Orclever Proceedings of Research and Development
https://doi.org/10.56038/oprd.v3i1.280
OR CLEVER Science & Research Group

Conference Article

A Comparative Study of Classification and Clustering Methods for Data Analysis in Digital Transformation and IoT Systems

Turgay Tugay BİLGİN[1] ,Süleyman Burak ALTINIŞIK[2*], Nihat Aydın ADIGÜZEL[2]

[1] Bursa Teknik Üniversitesi, 0000-0002-9245-5728, turgay.bilgin@btu.edu.tr
[2] Mert Yazılım, 0009-0005-0987-1798, burakaltinisik@trex.com.tr
[2] Mert Yazılım, 0009-0002-9201-679X, aydinadiguzel@trex.com.tr

* Correspondence: burakaltinisik@trex.com.tr; +90 444 3468

(First received October 12, 2023 and in final form December 21, 2023)

3rd International Conference on Design, Research and Development (RDCONF 2023)
December 13 - 15, 2023

Reference: Bilgin, T., T., Altinişik, S., B., Adıgüzel, N., A. A Comparative Study of Classification and Clustering Methods for Data Analysis in Digital Transformation and IoT Systems. Orclever Proceedings of Research and Development,3(1), 1-18.

Abstract

This study employs classification and clustering methodologies on datasets derived from digital transformation and Internet of Things (IoT) initiatives within the cable and automotive sectors. The analytical procedures are conducted utilizing the KNIME platform, employing Support Vector Machines (SVM) and K-Means algorithms. The results indicate that SVM exhibits superior accuracy rates compared to K-Means within both industries. The data collection methodology facilitated by the Mert Software IoT platform is identified as reliable and efficacious. The primary objective of this article is to augment decision-making precision in digital transformation software and contribute to the scholarly discourse within this domain.

Keywords: Machine Learning, Classification, Cluster Analysis, Industry 4.0.

https://doi.org/10.56038/oprd.v3i1.280

1. Introduction

Digital transformation has emerged as a significant topic in factories in the world over the recent years. The concept of the Internet of Things (IoT) represents all entities connected to the internet through a network [1]. Digital transformation begins by strategically determining the items that will undergo production on assembly lines within a factory. Subsequently, it encompasses the identification of the staff scheduled for duty during production, specifying their cycle times, and recognizing instances of downtime in the course of the process. Digital transformation assumes a pivotal role in these dimensions, given its paramount importance for industry economies. These changes are imperative for the future and have garnered increased prominence in light of technological advancements. Detecting downtime during production is a critical factor that directly affects manufacturing. When a machine fails, this directly impacts the whole factory, and the length of duration is often unpredictable. The software solutions aim to provide real-time notifications to reduce downtime. The ultimate goal is to increase production quantities. In pursuit of this goal, the data collected by Mert Software has been transformed into knowledge discovery tool through classification and clustering analyses. Clustering analysis is used to classify data based on similarities, particularly for data with an unknown number of groups and unclassified data. It is a technique that groups data into discrete clusters based on similarities with respect to units or variables [2]. This study concentrates on two distinct industries: the automotive industry and the cable production industry. The rationale behind choosing two disparate manufacturing sectors lies in the distinct operational characteristics inherent to each. Within the cable production industry, a singular product is manufactured sequentially utilizing a solitary machine. In contrast, the automotive industry permits the simultaneous production of multiple products. A pivotal demarcation lies in the StockId field within the datasets, exhibiting variation for each record. This study endeavors to assess the efficacy of identical algorithms by applying them to these sectors characterized by disparate production methods.

Klein M., in her article titled "Scenarios of Digital Transformation in Enterprises - A Conceptual Model Proposal," underscores the imperative for businesses to formulate comprehensive digital transformation strategies encompassing processes, business models, customer relationships, and management. Within her study, she delineates the trailblazers of diverse digital transformation methodologies, foreseeing their role in aiding businesses in the development of robust digital transformation strategies [3]. In their article titled "Formation of Digital Factories with Production Tracking Systems and

https://doi.org/10.56038/oprd.v3i1.280

Conceptual Data Analysis," published in 2022, Mustafa and Halil discussed the significance of production tracking, including planning on workstations, raw material inputs, inventory tracking, quality control processes, and maintenance processes, as well as the importance of Key Performance Indicators (KPIs) for businesses. Their study delved into the relationship between Manufacturing Execution Systems (MES) and Enterprise Resource Planning (ERP) systems, highlighting the advantages of digitizing every moment within businesses through these processes. They found that this digitization would significantly reduce production costs, increase labor productivity, decrease production expenses, and facilitate inventory tracking [4].

Altuntaş, conducted a quantitative study involving 114 corporate executives in the article titled "The Impact of Digital Transformation Practices on Corporate Brand Value". The study revealed that 89% of the executives had engaged in digital transformation initiatives, and 34% reported increased revenue and improved customer relationships as a result. Altuntaş argued that the Industry 4.0 revolution would yield positive outcomes for every sector. Furthermore, the author anticipated that digital investments, particularly those enabling brands in the fast-moving consumer goods and retail sectors to offer personalized services to customers, would strengthen brand loyalty and contribute to an increase in brand value [5]. In the article titled "The Impact of the Internet of Things (IoT) Technology on Businesses in the Digital Transformation Process," published by Çark in 2020, the author conducted a study to understand the influence of the Internet of Things (IoT) technology on businesses and to assess the existing literature on the subject. The study utilized content analysis as its methodology and reviewed data from the Web of Science (WoS), Ulakbim, and DergiPark databases. Within the context of the digital transformation process, Çark provided recommendations for preparing and adapting individuals, organizations, and society as a whole to Industry 4.0 technologies. These recommendations aimed to facilitate a healthy transition and adaptation to the digital transformation era [1].

