Application of Data Mining Techniques for the Investigation of Factors Affecting Transportation Enterprises

Mahdi Yousefi Nejad Attari, Assistant Professor, Department of Industrial engineering, Bonab branch, Islamic Azad University, Bonab, Iran. Babak Ejlaly, Adjunct Professor, Kooshyar University, Hannaneh Heidarpour, Department of Industrial Engineering, Seraj institute of higher education, Tabriz, Iran, Ali Ala, Department of Industrial Engineering & Management, Shanghai Jiao Tong University, Shanghai, China

Abstract— In land transportation, due to the importance of this business and competition between active organizations in this field, applying new technologies in management and making better decisions can be beneficial. We have presented three data mining techniques of clustering, association rules, and classification to investigate the factors affecting the cost and time of road and rail transportation. Using methods based on the K-means algorithm with comparing four clustering, we have proposed Naive Bayes (probabilistic) classification to determine the total accuracy of transportation percent to 97.91%. Finally, classification tree algorithms such as Bayesian theory and random forest have been used, and the results and output rules have been compared. This article is comprehensive and new to use various effective parameters inland transportation. We will confirm its efficiency by using the criterion(5v)(which we will explain in its place) and then the results in the field. A larger one, called land transit, could be generalized between the two countries. In the end, we have discussed more in methodology and results

Keywords — Data mining, Time, road and rail transportation, clustering, cost.

1. INTRODUCTION

In today's world, data and information are considered as organizational wealth and organizations. The world's largest and most successful companies are always looking for more appropriate and commercial use of these virtual resources. On the other hand, with the complexity of business environments, the nature and volume of organizational data have become

very different. An integrated and managerial view of them becomes essential. In transportation engineering, a lot of data has been extracted in traffic, accident analysis, asphalt conditions, road durability, bridge management, and so on. One of the solutions adopted by successful organizations in this field is to create a comprehensive system of data and statistics and the correct use of data mining and knowledge discovery techniques.[1]

One solution that can speed up decision-making in activities and increase decision-making accuracy is to correctly use data mining techniques on existing data to discover organizational knowledge or predict future situations. Transportation plays an essential role in economic growth and globalization.

Research shows that only 6.35% of the data mining method has been used to examine the cost-effectiveness of transportation (Izadi and et al 2020). Due to importance of cost and its classification into three parts, (1) from the freight transport operators 'perspectival costs, (2) from the freight owners 'perspective and (3) from the national perspective. In this research, transportation costs have been considered as a combination of these costs and we played a bigger role in the cost factor, but we did not neglect to pay attention to other factors.

Table1. Iran-Turkey non-oil trade balance (by land) (from 1997 to 2006)

Year	Export	Export	Import	Import	Trade balance (million dollars)
	Value (million dollars)	Weight (million dollars)	Value (million dollars)	Weight (million dollars)	
1997	741	326	673	887	-561
1998	550	566	768	1246	-681
1999	389	530	1007	1506	-976
2000	562	593	1371	2017	-1424
2001	692	1056	1820	2464	-1408
2002	1121	1430	1816	3080	-1649
2003	1209	1476	2746	3782	-2307
2004	1539	1649	2662	3605	-1956
2005	2164	1976	3564	3873	-1898
2006	1631	1314	2601	2995	-1681

Existence of extensive land trade Table 1. between the two neighboring countries of Iran and Turkey emphasizes the

This paragraph of the first submitted on December 12, 2020.

Mahdi Yousefi Nejad Attari is an Assistant Professor, Department of Industrial engineering, Bonab branch, Islamic Azad University, Bonab, Iran (e-mail: myousefi@bonabiau.ac.ir). Babak Ejlaly is an adjunct professor in Kooshyar University, Rasht, Iran (e-mail: babak_ejlaly@kooshyar.ac.ir). Hannane Heidarpour in Department of Industrial Engineering, Seraj institute of higher education, Tabriz, Iran, (e-mail: h_heydarpoor1993@yahoo.com).

Ali Ala, Department of Industrial Engineering & Management, Shanghai Jiao Tong University, Shanghai, China (email: Kanban@sjtu.edu.cn)

importance of planning and paying attention to the effective factors in the field of transportation.

In this article, data mining (in the organization under study) is used as a planning technique to generalize the results in the field of transit between the two countries.

In this article, we will confirm the efficiency of our method using the criteria (5 v) (Li Zhu and et al. 2018). Below we briefly define the above criteria

Volume: refers to the quantity of data produced by various sources.

Variety: is mainly focused on all kinds of data by detector, sensors and even social media.

Velocity: data in transportation has increased due to improved communications technologies, increased processing power and speed of monitoring and processing.

Veracity: Refine data and delete unfamiliar data.

Value of Time: value of travel time.

The Promotion organization of Iran It is in charge of promoting non-oil exports in Iran has a significant role to export the goods between Iran and Turkey. In 2021, Turkey was the largest trading partner of Iran in the European region with 5 billion 46 million and 49 thousand dollars of imports from Iran and 4 billion and 995 million and 631 thousand and 963 dollars of exports. Through transportation planning, risk assessment, road conditions, highway feature inventory, lane markings, and signal inventory, weather conditions, road features inventory, and other studies in transportation engineering, large amounts of data are produced. Decision-makers make decisions based on these data in order to address a specific issue. Decision-makers are constantly on the hunt for ways to make accessing and applying diverse datasets easier. Categorizing what data is available, evaluating the pattern of the phenomenon, extracting the data of interest, and converting the data into formats required for the application are the basic requirements. Various data fields must be gathered in the real-world condition of the transportation domain to incorporate and arrive at solutions. A recent data mining analysis has opened a new frontier for transportation engineers' decision-makers.

The paper's main objectives are as follows:

- To introduce briefly the data mining and commercially available software
- To review data mining applications in transportation engineering.
- To demonstrate a data mining problem with commercially available software for the 'transportation cost loading data' problem.
- Using the Probabilistic method to get efficiency percent to total cost rather than another related method.

2. Theoretical Foundations and Research Background

So far, several definitions have been proposed for data mining.

In some of these definitions, data mining is a tool that enables users to communicate directly with large volumes of data. In some more precise definitions, data exploration has also been considered. Here are some of these definitions.

The term data mining refers to the semi-automated process of analyzing large databases to find functional patterns. [2]

Data mining means searching in a collection of data to find patterns among them. [3]

Data mining means extracting new and citable knowledge from a large database set. [3]

Data mining means analyzing the observable data set to find reliable relationships between data. [4]

2-1. Data Mining Process

According to Cross-Industry Standard Process for DM, five phases can be considered for data mining projects: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. [5]

This standard process, shown in Figure 1, does not move linearly; instead, it is possible to go back at any stage. And this process is so repeated until we finally achieve the best result. We have also, define the acronyms of each factor of influence on transportation in Table 2. Based on this we have to easily understand what are factors meaning in each part.

