of Data
- Modelling
- Evaluation of the Model
- Application of the Model
- Updating Model

Application is seen in the literature, the 2nd and 3rd stages of the data is understood to be the temporal and preparing a project comprises 75% to 80%. This gives us access to the database and shows how important the quality of the data is. The purpose of going for data mining and knowledge discovery through large and complex databases is to ease these difficulties and making data decisive and helpful in decision processes for management. The steps of Knowledge Discovery in Database (KDD) are given in Fig. 1.

III. The Steps of the KDD Process in the Study

A. Selection

Selection is the first step in which the most important data fields are opted for evaluation subject to determined selection criteria. Even though a data cluster comprises very similar data points, more important/representative ones should be filtered to avoid high complexity.

B. Data Pre-Processing

In order to achieve higher data quality and reliability in data mining process, data pre-processing, which is another key stage of data mining, is conducted. Data pre-processing operations consist of data cleaning, aggregation,

transformation and reduction. Data cleaning is a set of operations to conduct certain functionalities such as filling missing data, identifying outliers, and removing the incoherency. Also, these steps are applied to remove various noisy data, which remain out of range and on the far extremes. In this study, missing data and outliers are identified by using Audit processor in SPSS, and then, all deficient values, outliers and incoherent data are removed. Additionally, cleaning and normalization processes are performed to obtain a mean distribution.

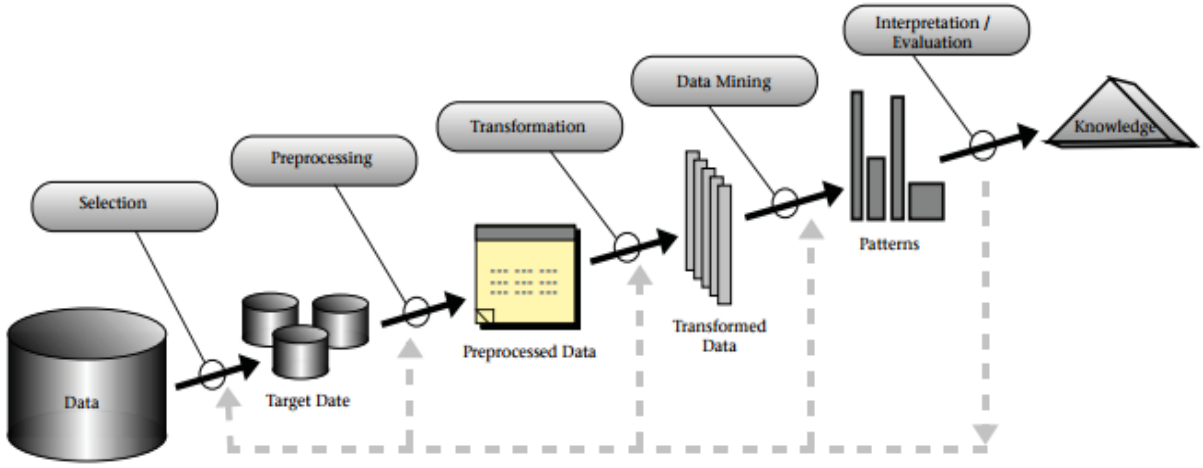

Fig. 1 An Overview of the steps that compose the KDD Process [4]

C.Data Transformation

Data transformation is the stage of transforming obtained data into appropriate forms in which various operations such as correction, integration, generalization and normalization can be undertaken depending on the quality and state of the data.

The next step is to convert the data without clearing the data. Expert opinion has been referred to this data conversion process. The following transformations were made after the production dates per person were assessed as seasonal or monthly, the amount of daily production, total working time, overtime time, working time, number of employers, and accordingly, the original data were deducted from the lost values.

It is determined that the date data is the data of the first 5 months of January, February, March, April, May of 2017. In January, February, March, April and May, in order to ensure the effective functioning of the program and proper classification, it has been gathered in 5 basic groups.