Gürkan's thesis, titled "Development of an Intelligent Factory Management and Information System Supported by Industry 4.0 and Digital Transformation Technologies," introduces a novel approach [6]. The primary goal is to establish a high-tech automation system infrastructure in industry while minimizing human intervention. This approach aims to reduce error rates during the production phase to a minimum and produce high-quality products. It also focuses on enabling factories, especially in terms of production performance, to self-optimize via network connectivity. The objective is to maximize production speed in factories and minimize production costs. In the course of

https://doi.org/10.56038/oprd.v3i1.280

this research, Gürkan developed an Android-based IoT application and implemented it, using electronic cards, specifically in smart marble factories, particularly in the marble drying phase. As a result of this study, several advantages were realized, including the ability to facilitate planned production, contribute to high-quality production, accelerate mass production processes, and minimize production losses [6].

Kaynar and his colleagues, in their 2016 article titled "Sentiment Analysis with Machine Learning Methods," conducted sentiment analysis on datasets containing movie reviews from IMDB using classification analyses, specifically the Support Vector Machines (SVM) algorithm and Multilayer Perceptrons (MLP) algorithms. The study revealed that the SVM analysis had a significantly higher accuracy rate compared to other methods [7].

2. Materials and Methods

2.1. KNIME Platform

KNIME is a platform that processes data and enables reporting through relationships between nodes. Being open source, it is open to further development by programmers who can add additional features to the system. It is generally used in data analysis applications in business intelligence processes. It has various components, and these components are called nodes. Analyses are performed through nodes without the need for coding. The outputs of each node can be viewed and interpreted separately. It is widely used in research related to data analytics. With its powerful features and open system architecture, it is becoming increasingly popular [8].

2.2. Support Vector Machines (SVM)

Classification refers to the process of appropriately distributing data into predefined classes within a dataset. Classification algorithms are used to learn the characteristics of classes from training data and to predict to which class incoming test data belongs. Among classification algorithms, the most commonly used ones in the literature include Naive Bayes, Artificial Neural Networks (ANN), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and the KStar algorithm. In this study, SVM was employed.

SVM is generally used for pattern recognition and classification problems. It trains a support vector classifier using a multi-term kernel. It normalizes all attributes with predefined data [9]. The SVM algorithm is based on Lagrange multiplier equations. Its primary objective is to find the optimal separating hyperplane that best divides data points into different classes. Support Vector Machines accommodate two scenarios: linear

https://doi.org/10.56038/oprd.v3i1.280

and non-linear. Linear SVM is applied to problems that are linearly separable. The structure of linear SVM is illustrated in Figure 1(a).

Let (y_1, y_2, ..., y_n) be the dataset, where $z_i \in (-1,1)$ represents the class labels, and b is the threshold value. In the SVM algorithm, data is separated by the hyperplane $(x) = w^T y + b = 0$.

$$W^T y_t + b \geq +1, \qquad \text{if } z_t = +1 \text{ (class 2)}$$

$$W^T y_t + b \leq -1, \qquad \text{if } z_t = +1 \text{ (class 1)}$$

The points lying below and above the hyperplane $(x) = w^T y + b = 0$ are calculated using Equations 2.1 and 2.2.