Table 2. acronyms definition for each factor in transportation topic

Factors of Influencing on Transportation	Acronyms	Definition
Demand Data	DD	The information of
Amount of demand	AD	Quantity demand of total amount good or service
Loading Type	LT	Type of vehicle or vehicle to be loaded
Amount of Supply	AS	Supply is a fundamental economic concept that describes the total amount of a specific good or service that is available to consumers
Loading Cargo	LC	services of loading cargo between any place or point of rest on a wharf or terminal, and railcars, trucks, or any other means of land transportation.
Estimated Transportation Time	ETT	Estimated time to transport cargo from the warehouse to the destination
Transportation Time for Loading	TTL	Loading time on transport lines
Transportation Cost (in Rials):	TC	The cost of transportation from A location to B location
Transportation Type	TT	Type of transportation used for loading include in Ship, plane and Truck
Weather Condition	WC	The type of weather effect on the total arriving time a cargo to location

Figure 1: Nonlinear data mining standard model [5]

2.2. K-Means Algorithm

Clustering is one of the branches of unsupervised learning. It is an automated process during which samples are divided into classes whose members are similar to each other, called clusters. Thus, a cluster is a set of objects in which things are similar to each other and are dissimilar to objects in other clusters. Different criteria can be considered for similarity; for example, the distance metric can be used for clustering, and closer to each other can be considered a cluster, which is also called distance-based clustering. One of these methods is k-means clustering. [2]

In a simple type of this method, first, the required number of clusters of points is randomly selected. Then, in the data, according to the degree of proximity (similarity), they are attributed to one of these clusters, and thus new clusters are obtained. By repeating the same procedure, new centers can be calculated for them in each repetition by accepting the average of the data and re-assigned the data to new clusters. This process continues until the data is no longer changed. [9]

Table 3. different literature review using different methods

Author title	Article title	Method	Year of publication	Description
Bhayaan and Rao	Defining los criteria of urban streets using GPS Data	K-means and K-medoid clustering	2010	Use only the speed criterion to divide urban areas
Zhu and et al.	Parameters Calibration of Traffic Simulation Model Based on Data Mining	K-nearest neighbor and locally weighted regression	2012	Method to calibrate the parameters of the speed-density traffic simulation to calculate the speed of vehicles
Tao Pei and et al.	Density-based clustering for data containing two types of points	Two-component clustering	2014	Two-component clustering method to recognize different types of clusters from corresponding regions(urban areas) of the density domain
Nezernko and et al.	The situation of transportation field in Baltic Sea Region	Bayesian analysis and affinity analysis	2015	Focus only the flow transportation
Li Zhu and Fei Richard and et al.	Big data analytics in Intelligent Transportation System:(A survey)	Use ANN (For example, SVM classification for kernel function)	2017	The ANN method used to predict incidents
Faezeh Ghofrani and Qing He and et al.	Recent applications of big data analytics in railway transportation system: A survey	Simulation classification clustering Estimation	2018	They selected a four-layer structure 1.operation and safety and maintenance 2.predictive and perspective and etc. 3. classification and clustering and etc. 4.SVM and SVR and ANN and etc.

This research as the Table 3. shows that in the study of factors affecting transportation from limited parameters such as vehicle speed or transport flow and from methods such as clustering and classification and ANN have been used, while the above research in terms of factors affecting transportation (Table 1) is comprehensive and a combination of clustering Rapid Miner classification tree algorithms such as Bayesian theory and random forest have been used and the results and output rules have been compared.

With reference to table number one, which includes the factors affecting transportation, and in comparison, with previous studies (table above), if from a simple weighting method (assuming the weight of all factors is equal), the score is as follows:

Table 4. different factors effecting on transportation based on various methods

Factors affecting transportation	K-means and K-medoid clustering(2010)	K-nearest neighbor and locally weighted regression(2012)	Two-component Clustering(2014)	Bayesian analysis and affinity analysis(2015)	Use ANN (for example, SVM classification for kernel function)(2017)	Simulation classification clustering Estimation(2018)	This study(NOW)
Demand data							✓
Amount of demand							✓
Demand tonnage(kg)							✓
Origin			✓	✓	✓	✓	✓
Destination			✓	✓	✓	✓	✓
Loading Type				✓	✓	✓	✓
Amount of Supply				✓	✓	✓	✓
Loading tonnage(kg)				✓			✓
Estimated Transportation time (per day)		✓				✓	✓
Transportation time from the day of demand to the day of unloading	✓	✓		✓	✓	✓	✓
Transportation Cost (in Rials)							✓
Transportation type							✓
Weather conditions							✓
License permission							✓
Positive score	10	20	20	40	50	60	140
Negative score	-130	-120	-120	-100	-90	-80	0

The parameters of the Table 4. which have been used to analyze the strengths and weakness of each of the proposed models, have been extracted from a review of the literature . This Table uses the simple scoring method(assume the weight of all these parameters is the same and consider it 10 for ease of operation). For example, the Two-Component clustering model has a positive score of 20 and has weaknesses, including inattention to demand data, amount of demand, demanded tonnage(kg), etc., which in total has a negative score of minus 120. While our new research is comprehensive, it has a positive score of 140. Then, a table has been provided from shipping exported during the last year and six months by an international transport company. First, data clustering has been done using Rapid Miner, and then association rules have been extracted. Finally, classification tree algorithms such as Bayesian theory and random forest have been used, and the results and output rules have been compared.

2.3. Clustering Assessment(Davies Bouldin Index)[2]

There are different criteria for evaluating clustering algorithms that can be divided into unsupervised and supervised classes. Unsupervised evaluation criteria sometimes referred to as internal criteria in scientific literature, are those operations that are used according to the information contained in the data set. The most important task of a clustering algorithm is to minimize the within-cluster distance and maximize the distance between clusters.

Hence, the two important factors used in all evaluation criteria are cluster density and cluster separation and meeting

the goals of minimizing the distance within clusters and maximizing the distance between clusters in the group, respectively, maximizing each cluster's density and maximizing the separation between sets.

One of these criteria is Davis Bouldin which is defined based on these two essential factors:

Separation factor or Separation in this criterion is defined as $dis(C_i, C_j)$ in which dis is the Euclidean distance between the center of two clusters.

The density factor or Cohesion is defined as $\sum_{x \in C_i} dis(x, C_i) \cdot \frac{1}{m}$. Where x is the point inside the cluster and c is the center of the cluster.

Davis Bouldin's general formula is as follows: The value of NC is for all points.

$$\frac{1}{nc} \sum_{i=1}^{i_{nc}} \left\{ \frac{coh(i) + coh(j)}{Sep(i, j)} \right\} \quad (1-2)$$

2.4. FP-Growth Algorithm

There are many algorithms for the production of association rules. Two of the most widely used methods are Apriority and FP-Growth algorithms. In this research, we use the second method.

It identifies frequent patterns without producing a set of candidate items and with the help of a data tree structure. This algorithm, which follows the Divide and Conquers strategy, first transforms the data set into a FP-tree. It then directly extracts frequent patterns from this tree.

Concisely, it can be said that one FP-tree is a compact representation of transaction data set. For this algorithm tree, transaction items are read one by one and mapped as a path on the tree.

Because transactions typically use common items; therefore, the transaction paths overlap on the tree, and because of this, the tree will be more compact than the initial data. In the first database scan, the set of one-member items and their support are identified, and also the set of repetitive items are arranged in descending order of their support. A tree is then constructed as follows: First, the tree's root with the null label is built. The dataset is then scanned a second time.

The items in each transaction are processed in L order, and a branch is created for each transaction. for facilitating tree navigation, a table is created in which each pen is pointed to its place in the tree. The tree is completed after scanning all transactions. [2][9].

2.5. Classification Algorithm

Bayesian theory is one of the statistical methods of classification.

In this method, different classes are each considered as a hypothesis with probability.

Each new training record increases or decreases the probability that the previous hypotheses are correct. Finally, the hypotheses that have the highest probability are considered as a class and are labeled. This technique combines Bayes's theory with a causal relationship between data. [2]

The tree-based random forest produces many decisions. Classify a new object, the input vector is placed at the end of each of the random forest trees. For each tree that provides a classification, this tree gives that class a "vote". The forest chooses the classification that has the most votes (among all forest trees). [10] Each tree is formed as follows:

1. If N is the number of nodes in the train data set, N nodes randomly, with the substitute of the primary data, are sampled. This is a sample work developed for this tree.