D.Data Reduction

This stage is to filter the data components with respect to the relevancy, where the irrelevant and noisy items/fields/features are left out and repetitions are removed. The query-time is also optimized for data mining processes with data reduction operations so that better results are produced in shorter time.

E. Data Mining

At this stage of the whole process, an appropriate data mining algorithm is decided to be recruited to achieve the overall aim and objectives of the project. The aims may include data classifications, clustering and analysis towards aimed knowledge discovery needed for decision makers. In this study, decision trees algorithms were adapted as data mining classifiers. Decision trees are popular functionalities and tools for data classification and forecasting in which a dependent variable (the decision variable here) is related to independent variables via the branches of the tree, where the branches make up the complete tree representing all relationships. The trees are constructed with a top-down approach, where the most important independent variables may split further sub-branches. A tree can grow very hugely depending on the complexity of the relationships. The CART **(**Classification and Regression Tree**)** algorithm is an approach which stops the enlargement of the tree with irrelevant data and relationships.

F. Interpretation and Evaluation

Interpretation and evaluation step is to assess the information retrieved and included in reports. The information expected and the one found may not be completely compatible due to some technical and procedural reasons. In order to facilitate the decision making process with meaningful knowledge discovery, a substantial interpretation and evaluation mechanism should be made available. This is to avoid inconsistencies and deficiencies between discovered knowledge and the expected knowledge.

Data mining facilitates to obtain meaningful and sound knowledge from huge piles/collections of data. The knowledge acquired as the result of long-lasting and difficult processes can be obtained using data mining technologies in shorter periods of time. This acquired and discovered knowledge can be used for strategic decision making and/or objective evaluations in which it helps analyzing the enterprise

data resources and forecasting the outcome of various business approaches. Variables used in the study are:

- **Date:** January, February, March, April and May of the year 2017 are indicated for a total of 5 months.
- **Number of Employees:** Indicates the number of people working during normal working hours.
- **Working Hour:** Normal This is the amount of man hours during the workday.
- **Overtime:** It is stated that the worker continues to work on completing the normal weekly period.
- **Total Working Time**: The total working time is specified in normal working hours and overtime.
- **Number of Products Produced in Daily Batch**: The number of products produced from the production band is indicated within 1 day.
- **Per Person Production:** Indicates the average product that a person working during the day's work day contributes to the production.

G. Modelling

In order to understand the data in the establishment of the data warehouse and to understand the data, it is necessary to establish a data model first. The data model does not express the definition, structure, and patterns of the data. The purpose of data modeling can be generalized as to what the meaning of the data is, to the relationships between the data, and to clearly specify the properties of the data. A good data model asset should be able to reflect well the qualities of assets, relationships between data, primary, secondary or alternative keys, relationships with entities, and relationships between them [5]. It can be explored with data, fitting different models and investigating different relationships, until you find useful information. Also, it can be found with use of data mining algorithms to extract the information and patterns derived by the KDD process. The result of the model has been shown in Fig. 2.

IV. Research Findings

The result of descriptive statistics can be summarized as:

- When the descriptive statistics regarding the number of workers are examined; The minimum number of workers is 228, the maximum number of employers is 420 and the average number of employers is 351.
- Regarding the study period; The minimum number of employees is 1586 and the maximum number of employees is 2907. The average value is approximately 2442 hours. Conversion not done.
- When the minimum overtime is 0, the maximum working hours are 477 and the average working hours are about 209 hours. Conversion not done.
- The minimum working hours is 1779, the maximum working hours is 3062 and the average working hours is about 2650. Conversion not done.
- Daily production is a minimum of 3417, a maximum of 7050 and an average of about 5966. Conversion not done.

The result of the model has been shown in Fig. 2.