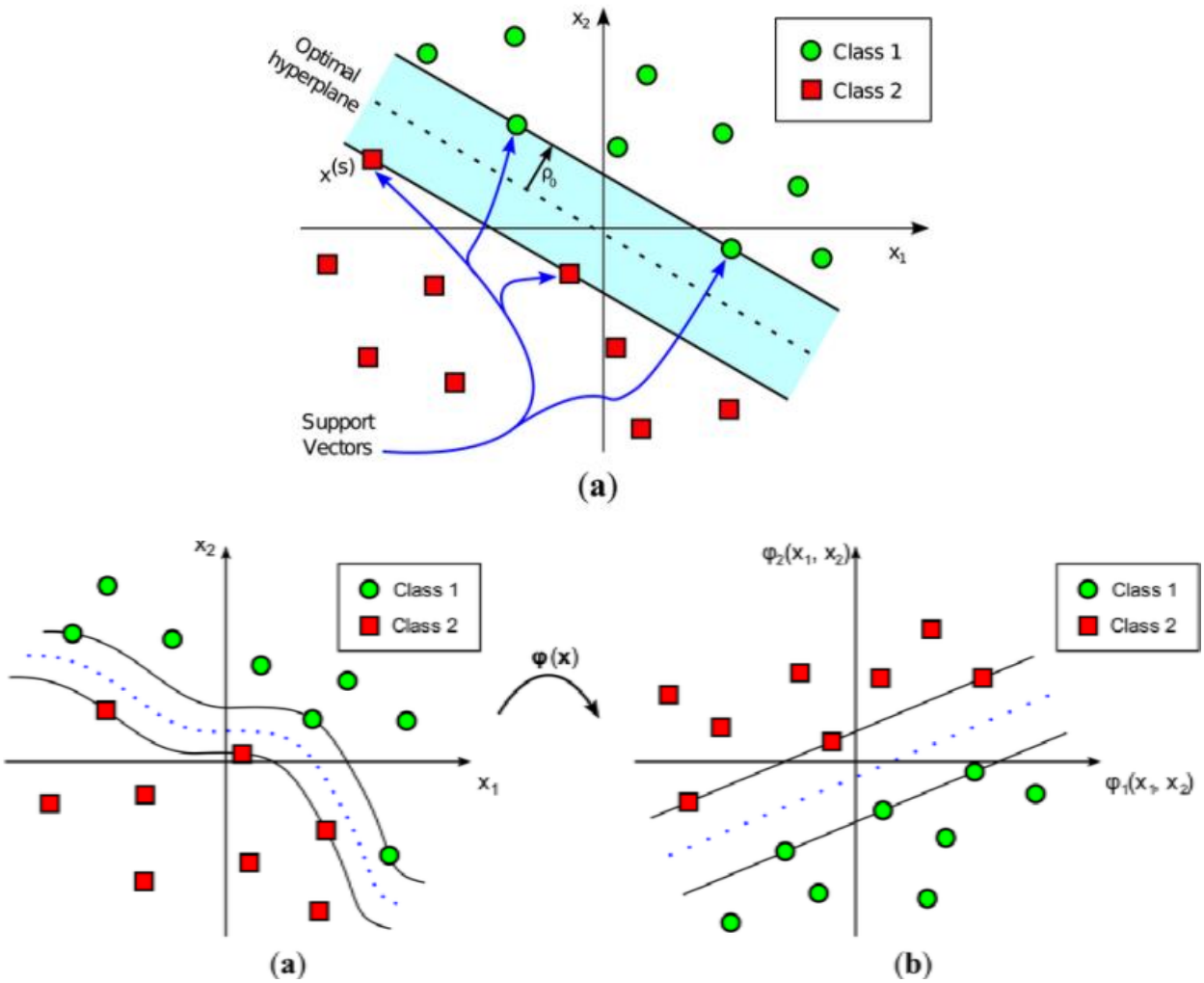

Figure 1. Geometric representation of SVM (a) Linear SVM (b) Non-linear SVM [12]

https://doi.org/10.56038/oprd.v3i1.280

When there is no linear separability, moving the data to a higher-dimensional space can be considered as a solution. This is the fundamental theory behind non-linear Support Vector Machines. SVM achieves these transformations using kernel functions. In this case, a dataset of dimension a is transformed into a new dataset of dimension b, where $b > a$. The structure of this SVM is presented in Figure 1(b). There are numerous kernel functions developed for SVM, and in this study, a radial basis kernel function is utilized. The equation of this function is presented in Equation 1. [13]

$$K(y_i, y_j) = e^{-\frac{\|y_i - y_j\|^2}{2\sigma^2|}} \quad (1)$$

2.3. K-means Algorithm

Cluster analysis is a method that groups units under investigation in a research study into specific categories based on their similarities, enabling classification, revealing common characteristics of units, and making general descriptions related to these classes [10]. In cluster analysis, grouping is done based on similarities and differences. The inputs involve similarity measures or the necessity of calculating which similarities can be applied to the data. Depending on the purpose and field of use, the objectives of cluster analysis are as follows:

1. Identifying the correct types.
2. Building models.
3. Predictions based on groups.
4. Hypothesis testing.
5. Data exploration.
6. Hypothesis generation.
7. Data reduction.

In cluster analysis, distances are calculated between rows of the data matrix. In the formulas, "i" and "j" represent rows of the data matrix, "k" represents columns, "x_ik" represents data in the "i"th row and "k"th column, and "p" represents the total number of variables.

The K-Means clustering method is an effective technique for evaluating many commercial datasets due to its efficiency in clustering large datasets. Being widely used in practical applications for over fifty years, K-Means has become the preferred clustering

https://doi.org/10.56038/oprd.v3i1.280

method in this study for several reasons, such as its speed in comparison to hierarchical clustering methods when there is an expectation of forming a low number of clusters, and its ease of implementation.

The K-Means clustering method is a simple yet effective algorithm for creating clusters based on the available data. The application steps for this method can be outlined as follows [14]:

Step 1: Determination of the number of clusters to which the dataset needs to be partitioned.

Step 2: Random assignment of initial cluster centers for k clusters.

Step 3: Identification of the nearest cluster center for each cluster data point.

Step 4: Calculation of the cluster centroid for k clusters and updating the position of each cluster center based on the new centroid value.

Step 5: Iteration of the processes between steps 3 and 5 until convergence or termination is achieved.

In the third step, the distance from each cluster data point to the nearest cluster center is determined using the Euclidean Distance formula:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2} \tag{2}$$

In the fourth step, each new cluster center is obtained by taking the averages of x and y coordinates using the formulas $\frac{\sum x}{n}$ and $\frac{\sum y}{n}$ respectively.[15]

3. Dataset

In manufacturing facilities where production processes occur, signals generated by machinery, including production and stoppage signals, are conveyed to operator panels through an electronic card known as IOCARD. The IOCARD serves the function of converting signals originating from the machinery into readable data strings, which are subsequently transmitted to the panel software. The panel program, developed in C#, undertakes the reception of this data and parses it based on station-specific definitions. For instance, upon receipt of a signal indicating an increase in production count, the program increments the production accordingly. Similarly, if a signal denoting operational status is received, the program sets the status to 'operational.' In the event of

https://doi.org/10.56038/oprd.v3i1.280

a stoppage signal, the system state is modified to a predetermined 'stop' status. Notably, the system accommodates the establishment of approximately 40 distinct definitions.