2. If we have M variables and consider m smaller than M so that in each node, the M variables are randomly selected from M , and the best separation is used on these m variables to separate the node. The value of m is considered constant during forest construction.

3. Each tree grows as large as possible. There is no pruning.

2.6. Data Mining in Transportation

In transportation engineering, a lot of data has been extracted in the areas of the field of traffic, accident analysis, asphalt conditions, road durability, stair management, and so on. Based on this data, decision-makers decide to solve a related problem. Decision-makers are always looking for ways to access data and use different data easily. The ability to identify available data, determine data features, extract required data, and convert data to the required formats is among the requirements for planning. In the real world of Transportation, many different information has to be merged to reach a solution. Recent research in transportation data mining has created new horizons for decision-makers of transportation engineers.

In 2014, Li and Chen combined three data mining methods: K-means clustering, decision trees, and neural networks that predicted freeway travel time with non-recurrent congestion. [11]

Ahmadian et al. used the Priority algorithm to obtain the available association rules in the time/place intervals of problems. They also used a 2-step technique to identify simultaneously over a certain period. Finally, they showed that the resulting laws be used to determine the needs of citizens and improve the management of municipal services. [12]

In his research about time series prediction, Peter Zhang used a combined Autoregressive Integrated Moving Average (ARIMA) model and neural networks. He concluded that the use of time series models alone does not have the accuracy and power of neural networks.

He proposed combining artificial neural network models and ARIMA as an optimal solution in predicting time series. He claimed that a combination of artificial neural network models as a nonlinear model and ARIMA model as a linear model in time series prediction performs better than the prediction made using each model alone. For proving this claim using empirical data, he first modeled the data set using the ARIMA model and compared the output with the output of his hypothesis model. Based on that, he proved his hypothesis. [13]

Lathia, Froehlich, and Capra, in an article Mining Public Transport Usage for Personalized Intelligent Transport Systems, used data mining tools and techniques and reduced the travel time of passengers to reach their destination and even predicted the travel patterns of travelers. A case study of this research has been done using the data available in the London Subway transportation system database. It has estimated the travel time of each passenger and extracted the stations that are used most by each passenger. By presenting a model, it offers the passenger the route that has the shortest travel time. Another purpose of this article was to determine the most used stations because this information effectively improves and changes the travel plans of that station. [15]

In their article, Adams, Akano, and Asemota predicted the amount of electricity needed in Nigeria. The goal was to predict the amount of electricity by the country in the next ten years, knowing the amount of electricity generated between 1970 and 2009. The reason for this research was that in 2005, due to insufficient electricity production and reduced rainfall, only 40% of the country's electricity needs were provided. Therefore, the use of a model that can make this prediction was the group's objective. In this study, the group used the ARIMA model for forecasting and explained the reason as follows: ARIMA model is a general model and can model data and series with seasonal and non-seasonal period and determine the degree of this model they used ACF and PACF. After modeling test data with different degrees of autoregressive, moving average, differentiation, evaluation of outputs based on parameters Stationary R2 and Bayesian Information Criteria (BIC). this group selects the best model on the amount of required electricity. They predicted the electricity for the next ten years. [17]

3. Crisp Process

America's most extensive software and hardware Electronics Company (National Cash Register, NCR), as part of its goal, has established teams of consultants and data mining technicians to serve the needs of its customers to give more value to its data warehouse customers (Teradata) [29] The data mining process in the Crisp data mining standard is carried out in six stages by the data discovery method presented by Osama Fayyad et al. [6]; In the following, the method of conducting the research based on these steps will be described.

Figure 2: Crisp process model [5]

3.1. Understanding the Business Environment

Understanding Business Objectives: The first step in any data mining project is to determine what is expected from the implementation of the project.

Understanding the goals:

Business goals are proposed in business or health, and data mining goals should be examined in the field of data mining and its techniques. The most critical output at this stage is the list of data mining objectives. Also, the criteria for data mining success must be determined.

Preparing a project program:

At the end of the first stage of Crisp, a program similar to other studies should be set. In addition to specifying the details of research activities and the time of its implementation, the required methods and tools in the project must be selected in this program. The output of this part of the process is in the form of implementation documents and a research schedule. The program of the considered research is presented in the form of a proposal in advance.

3.2. Understanding The Data

Collection of initial data: In this stage, depending on the type of project, data are collected by different sources with a questionnaire or as systematic.

Data Description: Data description can be defined as the preliminary examination of the data and the extraction of the initial features of the data set.

In this research, the data format, number of variables, type of variables (continuous and discrete) data, and the concept of variables are examined in this stage. Also, in this section, according to the number of records, it was determined whether or not the type and volume of data meet the requirements of the research.

Data Discovery: After a preliminary description of the data, a closer look at the data should be provided. At this stage, the main features of the data should be extracted using statistical methods, and the features of the data should be depicted with the help of visualization tools.

Examining the quality of data:

The output of this step is usually in the form of a report in which data errors, missing values, and duplicate records are identified. Data errors are meaningless values and remote data that affect results.

3.3. Data preparation

Data selection: In some problems, due to the large volume of records available, sampling methods are used to reduce the volume of data.

Data wiping: In this section, the lost values should be replaced by using the output report of the data quality check in the previous step. Also, in the second stage, the meaningless values of the variables should be deleted or replaced with meaningful values according to the report of the concepts of the variables in the previous step.

Reducing variables: At this stage, after data wiping, the existing methods of feature selection should be used to find influential variables and eliminate low-impact and ineffective variables.

This step is efficient, and when the number of variables is large, it reduces the implementation time and modeling efficiency, so it is a significant step.

3.4. Modeling and Evaluations

Selection of modeling method: In this step, the first several classifications or clustering techniques will be selected based on the data type. Depending on the preprocessing stage, we may need to go back to this stage based on the evaluation parameters and examine the different classification techniques and parameters.

Evaluation of results: In this step, the evaluation criteria of the model can be selected according to the field of research and study of valuable articles and analyses done on the data; and the best method and class in each step was selected based on this feature. In cases where the cost is low, and the budget is sufficient, the created model can be tested on other real data sets, and the accuracy of the developed model can be further examined.

4. Model data Collection and Information

The collected data set includes 1389 information related to the Transportation of various products of a transport company on multiple types of railways and roads. It exports products including zinc ingots and aluminum ingots to Turkey. The data consists of 14 numerical and discrete columns. The features are defined as in Table 5.

Table 5: Data and their type

TITLE	VARIABLE TYPE
Demand data	SOLAR DATE
Amount of demand	NUMERICAL
Demanded tonnage(in KG)	NUMERICAL
Origin	DISCRETE
Destination	DISCRETE

Loading Type discrete	DISCRETE ZINC INGOT: CONTAINER: ALUMINIUM INGOT:
Amount of Supply	NUMERICAL
Loading tonnage(in K.G)	NUMERICAL
Estimated Transportation time (per day)	NUMERICAL
Transportation time from the day of demand to the day of unloading	NUMERICAL
Transportation Cost (in Rials)	NUMERICAL
Transportation type	DISCRETE TWO VALUES RAIL TRANSPORT: 0 TRUCK TRANSPORT: 1
Weather condition	DISCRETE TWO VALUES TEMPERATE: 0 INTEMPERATE: 1
License permission	YES: 1 NO: 0

In this step, after describing the data with a diagram, if there are missing data, based on the number of missing data, the approach of deleting or replacing the lost data is selected. Also, the duplicate rows and extra variables in the software are checked.