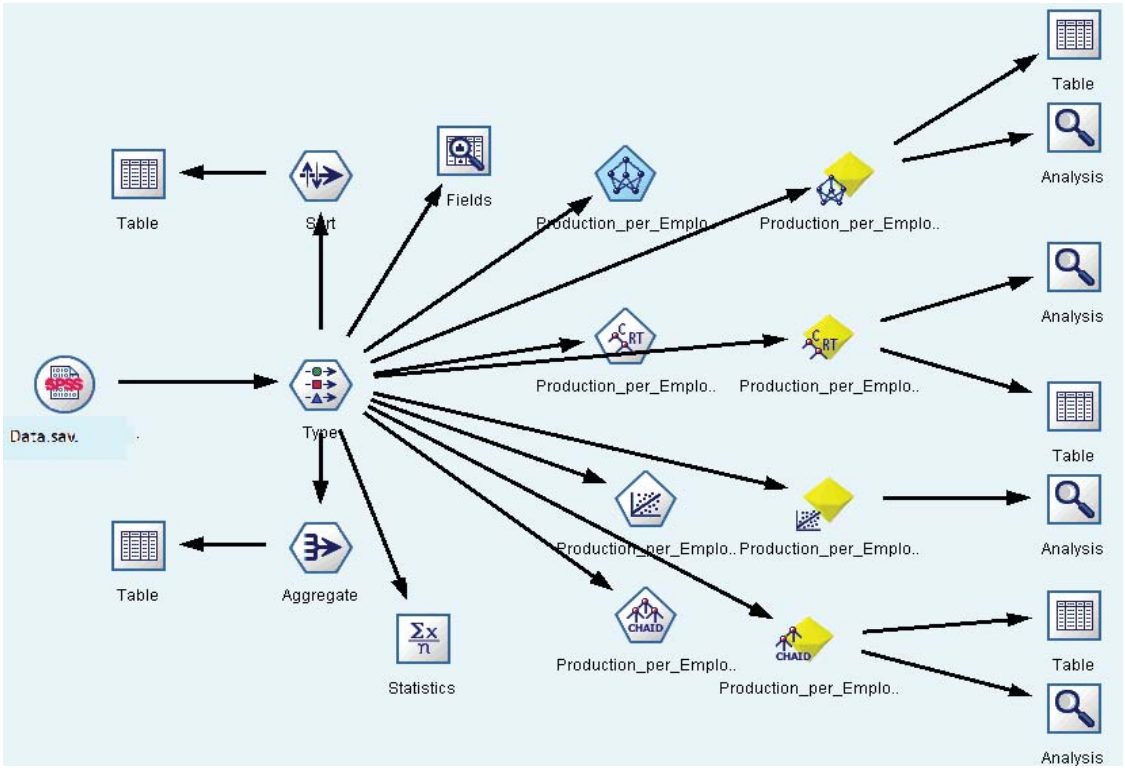

Fig. 2 IBM SPSS Modeler program view of the models used

In Table I, the model errors of 198 personnel were compared according to the per capita production findings, and the model results were given by selecting the most optimal algorithm.

TABLE I
PERFORMANCE RESULTS FOR USED MODELS

	Cart Algorithm	Regression Analysis	Artifical Neural Networks	Chaid Algorithm
Minimum Error	-0,917	-0,747	-0,949	-0,994
Maximum Error	0,5	0,331	0,08	0,2
Average Error	0	0	-0,041	0
Average Absolute Error	0,014	0,119	0,069	0,026
Standard Deviation	0,085	0,179	0,211	0,115
Linear Corelation	0,929	0,628	0,638	0,866

Mean absolute error and standard deviation of the CART model is considered to give good estimation results when the linear correlation showing the prediction success of the developed model is considered.

A. Model Comparisons and Model Selections

The most important variables affecting the production of pants per person in the study were investigated with different classifier models and the results were given.

According to the results of these sequences; While daily production for YSA is among the most important with 21%, number of employees 9% and working time 2%, date and total working time are also effective for CART. In the CHAID algorithm, overtime is the most important, while the regression results differ according to the other three algorithms. While daily production is less important, number of employees, duration of work and overtime are among the most important lists.