In this study, the accuracy of the production data collected by Mert Software IoT platform was used for classification and clustering. This study was conducted using datasets collected in the automotive and cable industries. In Table 1, an example dataset used in the cable industry is provided. In Table 2, an example dataset used in the automotive industry is provided. Table 3 explains the variables, and Table 4 specifies the data types.

Table 1. Data set used in the Cable Industry

StockId	PID	Qty	EmpId	StopId	Situation
1792	30110	0	0	69	Stopped
1792	30110	0	0	69	Stopped
1792	30110	0	0	69	Stopped
1792	30110	1	0	69	Working
1792	30110	0	69	-999	Stopped
1792	30110	0	69	-999	Working
1792	30110	0	69	-999	Working

Table 2. Data set used in the Otomotive Industry

StockId	PID	Qty	EmpId	StopId	Situation
627709	278935	1	297	-999	Working
627707	278935	1	297	-999	Working
627709	278935	1	297	-999	Working
627707	278935	1	297	-999	Working
627709	278935	1	297	-999	Stopped
627707	278935	1	297	-999	Working
627709	278935	1	297	-999	Working

Table 3. Fields and descriptions in the dataset

Fields in the dataset	Description
CompanyId	PrimaryKey, Company Code
workStationId	Station id from which data was collected
StockId	Id of the material produced
PID	Unique id generated by the system
Cycle	Produced stock cycle time
AvgCycle	Avarage cycle time of stock produced
QTY	Produced quantity
QTY2	Produced quantity2 (packet)
QTY3	Produced quantity3 (pallet)
EquipmentId	Equipment id
EmpId	Employee id
Shift	İnstant shift
StopId	Stoppage id
StopStartTime	Stop start time
StopSection	Reason for registration

https://doi.org/10.56038/oprd.v3i1.280

ErrorsId	Error id
EquipmentCoe	Multiplier
ReasonPeqId	If the machine on which production is made connot produce by another machine, the station id that caused it
ID	Unique field
Situation	Machine condition

Table 4. Fields and data types in the dataset

Fields in the dataset	Data Types
CompanyıD	İnt64
WorkStationId	İnt64
StockId	İnt64
PID	İnt64
Cycle	Float64
AvgCycle	Float64
QTY	Float64
QTY2	Float64
QTY3	Float64
EquipmentId	İnt64
EmpId	İnt64
Shift	İnt64
StopId	İnt64
StopStartTime	Datetime
StopSection	İnt64
ErrorsId	İnt64
EquipmentCoe	İnt64
ReasonPeqId	İnt64
ID	Bigint
Situation	string

4. Evaluation

The process of collecting data from production and reporting the classified and clustered analyses of the collected data is as follows:

- Determination of points to receive signals from the machine.
- Installation of SQL SERVER for recording data collected from the machine. Activation of the system.
- Analysis of the collected data. In this phase, the system conducts initial checks before system analyses, verifying the existence of shift definitions. If a shift definition exists, the system initiates operations and performs clustering analysis for Personnel, Production, and the Manufactured product. The system operates synchronously, conducting Classification analysis for Production and Time.
- Following these analyses on the raw data, the results are utilized in reporting and dashboard tools for faster and analyzed data presentation.

https://doi.org/10.56038/oprd.v3i1.280

The workflow of the collected data is shown in Figure 2.

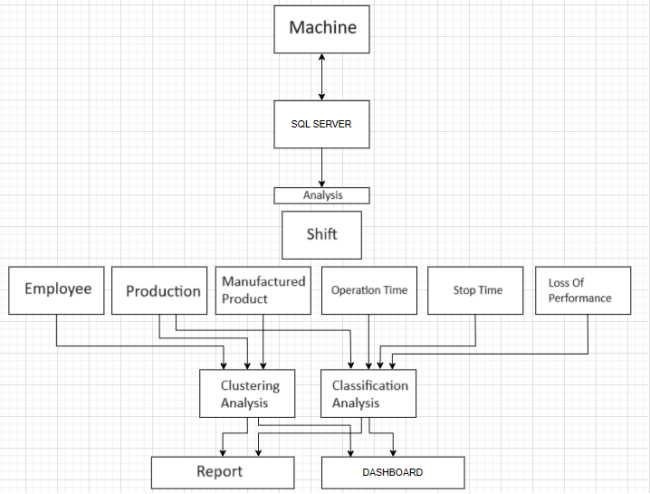

Figure 2. Analysis flow chart

4.1 Confusion Matrix

It is a metric used to evaluate the performance of an algorithm in classification problems. The confusion matrix visualizes the number of correct and incorrect classifications by comparing the actual class labels with the predicted class labels. The confusion matrix is typically presented in the form of a 2x2 table, as shown in Figure 3, but it can be larger for multi-class classification problems. The 2x2 confusion matrix includes the following four elements:

True Positive (TP): The number of true positives. TP increases when the actual class label is positive, and the predicted class label is also positive. True Negative (TN): The number of true negatives. TN increases when the actual class label is negative, and the predicted class label is also negative. False Positive (FP): The number of false positives. FP increases when the actual class label is negative, but the predicted class label is positive. It is also known as a false alarm. False Negative (FN): The number of false negatives. FN increases when the actual class label is positive, but the predicted class label is negative. It represents the cases that were missed.

https://doi.org/10.56038/oprd.v3i1.280

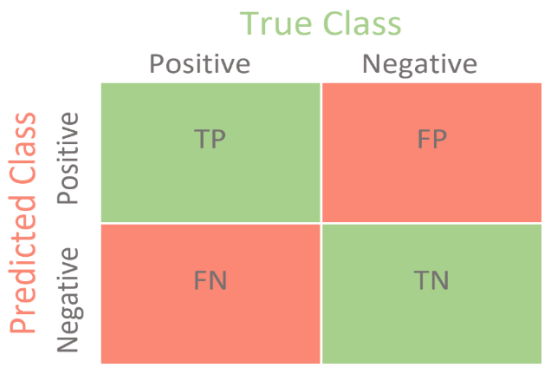

Figure 3. Confusion Matrix for Binary Classification

Confusion matrix is used to calculate various performance metrics using these four elements. Among these metrics, accuracy, precision, recall, and F1 score values are included. The confusion matrix is an important tool for understanding the performance of classification algorithms, assessing the strength of the model, and understanding classification errors [11].

4.2 Performance Metrics

To obtain reliable accuracy results, some measurements are made using the values in the confusion matrix. These measurements are achieved with the accuracy, precision, recall (sensitivity), F1 score, and specificity formulas [11].

4.2.1 Accuracy

Accuracy represents the accuracy value, which is the ratio of correct classifications to the total number of classifications.

$$\text{Accuracy} = \frac{\text{TP}+TN}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100\% \tag{3}$$

4.2.2 Precision

Precision gives the ratio of correctly classified data to all the positives. Here is the formula for precision:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \tag{4}$$

4.2.3 Sensitivity (Recall)

Sensitivity provides the ratio of data correctly classified as positive to the actual positive data. Here is the formula for sensitivity (recall):

https://doi.org/10.56038/oprd.v3i1.280

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{5}$$

4.2.4 F1 Score

F1 Score is a value calculated by taking the harmonic mean of precision and recall values. Here is the formula for calculating the F1 Score:

$$\text{F1 Score} = 2\, x \frac{\text{recall x precision}}{\text{recall} + \text{precision}} \tag{6}$$

4.2.5 Specificity

Specificity is a value calculated by taking the ratio of data correctly classified as negative to the actual negative data. Here is how you can calculate specificity:

$$\text{Specificity} = \frac{\text{True negative}}{\text{True negative} + \text{False positive}} \tag{7}$$

5. Results

5.1 KNIME Workflow

The workflow developed on the KNIME platform is illustrated in Figure 4. Initially, data is read using the "Excel Reader" node, and then irrelevant elements in the dataset are removed using the "Column Filter" node. Elements such as primary key fields like "CompanyId" and "ID" are among those removed from the dataset. To facilitate learning, the dataset is divided into 80% for training and 20% for testing using the "Partitioning" node. Data is normalized for classification analysis. The "Normalizer" node is used to normalize the dataset, applying min-max normalization to the fields "PstopId," "PID," and "QTY," scaling values to a range between 0 and 1. Subsequently, the dataset is split for training and testing through the "Partitioning" node. The training set provides insights into how well the model explains information in the target variable, while the test set indicates the model's performance with unseen observations. In the modeling phase, learning and prediction nodes for Artificial Neural Networks are added to the model. Following the Normalizer node, for classification analysis, the data is linked to the "SVM Learner" node. This node selects the "Status" field for classification. Then, the output of the Normalizer node and the SVM Learner node are combined in the "SVM Predictor" node. This node aligns the learned data from the SVM Learner with the test data input. Subsequently, it connects to the "Scorer (JavaScript)" node for evaluation. Simultaneously, for clustering analysis, the "k-Means" node is employed, receiving input from the Partitioning node based on the fields "QTY" and "QTY2." This node generates two classes with a defined iteration count of 90. The output from the k-Means node, the

https://doi.org/10.56038/oprd.v3i1.280

Clustering Model, is linked as input to the "Cluster Assigner" node. Following the clustering analysis, to interpret the cluster names as "Working" and "Stopping," the "String Replacer" node is connected. This node replaces the name of cluster_0 with "Working" and cluster_1 with "Stopping." Subsequently, it is connected again to another "String Replacer" node, where cluster_1 is named "Stopping." Finally, the workflow connects to the "Scorer (JavaScript)" node and the "Scorer" node for evaluation.