4.1. Clustering

For analyzing transportation data, the first step is clustering and classification offload data. Then the model k-means clustering, is used. First, the number of parameters is entered manually, and the optimal number of clusters is determined with the evaluation parameter of Davis Boldin. In the analysis step, the optimal number of sets is evaluated with the output of the central table.

4.2. Extracting Data of Association Rules

The process of discovering the dependency rules is one of the fundamental approaches in modern data mining science to find directions and patterns in the database that researchers have highly regarded. In the data extraction step, the objective is to generate association rules from the data set to extract the relationships between features in Transportation.

The basic assumptions are as follows: Consider the set $I = I_1, I_2, I_3$ as a set of items.

Consider set D as a set of data that is, database transactions so that each transaction contains a collection of items; that is, each transaction T is a subset of I ($T \subseteq I$). Each transaction has an identifier called a TID. If A is a set of items, we say that transaction T contains A if and only if $A \subseteq T$.

An association rule is a statement in the form $A \Rightarrow B$, while $A \subseteq I$, $B \subseteq I$, and $A \cap B = \emptyset$. Criteria for the acceptability of association rules are two important parameters of the support and confidence of the rules.

The $A \Rightarrow B$ rule in the set of transactions D has support equal to s if s percent of transactions D include $A \cup B$, and this rule has the same reliability as c , if c percent of the transactions that contain A , also include B . In other words:

$$Support(A \Rightarrow B) = P(A \cup B)$$

$$Confidence(A \Rightarrow B) = P(B|A)$$

Rules that have a low Support or min-sup limit and low confidence or min-conf limit are called strong association rules, and the objective of all algorithms is such rules (or, as will be seen in some cases, a subset of these rules). For example, the following rule states that 70% of those who buy a TV and a video player and 80% of all transactions involve both the TV and the video player.

$$\begin{aligned} &Television \Rightarrow VCR \\ &[Support = 80\% \text{ Confidence} = 70\%] \end{aligned}$$

4.3. Data Extraction (Classification Algorithms)

After the classification step and the classification of the load data and analyzing the clusters, collective tree classification algorithms such as random forest and Naive Bayes Classifiers probabilistic method are used. The results and output rules are compared.

Figure 2: Model of cumulative algorithms

animals, or commodities from one place to another. In other words, to implement the classes, we have to divide the data into two categories: training and test. (Figure 2) Using the cross-validation method, we divide the data set into two classes:

training and examination. In the simple validation method in all stages of implementation, for example, 70% of the training data set and the rest can be used to test and calculate the indicators. Still, more innovatively, the cross-validation method was used, which was also used in this study, to make the whole data set both in training and in the test. In this type of validation, the data is subdivided into K subsets; of these K

subsets, one is used for confirmation and the other K-1 for training. This procedure is repeated K times, and all data is used exactly once for training and once for validation. Finally, the average result of this K validation load is chosen as a final estimate.

4.4. Criteria for Evaluating Classification Algorithms

This matrix shows how the classification algorithm works according to the input data set by classifying the classes of the classification problem. According to Table 6, the concepts of True Positive-TP, False Positive-FP, True Negative (TN), and False Negative(FN) are as follows.

Table 6: Confusion matrix

		Estimated records	
		Negative class (-)	Positive class (+)
Real records	Negative class (-)	TN	FP
	Positive class (+)	FN	TP

True Positive: This value indicates the number of records whose actual class is positive and the classifier has correctly identified their class as positive.

False Positive: This value indicates the number of records whose actual class is negative, and the classifier has incorrectly identified their class as positive.

True Negative: This value indicates the number of records whose actual class is negative, and the classifier has correctly identified their class as negative.

False Negative: This value indicates the number of records whose actual class is positive, and the classifier has incorrectly identified their class as negative.

The essential types of evaluation criteria for classification algorithms are now described. The most critical measure is the accuracy or rate of classification, which means that the classifier has correctly classified the percentage of the test records. This accuracy is calculated based on the concepts of the matrix according to the following relation:

$$Accuracy = \frac{Tp+TN}{Tp+TN+Fp+FN} \quad (1)$$

Figure 3: Classification model (data extraction)

Next, based on the definition of the confusion matrix and the parameters of total accuracy evaluation, the results are reported from the software output and then analyzed.

Finally, with the conclusion and analysis of clustering, the strategies in Transportation will be reported.

5. Analysis of Total Accuracy Data Mining Process Results

In this section, based on the Crisp process model, the steps of data comprehension, data visualization, data preparation, and wiping are described by running in the software.

Then the clustering and classification of the results are reported in the form of a table. Results of Optimal Cluster Analysis and the central plain of clusters are described at the end. Then, in the classification section, the two random forest and Naive Bayes methods are applied to the clustered data set. The accuracy of the models is reported and then evaluated.

The statistical information of the data set according to Figure 4 is as follows:

Name of data	Data type	Number of missing data	Max/min number of normal values	Max/min discrete of variables	Value of discrete variables
Demand date	Polynomial	0	96/4/31 (1)	96/1/6 (46)	96/1/6 (46), 97/12/28 (40)
Number of demands	Integer	0	1	18	1.7
Customer Nottage	Integer	0	5000	480000	500007.940
Location	Polynomial	0	RASHT (1)	SAHLAN (742)	SAHLAN (742), ZANJAN (425)
Destination	polynomial	0	DENIZLI (1)	VAN (532)	VAN (532), ANKARA (181)
Loading Type	Polynomial	0	WIRE (1)	Container (444)	Container (444), Zing Ingot (333)
Supply Number	Integer	2	1	18	1.7
Loading Cargo	Integer	0	4900	487070	47853.705
Estimate transportation time	Integer	0	3	9	4.6
Transportation time from loading day	Integer	0	3	12	5.4
Transportation Cost	Integer	0	28.74000	156.703	11221154.5
Transportation Type	Binomial	0	1 (649)	0 (739)	0 (739), 1 (649)

Figure 4: Statistical data

According to Figure 4 in the statistical information section, the first column is the name of the data. the second column is the data type. the third column is the number of missing data of each variable. the fourth and fifth columns are maxim and minimum of numerical variables and the minimum and maximum values of discrete variables and the last column shows the mean and standard deviation of numerical variables and the mode value of discrete variables.

There are only two rows of missing data in the supply column. The demand date variable is the solar data type, considered discrete variable in the software.

The amount of supply and demand ranges from 1 to 18, and the average is 1.698.

Loading tonnage is defined from 4900 kg to 487070 kg.

Index	Nominal value	Absolute count	Fraction
1	SAHLAN	742	0.535
2	ZANJAN	425	0.306
3	QAZVIN	104	0.075
4	TABRIZ	69	0.050
5	ISFAHAN	13	0.009
6	YAZD	13	0.009
7	QOM	4	0.003
8	TEHRAN	4	0.003

Index	Nominal value	Absolute count	Fraction
11	SALAFCHEGAN	2	0.001
12	ARAK	1	0.001
13	BANDAR EMAM	1	0.001
14	ESFAHAN	1	0.001
15	ESHTEHARD	1	0.001
16	HAMEDAN	1	0.001
17	KASHAN	1	0.001
18	RASHT	1	0.001

Figure 5: Frequency of origin cities in Transportation

According to Figure 5, 53% of Transportation from the original unit is from Sahlan city, 30% from Zanzan. The rest is defined with a much lower percentage in Qazvin, Isfahan, Yazd, Qom, Tehran, Ilam, Salafchegan, and Arak.