B. CART Classification Model Conclusion

In the classification of fabrics according to production quantity. Dependent variable is per capita production, while independent variables are selected as production indicator variables. The result of the CART model used in the study is given in the Fig. 3. The model result given in the Fig. 3 is visually given a decision tree view and interpreted below:

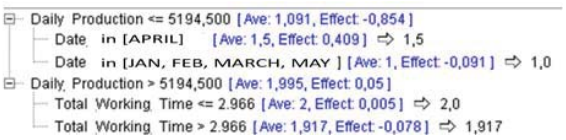

Fig. 3 CART model results for production per person

According to Fig. 3, the most influential variable in the classification of the plant to the production amount is that the "Daily Production" is variable, total production and production history. It was seen that the number of employees, working hours, overtime hours did not appear as important variables in determining per capita output.

Fig. 5 shows that per capita production is less (1.0) in January, February, March, and May than in "fewer producers from around 5195 per person". However, apart from these months (April), it is seen that the production personnel work more efficiently and have more per capita production (1.5).

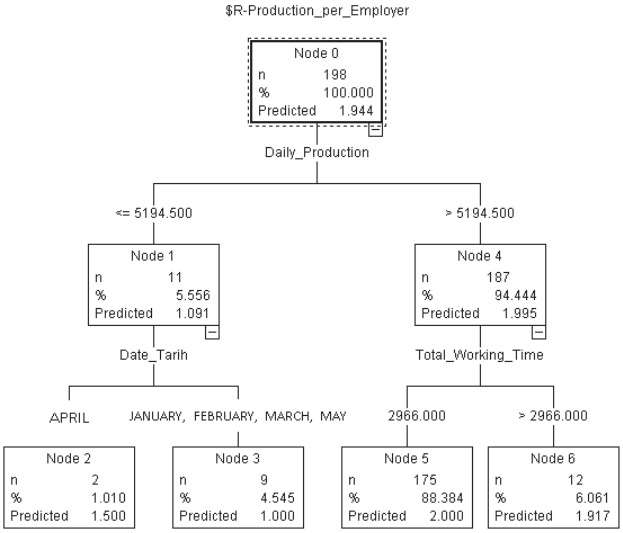

Fig. 4 CART model decision-making result per capacity

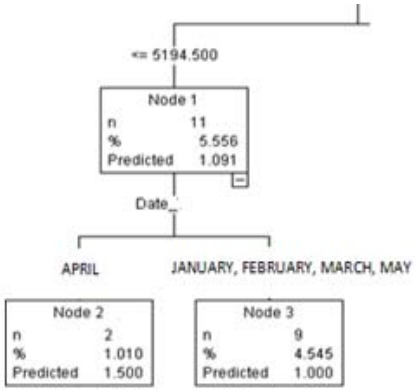

Fig. 5 Decision tree production per person

Node 1: The number of staff members (11 staff members) who make up the profile, the daily production amount is less than 5195 and consists of the working date variable. 5.6% of the staff members in this profession produce 1.1 trousers per person. Node 2: Profiler's staffs are producing in April and 1.01% (2 staff members) of them produce 1.5 pants per person. Node 3: 4.55% (9 staff members) of the 3 rd person who makes the profile makes 1 pants per person in January, February, March and May.

As shown in Fig. 6, it is seen that the number of people who produce "more than about 5195 per person" produces more than 2966 hours of production (2.0). However, it is seen that more than 2966 hours of production are less (1.92). In the profile 4, where there are 187 personnel, the daily production amount is the total working hours variable. 94.44% of the personnel in this profession 187 personnel) produces approximately 2 trousers.