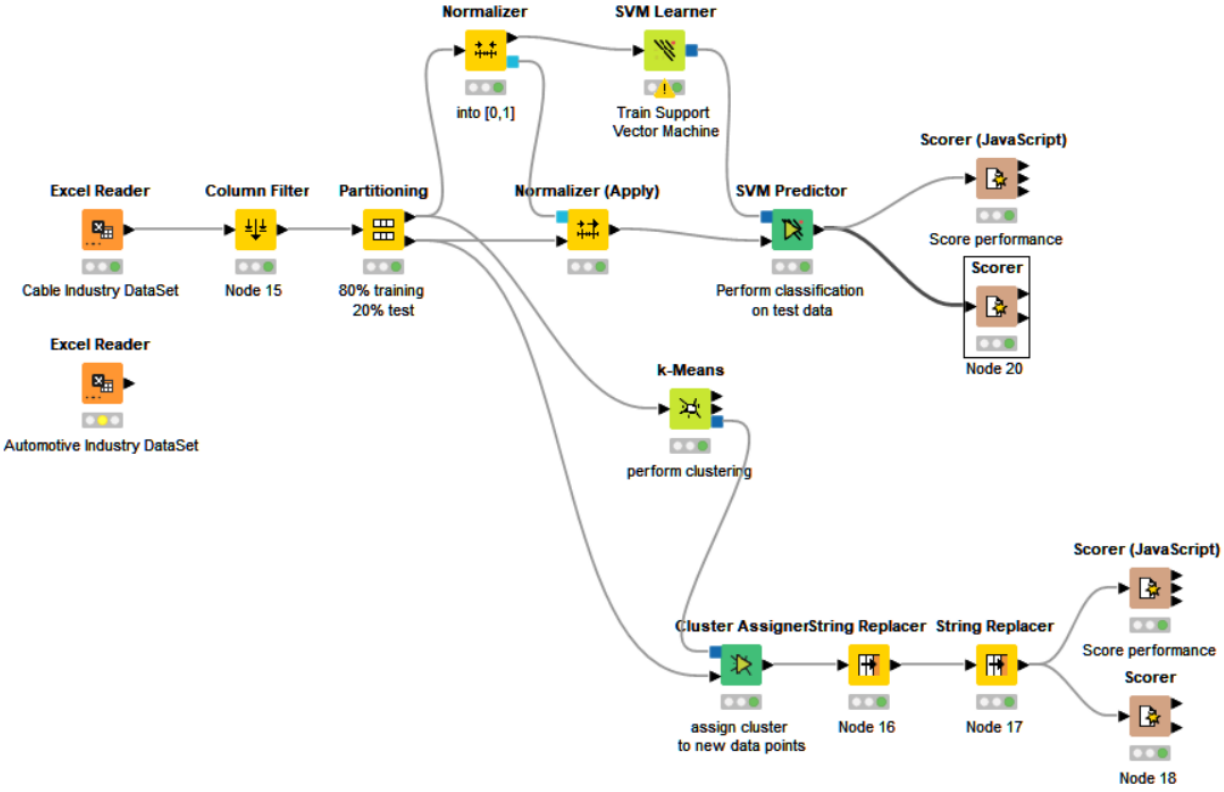

Figure 4. The KNIME Workflow used in this study.

5.2 Support Vector Machine Results

After the model completes its job, the Scorer node produces results, and as shown in Figure 5, an Overall Accuracy of 74,14% and an Overall Error of 25,86% were obtained.

https://doi.org/10.56038/oprd.v3i1.280

Scorer View

Confusion Matrix

	Stopping (Predicted)	Working (Predicted)	
Stopping (Actual)	22	13	62.86%
Working (Actual)	2	21	91.30%
	91.67%	61.76%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
Stopping	22	2	21	13	62.86%	91.67%	62.86%	91.30%	74.58%
Working	21	13	22	2	91.30%	61.76%	91.30%	62.86%	73.68%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
74.14%	25.86%	0.501	43	15

Figure 5. Cable industry results with classification

The performance metrics for the learning results from classification algorithms in the project are shown in Figure 6.

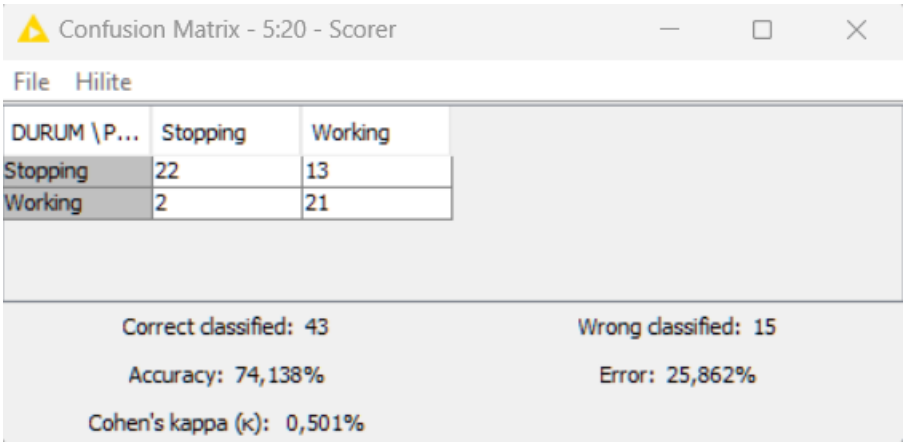

Figure 6. Cable industry SVM algorithm Performance

After the model runs, the Scorer node yields Overall Accuracy of 98.70% and Overall Error of 1.30%, as shown in Figure 7.

https://doi.org/10.56038/oprd.v3i1.280

	Stopping (Predicted)	Working (Predicted)	
Stopping (Actual)	140	21	86.96%
Working (Actual)	7	1993	99.65%
	95.24%	98.96%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
Stopping	140	7	1993	21	86.96%	95.24%	86.96%	99.65%	90.91%
Working	1993	21	140	7	99.65%	98.96%	99.65%	86.96%	99.30%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
98.70%	1.30%	0.902	2133	28

Figure 7. Automotive Industry results of classification

The performance metrics for the learning results from classification algorithms in the project are shown in Figure 8.