Index	Nominal value	Absolute count	Fraction
1	VAN	532	0.383
2	ANKARA	181	0.130
3	MERSIN	141	0.102
4	GAZIANTEP	94	0.068
5	MURATBEY	93	0.067
6	KONYA	80	0.058
7	ADANA	75	0.054
8	ISKENDERUN	53	0.038

Index	Nominal value	Absolute count	Fraction
10	DERINCE	33	0.024
11	IZMIR	18	0.013
12	KAYSERI	18	0.013
13	BURSA	6	0.004
14	ERENKOY	5	0.004
15	GEBZE	5	0.004
16	HATAY	4	0.003

Figure 6: Frequency of destination cities in Transportation

According to Figure 6, 38% of the destinations are frequently to Van, Turkey, 13% to Ankara, 10% to Mersin, 0.06% to Gaziantep, 0.06% to Muratbey, 0.05% to Konya, and the rest to Adana, Izmir, and other border cities of Turkey.

Estimated transportation time is defined as a minimum of 3 days to 9 days and an average of 4.61.

Transportation time is defined from the day of demand to unloading from a minimum of 3 days to 12 days.

The weather condition is defined in 232 lines in intemperate mode and 1156 lines in temperate mode.

46% of loads are in truck mode and 53% in rail mode.

96% of loads do not need a license for Transportation.

Information is defined from 2017 to 2019.

The maximum range of demanded tonnage and loaded tonnage is defined for truck transportation up to an average of 475,000 and for rail transportation only from 25,000 to 50,000 kg.

For rail transportation, only the city of Sahlan is the origin of Transportation, and other loadings are defined in truck transportation in other cities.

In rail transportation, only the cities of Van, Gaziantep, Mersin, and Iskenderun are transportation destinations, and the rest are related to truck transportation.

The maximum transportation time is from the day of demand to unloading, in a time interval of three days to 12 days for truck transportation and rail transportation up to an average of 5 days and a maximum of 8 days.

The highest transportation cost is related to truck transportation up to 1,500,000,000 Rials, and the difference in cost between rail and truck transportation is noticeable.

5.1. Pre-processing and Preparation of Data

5.1.1. Additional Variables

In the transport data analysis section, the first step in the clustering section is the need for numerical and continuous descriptive variables. Due to estimated time variables and the time interval between Transportation and unloading, the demand date feature, which is of the solar type, is left out of the data in modeling and extracting additional knowledge.

5.1.2. Deleting the Lost Value

As mentioned in the data description section, only two rows of data are lost in the supply column. One way is to delete a row with missing data. The disadvantage of this method is the removal of useful information in other features that have a value. One of the best ways is to replace the lost weight with the average value of the same column to maintain the standard deviation of the data. Therefore, in the software, the missing data alternative for the numerical supply variable is used.

Loading Type	Polynomial	0	WIRE (1)
Supply Number	Integer	2	1
Loading Cargo	Integer	0	4900
Estimate transportation time	Integer	0	3
Transportation time from loading day	Integer	0	3

Figure 7: Replacement of missing data

5.2. Results of K-Means Modeling (clustering)

In this step, the K-Means clustering algorithm is applied to the wiped data. The number of k (clusters) is simulated and evaluated based on Davis Boldin parameter in each step. The optimal number of clusters is based on the evaluation value and the analysis of the results section.

For clustering, the data should be converted to a numerical value

The results of the Davis Boldin evaluation parameter for the number of clusters 3 and 4 are reported and analyzed in the comparison section

Table 3: Comparison of two clustering methods

Model (K)	Davis Bouldin
K-means(k=3)	0.007
K-means(k=4)	0.005

The Davis Boldin evaluator reports the distance between clusters to the distance between clusters and obtains a number between zero and one. In this case, a successful model is when the samples in one cluster are closer and denser to each other and farther away from the samples in different clusters. It means that the closer Davis Boldin is to zero, the more successful the model and the more optimal the number of clusters. According to Table 4, the lowest value of the evaluator is with 4 clusters and a K-means algorithm. Therefore, in the next section, the results of 4-means clustering for classification are reported

5.2.1. K-means Clustering Analysis

In this step, according to the software results to analyze the clusters, the number of samples in each cluster is discussed. According to Figure 8, the number of samples is determined separately by clusters. In cluster zero, there are 1180 items,

equal to most of the samples; in cluster 1, there are only four items; in cluster two, there are 24 items, and in cluster three, 180 items.

Figure 8: Cluster sample separation

The following output (Figure 9) also indicates what cluster each load transaction is in.

This table with the 4-means method is as Table 7

Table 7: Central table of clusters based on numerical and qualitative indicators of Transportation

Transportation indexes	cluster0 score	cluster1 score	cluster2 score	cluster3 score
Transportation cost	72048463.2	1471449750	667131847.5	27227713 1.5
Transportation time from the day of demand to the day of unloading	5231	6750	6350	6661
Estimated Transportation time	4445	6250	5600	5271
Demanded tonnage	37751	195627	148278.1	82338.2
Amount of supply	1254	825	6150	3327
Demanded tonnage	409975	198000	147075	79833.3
Amount of demand	1256	8250	6150	3321

According to Table 7, the analysis of each cluster is as follows:

Cluster 1 (0.28%): loads transported in this group are the most expensive load in terms of transportation features compared to other clusters (up to 1 billion more), and the average loading tonnage is the highest, as well as the highest amount of supply and demand, is in this group.

100% of loads of this group do not require a license, and the type of Transportation of all transactions is 100% truck type and does not include rail transportation. The origin of all loads is Qazvin city and 75% to Adana city and 25% to Murat Bay city. The type of load is 75% rebar and 25% zinc ingot. The maximum transportation time from the day of demand to unloading is in loads of this cluster. Only four items with very different and unique features are in this group.

Cluster zero (frequency of 85.01%): The most loads (1180 items) are in this group, which on average, the lowest cost compared to other clusters and the shortest transportation time from the day of demand to the day of unloading are for loads of this group. 99% of rail loads are in this cluster and other groups, except for a tiny percentage in cluster 3, there are not rail loadings. But also, there is 37% truck loading. By far, the lowest amount of supply and demand is in this cluster. The origin of cities is 62% from Sahlan, and 26% from Zanzan, and the rest with a small percentage from Qazvin, Tabriz, and Yazd and there are destinations in almost all cities, but the place with most unloading is in Van, Turkey (42%). 37% of the load type is the container, and 22% is zinc ingot.

Cluster Two (1.72%): After Cluster 1, relatively high transportation costs are related to loads of this group. The highest loading tonnage after cluster 1 and the highest amount of supply and demand with an average of 6.1 is for items in this group. 0.05% of the loads often need a license for Transportation in this group, and 40% of the Transportation is in intemperate weather conditions. While in other clusters, most of the transportations were in temperate conditions. 41% of the loads were sent from Rasht, 29% from Qazvin, and 20% from Qazvin, and 33% of destinations were to Van and the rest to Adana and Ankara. 100% of the loading was of truck type.

Cluster 3 (13.07%): transportation costs in this group are average and less than clusters 1 and 2 and more than cluster zero. The difference between the index of this group and other groups is the difference in the estimated transportation time and the transportation time from the day of demand and the day of unloading, which is more than other groups. Loading cargo, supply and demand in this cluster is average, and 0.06% of the loads need a license. 23% have been in intemperate weather conditions. 0.06 of the loading is often of the rail type 56% of the loading is sent from Zanzan to all Turkish cities on average.