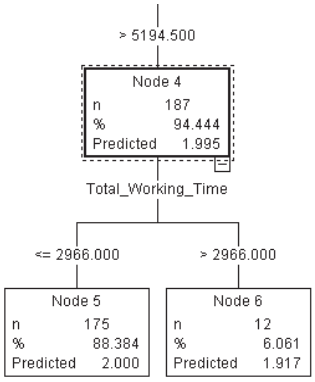

Fig. 6 Decision tree according to production per person

The production of the profile is bigger than the daily production quantity of 5195, the total working time is less than 2966 hours. 88.39% (175 personnel) of this staff are producing 2 pants per person. The daily production amount is larger than 5195, the total working time is 2966 hours. 6.01% (12 staff members) of the staff of this profession produces about 1.92 pieces of 2 pants.

V. Results

It is a matter of data mining to discover hidden information, to determine if the experiences overlap with the analysis result. In this study, decision tree algorithms were used to derive CART algorithm results. The most important advantages of decision tree algorithms are the lack of statistical assumptions that need to be met in other variant techniques because they are among nonparametric methods. Because it is not possible for us to come to a conclusion that all the rules that we have formed as a result of data mining can never be used or used. In addition, decision tree algorithms are among the other advantages of visualizing the direction of relations between dependent and independent variables. This feature makes the interpretation of the results obtained especially simple, making it more concrete and useful.

In this study, data on 198 personnel of a firm producing trousers for data mining techniques were taken and it was found that the long working hours of this firm producing trousers decreased the per capita production without regard to normal or overtime, and the production increased in April End result. Today, the most important variables affecting the amount of per-person pants production in the textile sector are defined as daily production amount, date and total working time.

Textile factories that want to be ambitious in this environment should not only inform the decision-making processes but also work on a certain employee, to make the personnel work more efficiently and to make innovations that will bring the company to an advantageous position.

In this study, we tried to estimate by using daily production amount, working time, overtime, total working time, date, number of employees as estimating variables and compared the importance order of estimations of Artificial Neural Network, Multiple Linear Regression Analysis, CART Algorithm and CHAID Algorithm models created for this purpose.

The estimates of accuracy can be increased by including genetic algorithms and nonlinear regression methods in the data mining process to be applied to estimate the per capita output in studies performed on this field. In addition, similar studies may be used to estimate the amount of per capita output produced by staff during the quality control process as an estimator variable, and to incorporate data collected from different textile factories into the study.

References

[1] Çakır, O "Comparison of Classification Methods in Data Mining", An Application on Banking Customer Database, Ph.D. Thesis, Marmara University Institute of Social Sciences, Istanbul, 2008.
[2] Ersoz, Filiz, "Veri Madenciliği Teknikleri ve Uygulamaları", Dijital Basımevi, 72 Tasarım, Ankara (2015).
[3] Jiawei, H., Kamber M., "Data Mining:Concepts and Techniques", University of Simon Fraser, 2001.
[4] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. "Knowledge Discovery and Data Mining: Towards a Unifying Framework. In Simoudis, E., Han, J., and Fayyad, U., Editors, Proceedings of KDD'96", Second International Conference on Knowledge Discovery & Data Mining, pages 82-88, AAAI Press, Menlo Park, CA.1996
[5] Gürsoy, U. T. Ş. "Data Mining and Knowledge Discovery", Pegem Academy, Ankara (2009).

Filiz ERSOZ graduated from the statistics department of Anadolu University in 1989. She completed her master's degree in 1992 and doctorate degree in 1998 for the field of Biometry/Biostatistics at Ankara University. She worked as a statistician at Turkish Statistical Institute between the years 1989-2001, as an operations research expert at Turkish Land Forces between the years 2001-2006, as probability and stochastic processes expert between the years 2006-2012. She has due rewarded as the rank of Associate Professor quantitative decision-making methods from the Turkish Inter-University Council (UAK) in 2011. Afterwards 2012, she has been working at Karabük University as an academic person. Also, she has taken the Professor degree at Karabük University in 2017 by Industrial Engineering Department.

The applied statistics, multivariate statistical analysis, simulation and modelling, data mining and knowledge discovery, machine learning, statistical quality control and decision making are her research studying area.

.