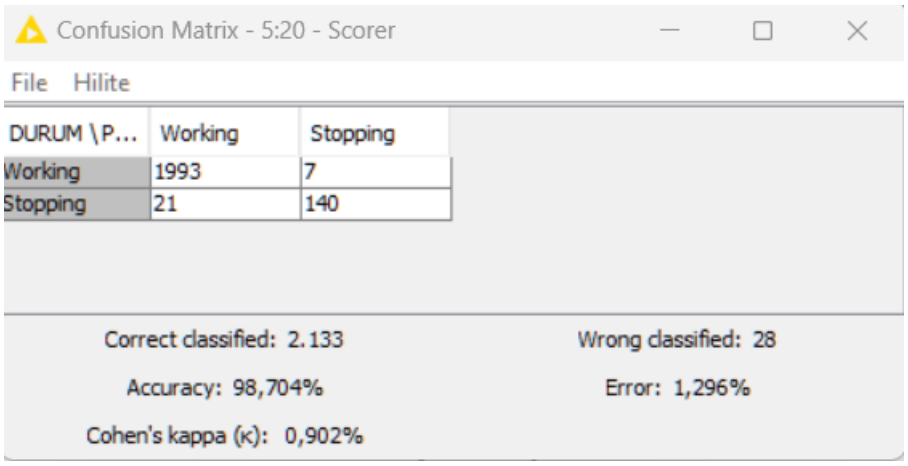

Figure 8. Automotive Sector SVM Algorithm Performance

5.3 K-means Results

Finally, the Scorer node produces results, and as shown in Figure 9, an Overall Accuracy of 67,24% and an Overall Error of 32,76% were obtained.

https://doi.org/10.56038/oprd.v3i1.280

	Stopping (Predicted)	Working (Predicted)	
Stopping (Actual)	30	5	85.71%
Working (Actual)	14	9	39.13%
	68.18%	64.29%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
Stopping	30	14	9	5	85.71%	68.18%	85.71%	39.13%	75.95%
Working	9	5	30	14	39.13%	64.29%	39.13%	85.71%	48.65%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
67.24%	32.76%	0.266	39	19

Figure 9. Cable industry result of cluster analysis

The performance metrics for the learning results of clustering analyses in the project are shown in Figure 10.

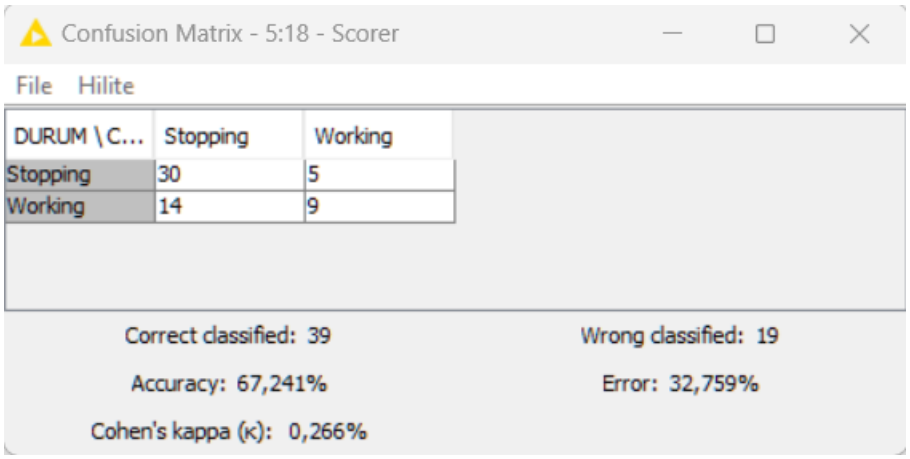

Figure 10. Cable industry cluster analysis Performance

After the model runs, the Scorer node yields Overall Accuracy of 85,47% and Overall Error of 14,53%, as shown in Figure 7.

https://doi.org/10.56038/oprd.v3i1.280

	Stopping (Predicted)	Working (Predicted)	
Stopping (Actual)	35	126	21.74%
Working (Actual)	188	1812	90.60%
	15.70%	93.50%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
Stopping	35	188	1812	126	21.74%	15.70%	21.74%	90.60%	18.23%
Working	1812	126	35	188	90.60%	93.50%	90.60%	21.74%	92.03%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
85.47%	14.53%	0.105	1847	314

Figure 11. Automotive Industry results of cluster analysis

The performance metrics for the learning results of clustering analyses in the project are shown in Figure 12.