5.3. Effective Variables in Load Analysis

In the discussion of analysis and classification of clusters, the deviation diagram (Figure 10) examines the factors affecting the classification of loads in the transportation system.

Figure 10: Cluster deviation diagram

In this figure, a unique color line is provided for each cluster; which specifies the average data scores for each column in the corresponding cluster. The low score of the blue cluster (cluster zero) shows the load with the lowest cost.

The red line is related to cluster 3 (medium cost loads and average supply and demand).

The green line is related to cluster 2 (loads with relatively high cost).

Indigo line has the highest score related to cluster 1 (high-cost loads and high cargo, and so on).

However, the principal analysis of this diagram is separating the lines from the transportation time and transportation cost. The loading tonnage shows that these three factors are more effective in separating the clusters relative to the factors of weather conditions, type of load and origin, and destination.

5.4. Modeling (Association Rules)

To generate association rules and frequent items, discrete and binary features are selected with the fp-growth algorithm in Rapid Miner to generate frequent rules and articles that are repeated together, and the results are as follows:

Size	Support	Item 1	Item 2	Item 3
1	0.54	Sahlan = origin	-	-
1	0.5	TT	-	-
1	0.38	VAN = Origin	-	-
1	0.32	Container = LT	-	-
1	0.31	Zanjan = Origin	-	-
1	0.25	Zing Ingot = LT	-	-
1	0.15	WC	-	-
2	0.35	Sahlan = Origin	VAN = destination	-
2	0.32	Sahlan = Origin	Container = LT	-
2	0.306	LT	Zanjan = Origin	-
2	0.25	LT	Zing Ingot = LT	-
2	0.32	VAN = destination	Container = LT	-
2	0.193	Zanjan = Origin	Zing Ingot = LT	-
3	0.32	Sahlan = Origin	VAN = destination	Container = LT
3	0.193	LT	Zanjan = Origin	Zing Ingot = LT

Figure 11: Frequent transportation items

Figure 11 shows a section of the most frequently repeated items, based on the value of the most commonly repeated support = 70%, as follows:

1. 53% of the transportation system originated from the city of Sahlan, and 38% of the destinations where the city of Van.
2. 46.5% of the loading of truck transportation type.
3. 32% of the load transportation system was container type, and 24% was zinc ingots.
4. In 34% of Transportation, the origin was from Sahlan, and the destination was the city of Van.
5. In 32% of Transportation originating from Sahlan city, the type of load is the container.
6. In 30% of trucks, the loading was zinc ingots.
7. In 30% of the truck transportation, the origin was Zanjan city.
8. In 32% of destinations to the city of Van, the loading was of container type.
9. In 32% of the origin was from Sahlan, Van city's destination, and the type of load was the container.

With confidence = 70% of the association rules; If-then is as follows

Size	Premises	Conclusions	Support	Confidence
1	VAN = destination	Container = LT	0.32	0.84
2	VAN = destination	Container = LT , Sahlan = LT	0.32	0.84
3	VAN = destination	Sahlan = LT	0.34	0.89
4	VAN = destination , Sahlan = Origin	Container = LT	0.32	0.93
5	Container = LT	Sahlan = Origin	0.32	1
6	Zanjan = Origin	LT	0.306	1
7	Zing Ingot = LT	LT	0.24	1
8	Container = LT	VAN = destination	0.32	1
9	Container = LT	VAN = destination , Sahlan = Origin	0.32	1
10	Container = LT, Sahlan = Origin	VAN = destination	0.32	1
11	Container = LT , VAN = destination	Sahlan = Origin	0.32	1
12	Zing Ingot = LT , Zanjan = Origin	LT	0.193	1

Figure 12: Association rules

1. With 83.5% confidence, if the loading destination is the city of Van, then the type of load is the container.

2. With 83.5% confidence, if the destination of the loading is the city of Van, then the type of load is the container, and the origin is the city of Sahlan.

3. With 89.5% confidence, if the destination of the loading is the city of Van, then the origin is from the city of Sahlan.

4 .With 93.3% confidence, if the destination is the city of Van and the origin is the city, the type of load is a container.

5.With 100% confidence, if the origin is the city of Zanjan, the type of load is truck.

6 .With 100% confidence, if the load type is ingot zinc, the transportation type is truck.

7.With 100% confidence, if the load type is the container, the destination is the city of Van.

8.With 100% confidence, if the load type is the container, the destination is the city of Van, and the origin is the city of Sahlan.

9. With 100% confidence, if the load type is ingot zinc, the origin is Zanjan, and the transportation type is truck.

5.5. Naive Bayes Modeling Results (Data Extraction)

In this section, two classifiers of the Naive Bayes Classifiers probabilistic method and the collective tree of random forest method are applied to the data set with labeled clusters. Then the results of each class are being reported.

The data set (cluster output) is divided into training and test data into ten steps by cross-validation method. And the label variable (objective) is considered a cluster, a multi-value discrete variable (Nominal).

After dividing the samples, the Naive Bayes is used for classification.

On the right side of the software, the accuracy parameters specified in the output are identified. The results of parameters of total accuracy and recall are as follows after implementing:

Accuracy: 89.77%

The result of the confusion matrix is as follows:

The data set (cluster output) is divided into training and test data into ten steps by cross-validation method. And the label variable (objective) is considered a cluster, a multi-value discrete variable (Nominal).

After dividing the samples, the Naive Bayes is used for classification.

On the right side of the software, the accuracy parameters specified in the output are identified. The results of parameters of total accuracy and recall are as follows after implementing:

Accuracy: 89.77%

The result of the confusion matrix is as follows:

accuracy: 89.77% +/- 3.00% (p-value: 89.77%)

	true cluster_0	true cluster_3	true cluster_1	true cluster_2	class precision
pred cluster_0	107	6	0	0	99.42%
pred cluster_3	94	147	0	2	60.48%
pred cluster_1	0	0	2	1	66.67%
pred cluster_2	10	27	2	21	35.00%
class recall	91.19%	81.67%	50.00%	87.50%	

Figure 13: Results of Naive Bayes accuracy and confusion matrix

Class accuracy of cluster zero: 91.19%

Class accuracy of cluster 3: 81.67%

Class accuracy of cluster 2: 50%

Class accuracy of cluster 1: 87.50%

5.5.1. Random Forest Results

Random tree forest generates many decisions. To classify, it places a new object from an input vector at the end of each random forest tree.

Each tree gives us a classification, and it is possible to say that this tree "votes" for that class. The forest chooses the sort that has the most votes (among all forest trees).

Figure 14: Random Forest operator

a) Cluster accuracy 0: 99.75%

Hence the number of samples that are really in the cluster zero and are correctly predicted: 1177 samples

Hence the number of samples that are actually in cluster 0 and are mistakenly predicted in cluster 1 : 0 samples

Hence the number of samples that are actually in cluster 0 and are mistakenly predicted in cluster 2: 0 samples

Hence, the number of samples that are actually in cluster 0 and are mistakenly predicted in cluster 3: 3 samples

b) Cluster accuracy 3: 91.67%

Hence the number of samples that are actually in cluster 3 and are correctly predicted: 165 samples

Hence the number of samples that are actually in cluster 3 and are mistakenly predicted in cluster 0: 15 samples

Hence the number of samples that are actually in cluster 3 and are mistakenly predicted in cluster 2: 0samples.