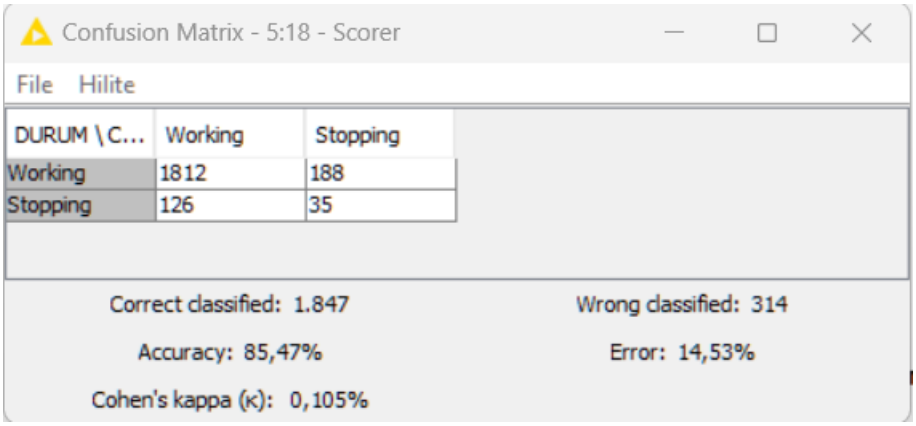

Figure 12. Automotive industry cluster analysis Performance

6. Discussion and Conclusion

In this study, an attempt has been made to classify and cluster the production, downtime, and staff data of companies involved in digital transformation projects by Mert Software. Machine Learning methods such as classification and clustering analyses were employed. In the data sets we used in this study, the SVM algorithm produced more successful results than the K-means algorithm in both the cable industry and the automotive industry. The study revealed that the method employed by Mert Software for digital transformation projects achieves high accuracy rates.

A major problem with industrial data is that sometimes operators make decisions that can lead to incorrect feedback. This affects the accuracy of the data and causes problems. In this study, we selected examples from factories where signals about production and downtime are collected automatically so that operator errors do not affect data quality. We believe that this study can contribute to the decision-making accuracy of digital transformation software and contribute to scientific research in this field.

References

[1] Çark, Ö. (2020). İşletmelerin dijital dönüşüm sürecinde "nesnelerin interneti" teknolojisinin etkisi. Turkish Studies-Economy, 15(3), 1247-1266.

[2] Çakmak, Z.. Uzgören, N., & Keçek, G. (2005). Kümeleme analizi teknikleri ile illerin kültürel yapılarına göre sınıflandırılması ve değişimlerin incelenmesi.

[3] Klein, M. (2020). İŞLETMELERİN DİJİTAL DÖNÜŞÜM SENARYOLARI - KAVRAMSAL BİR MODEL ÖNERİSİ . Elektronik Sosyal Bilimler Dergisi , 19 (74) , 997-1019 . DOI: 10.17755/esosder.676984 (Erişim Tarihi: 25.06.2023)

[4] Kılıç, H. & Timur, M. (2022). Üretim Takip Sistemleri ve Kavramsal Veri Analizi ile Dijital Fabrika Oluşumu. Avrupa Bilim ve Teknoloji Dergisi , (33) , 285-289. DOI: 10.31590/ejosat.996760 (Erişim Tarihi: 25.06.2023)

[5] Yılmaz Altuntaş, E. (2018). DİJİTAL DÖNÜŞÜM UYGULAMALARININ KURUMLARIN MARKA DEĞERİ ÜZERİNDEKİ ETKİSİ. Ege Üniversitesi İletişim Fakültesi Medya ve İletişim Araştırmaları Hakemli E-Dergisi , (2) , 1-18 . Retrieved from https://dergipark.org.tr/en/pub/egemiadergisi/issue/36758/384936 (Erişim Tarihi: 25.06.2023)

[6] Gürkan, Ç. Endüstri 4.0 ve Dijital Dönüşüm Teknolojileri ile Desteklenen Akıllı Fabrika Yönetim ve Bilişim Sisteminin Geliştirilmesi. (Erişim Tarihi: 25.06.2023)

[7] Kaynar, O., Görmez, Y., Yıldız, M., & Albayrak, A. (2016, September). Makine öğrenmesi yöntemleri ile Duygu Analizi. In International Artificial Intelligence and Data Processing Symposium (IDAP'16) (Vol. 17, No. 18, pp. 17-18). (Erişim Tarihi: 25.06.2023)

[8] Aksan, C.e. 2022. KNIME Nedir?. https://cekasan.com/tr/knime-nedir (Erişim Tarihi: 05.06.2023)

[9] Data Mining Software in Java, www.cs.waikato.ac.nz/~ml/weka/ (Erişim Tarihi: 16.06.2023)

[10] Ardıl, E., (2009). Esnek Hesaplama Yaklaşımı İle Yazılım Hata Kestirimi, Yüksek Lisans Tezi, Trakya Üniversitesi, Fen Bilimleri Enstitüsü, 86s.

[11] 14 - Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. Machine learning, **6(1):** 37-66.

[12] 15 - Ruiz-Gonzalez, R.; Gomez-Gil, J.; Gomez-Gil, F.J.; Martínez-Martínez, V. An SVM-Based Classifier for Estimating the State of Various Rotating Components in Agro-Industrial Machinery with a Vibration Signal Acquired from a Single Point on the Machine Chassis. Sensors 2014, 14, 20713-20735.

[13] Peker, M., & Özkaraca, O. (2018). Büyük ölçekli veri setleri için GPU hızlandırmalı melez bir GA- SVM: Cu-GA-SVM. Gazi University Journal of Science Part C: Design and Technology, 6(3), 581-591.

[14] Larose D. T., Discovering Knowledge in Data an Introduction to Data Mining, WILEY, ABD, 2005

[15] *Dündar, S. (2023). TR83 bölgesinde K-Means ve ARAS yöntemiyle kompost tesisi kuruluş yeri seçimi. Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 38(4), 2607-2624.*