Hence the number of samples that are actually in cluster 3 and are mistakenly predicted in cluster 1: 0 samples

c) Cluster accuracy 1: 100%

Hence the number of samples that are actually in cluster 1 and are correctly predicted: 4 samples

Hence the number of samples that are actually in cluster 0 and are mistakenly predicted in cluster 1: 0 samples

Hence the number of samples that are actually in cluster 0 and are mistakenly predicted in cluster 2: 0 samples

Hence the number of samples that are actually in cluster 0 and are mistakenly predicted in cluster 3: 0 samples

d) Cluster accuracy 2: 54.17%

Hence the number of samples that are actually in cluster 2 and are correctly predicted: 13 samples

Hence the number of samples that are actually in cluster 2 and are mistakenly predicted in cluster 1: 0 sample

Hence the number of samples that are actually in cluster 2 and are mistakenly predicted in cluster 3: 8 samples

Hence the number of samples that are actually in cluster 2 and are mistakenly predicted in cluster 0: 3 samples

6. Analysis of Results

The results of two decision tree and random forest models at the clustering output are summarized for analysis in Table 8:

Table 8: Classification Results

Model	Recall (cluster0)	Recall (cluster3)	Recall (cluster2)	Recall (cluster 1)	Accuracy
NB	91.19%	81.67%	87.50%	50%	89.77%
Random forest	99.75%	91.67%	54.17%	100%	99.91%

Looking closely at the results, it is evident that the accuracy of the collective random forest method, which uses ten decision trees for prediction, is 91.97% higher than the Naive Bayes probabilistic method. Because the ratio of samples in different clusters is very different and imbalance is observed in clusters, the accuracy of the models varies greatly in clusters. In the Naive Bayes method on cluster 1, which has only four items, there is low confidence of 50%.

But in other clusters, there is an average of 81%, which has a 15% error more than random forest. In general, in collective methods, due to the division of the data set, they are more accurate on imbalanced data and improve the results.

6.1. Data Analysis of Random Forest Trees

The decision tree is one method whose output is examined as analyzable rules, so it is a white-box method. And compared to black-box algorithms such as neural networks, the results can be studied as "if-then" rules. The relationship between factors can be analyzed in the field of Transportation. Some of the random forest tree models are mentioned in the classification section as follows:

Figure 15-Tree 1

Tree 1

204000<Demanded tonnage

|Transportation time from demand day to unloading day> 5.500

| |Demanded tonnage > 228000

| |Transportation time from demand day to unloading day>

7: cluster_0 {cluster_0 = 1,

cluster_3 = 0, cluster_1 = 1, cluster_2 = 0}

| |Transportation time from demand day to unloading day ≤ 7: cluster_1 cl cluster_0 = 0, cluster_3 = 0, cluster_1 = 3, cluster_2 = 0{

| | Demanded tonnage ≤ 228000: cluster_2 {cluster_0 = 0, cluster_3 = 0, cluster_1 = 0, cluster_2 = 2}

| Transportation time from demand day to loading day ≤ 5.500: cluster_3 {cluster_0 = 0, cluster_3 = 17, cluster_1 = 0, cluster_2 = 2}

Demanded tonnage ≤ 204000

| Demanded tonnage> 58750: cluster_3 {cluster_0 = 47, cluster_3 = 66, cluster_1 = 4, cluster_2 = 5}

| Demanded tonnage ≤ 58750

|Transportation time from demand day to unloading day> 11: cluster_3 {cluster_0 = 1, cluster_3 = 5, cluster_1 = 0, cluster_2 = 0}

| |Transportation time from demand day to unloading day ≤ 11: cluster_0 {cluster_0 = 1157, cluster_3 = 77, cluster_1 = 0, cluster_2 = 0}

Figure16-Tree2

Tree 2
cluster_1: 1141070000 <Transportation cost
{cluster_0=0, cluster_3=0, cluster_1=6, cluster_2=0}
1141070000 ≥transportation cost
|173113500 <transportation cost
| |519008850 <transportation cost :cluster_2
{cluster_0=0, cluster_3=0, cluster_1=0, cluster_2=13}
| |transportation cost ≤ 519008850: cluster_3 {cluster_0 = 0,
cluster_3 = 186, cluster_1 = 0, cluster_2 = 0}
| |transportation cost ≤ 173113500: cluster_0 {cluster_0 =
1183, cluster_3 = 0, cluster_1 = 0, cluster_2 = 0}

Figure 17: Tree 3

Tree 3

Loaded tonnage> 353165

|Loaded tonnage> 383360: cluster_3 {cluster_0 = 0,
cluster_3 = 9, cluster_1 = 0, cluster_2 = 0}

|Loaded tonnage ≤ 383360: cluster_2 {cluster_0 = 0,
cluster_3 = 0, cluster_1 = 0, cluster_2 = 7}

Loaded tonnage ≤ 353165: cluster_0 {cluster_0 = 1193,
cluster_3 = 157, cluster_1 = 1, cluster_2 = 21}

Figure 18: Tree 4

Tree 4

Amount of Demand > 14,500: cluster_2 {cluster_0 = 0,
cluster_3 = 2, cluster_1 = 0, cluster_2 = 3}

Amount of Demand ≤ 14,500

| Amount of Demand > 2,500

| | Demanded tonnage> 162750

| | | Estimated transportation time> 5,500

| | | | Estimated transportation time > 6,500: cluster_2
{cluster_0 = 0, cluster_3 = 1, cluster_1 = 0, cluster_2 = 2}

| | | | Estimated transportation time ≤ 6,500: cluster_1
{cluster_0 = 0, cluster_3 = 0, cluster_1 = 3, cluster_2 = 3}

| | | Estimated transportation time ≤ 5,500

| | | | Estimated transportation time> 4,500: cluster_0
{cluster_0 = 4, cluster_3 = 3, cluster_1 = 0, cluster_2 = 0}

| | | | Estimated transportation time ≤ 4,500: cluster_3
{cluster_0 = 4, cluster_3 = 5, cluster_1 = 0, cluster_2 = 5}

| | Demanded tonnage ≤ 162750: cluster_3 {cluster_0 = 30,
cluster_3 = 77, cluster_1 = 2, cluster_2 = 4}

| Amount of Demand ≤ 2,500: cluster_0 {cluster_0 = 1158,
cluster_3 = 79, cluster_1 = 0, cluster_2 = 3}

6.2. Data Extraction Analysis of results

The collected data set includes 1388 data related to the Transportation of various products of a transportation company in multiple types of rails and road. It exports

products including zinc ingots and aluminum ingots to Turkey. The data consists of 14 numerical and discrete columns. The features are the type of origin and destination, type of load, transportation cost, estimated transportation time and actual time, kind of Transportation, amount of supply, demanded tonnage, and loading cargo. The following information was extracted by describing statistical information and frequency charts:

The amount of supply and demand is from 1 to 18, and the average is 1,698.

Loading tonnage is defined from 4900 kg to 487070 kg.

53% of Transportation from the original unit is defined from Sahlan city, 30% from Zanjan. The rest have a much lower percentage from Qazvin, Isfahan, Yazd, Qom, Tehran, Ilam, Salafchegan, and Arak.

Estimated transportation time is defined as a minimum of 3 days to 9 days and an average of 4.61. Transportation time is determined from the day of demand to unloading from a minimum of 3 days to 12 days.

46% of loads are in truck type and 53% in rail type.

96% of loads do not require a license for Transportation, and the information is defined from 2017 to 2019.

The maximum range of demanded tonnage and loading tonnage is defined for up to an average of 475,000 truck transport and only from 25,000 to 50,000 kg for rail transport.

For rail transportation, only the city of Sahlan is the origin of Transportation, and other loads have been defined in truck transportation in other cities.

In rail transportation, only the cities of Van, Gaziantep, Mersin, and Iskenderun are destinations, and the rest are related to truck transportation.

The maximum transportation time from the day of demand to the day of unloading is three days to 12 days for truck transportation is up to an average of 5 days and a maximum of 8 days for rail transportation.

The highest transportation cost is related to truck transportation up to 1,500,000,000 Rials, and the difference in cost between rail and truck transportation is noticeable.

In the wiping data step, there are only two rows of missing data in the feature of the amount of supply, which instead of deleting rows that delete essential information from other columns, data substitution with the mean value of the column is used.

The variable of demand date is of solar-type, and due to the existence of time interval from the day of demand until the day of loading, there is no requirement to this feature and removed.

6.2.1.Data Extraction (clustering)

to optimize the number of clusters and the best amount of clustering, the Davis Boldin evaluator and clustering concepts were obtained; the smaller this evaluator, the more efficient the clusters, which means that the distance within clusters is less and the space within clusters smore. Based on Figure 5-1, there were the best clustering by K-means with $k = 4$.

Figure 19: Comparison of clustering evaluator with Davis Boldin

In the following, the results of K-means clustering classification are examined in terms of the number of samples per cluster, the central mean table of clusters; and the clusters are analyzed in terms of transportation concepts; Figure 20.

Figure 20: Frequency chart of transported loads

- Cluster 1 (costly truckloads) with only four items had the maximum amount in terms of transportation cost, loading tonnage per kilogram, supply and demand, full transportation time from the day of request to unloading were placed separately from other clusters in a group. All truckloads were sent from Qazvin to Adana, Turkey, and Murat Bay.

- Cluster zero (the lowest cost with the most rail loads): the highest number of loads (85%) which on average had the minimum amount in terms of transportation cost, loading tonnage per kilogram, amount of supply and demand, and the maximum transportation time from the day of demand to unloading.

Most of these loads are mainly of rail type, 37% of which are containers and the rest are zinc ingots, is from the city of Sahlan, and on average, loads are sent to all cities in Turkey.

- Cluster 2 (relatively expensive truck loads in intemperate weather conditions): After cluster 1, the highest transportation cost, highest loading tonnage after cluster 1, and the highest amount of supply and demand with an average of 6.1 are for items in this group with 24 samples.

The difference between this cluster and Cluster 1 is that 40% of the loads are in intemperate weather conditions, mainly sent to Van, Adana, and Ankara.

- Cluster 3 (average indicators in terms of Transportation with the most extended transportation time interval): Transportation costs in this group are average less than clusters 1 and 2 and more than cluster zero. The difference between the index of this group and other groups is the difference in the

time interval between the estimated transportation time and the transportation time from the day of demand and the day of unloading, which is more than other groups. In fact, on average 180 loads, mainly of zinc ingot type, have been sent to all Turkish cities from Zanjan.

In the data visualization section, the deviation diagram is used. By analyzing this diagram, it was found that transportation time, transportation cost, and loading cargo and weather conditions have a more significant impact than other factors in the separation of type of Transportation.

6.2.2 Data Extraction (Association Rules)

First, the repetitive patterns are created with fp-growth method and minimum confidence of 70%, and based on the repetitive patterns; nine association rules are extracted on quality indicators, which are as follows:

1. With 83.5% confidence, if the load destination is the city of Van, then the type of load is container type.
2. With 83.5% confidence, if the destination of the load is the city of Van, then the type of load is of the container type, and the origin is from the city of Sahlan.
3. With 89.5% confidence, if the destination of the load is the city of Van, then the origin is from the city of Sahlan.
4. With 93.3% confidence, if the load destination is the city of Van and the origin is from the city, the type of load is the container.
5. With 100% confidence, if the origin is the city of Zanjan, the transportation type is truck transport.
6. With 100% confidence, if the type of ingot load is zinc, the transportation type is truck transport.
7. With 100% confidence, if the type of load is the container, the destination is the city of Van.
8. With 100% confidence, if the type of load is the container, the destination is the city of Van, and the origin is the city of Sahlan.
9. With 100% confidence, if the load type is ingot zinc, the origin is Zanjan, and the kind of Transportation is truck transport.

6.2.3. Data Extraction (Classification)

The following results are reported from the analysis and comparison of the accuracy of the classifiers in Figure 21:

Figure 21: Accuracy of classifiers

Looking closely at the results, it is that the collective random forest method, which uses ten decision trees for

prediction, with 91.97% accuracy, has higher accuracy than the Naive Bayes probabilistic method. Then, the ratio of samples in different clusters is very different, so the imbalance is observed in clusters. The accuracy of the models varies greatly in sets. In the Naive Bayes method, cluster 1, which has only four items, has low confidence of 50%. But in other clusters, there is an average of 81%, which has a 15% error more than random forest. Due to the imbalanced data, the rule generation, white box, and hybrid random forest method were used. This method balances the data due to the division of the data set between 10 decision trees and reports accuracy on all classes with a very high value. The transportation cost variable being a parent in most trees indicates the impact of this variable on clustering. Some of the generated and valuable rules of random decision trees and forests are as follows:

1. If the transportation cost is more than one billion, the loads are in cluster 1 (high-cost, high-demand truckloads).
2. If the transportation cost is less than 173 million, the load is in the cluster with low cost and low rail demand.
3. If the transportation cost is more than 519 million, it is in cluster 2. If transportation cost is less than 519 million and more than 200 million, it is in cluster 3. (Direct impact of transportation cost feature).
4. If the loading cargo is less than 363165 kg, it is in cluster zero, and if it is more than 383360 kg, it is in cluster 3.
5. If the demanded cargo is less than 58750 kg and the transportation time from the day of demand to unloading is less than 11 days, it is in cluster zero.
6. If the demanded cargo is more than 203,000 kg and the transportation time is more than five days, it is in cluster 2.
7. But if the cargo is more than 228,000 and the transportation time is more than seven days, it is in cluster 1.
8. If the amount of demand is less than 2.5, it is in cluster zero.

7. Conclusions and Future Suggestions

In this paper, three data mining techniques of clustering, association rules, and classification were used to investigate the factors affecting the cost and time of road and rail transportation and data extraction. In the first step, by clustering method based on the K-means distance of 1388 transactions, road and rail transport were divided into four optimal clusters. The results showed that by far, transportation cost, cargo, and the amount of supply and demand features are the most practical features in cluster separation. The weather and type of transport and kind of load had a negligible effect on the classification of transport loads. In clusters zero and one, transportation costs and supply and demand were significantly different from other groups. In cluster 2, the intemperate weather conditions, and cluster 3, the time intervals from the day of demand to unloading caused the difference between a cluster and other clusters. In the association rules phase, nine rules with confidence above 85% were reported on the type of load and destination's qualitative variables. In the classification step, two Naive Bayes

(probabilistic and collective random forest methods) were used on the clustering output.

The results showed that the random forest with a total accuracy of 97.91% has much more accurate results than Naive Bayes for cluster classification. This method also proved the importance of collective (voting) algorithms to improve the accuracy of imbalanced data. With random forest trees, eight proper rules were extracted. The rules proved that in the parent of most trees, transportation cost and loading tonnage are the criteria for comparing clusters and tree generation, which confirmed the importance of these factors in the road and rail transportation system classification.

In general, if researchers want to propose development on the results of this study in the future, they can consider the following: 1) using other clustering methods such as SOM (self-organizing map) and fuzzy clustering. 2) Using methods to optimize the number of clusters, such as meta-heuristic genetic algorithm and particle swarm optimization.

REFERENCES

- [1] Olson, D. 2003, "Managerial Issues of Enterprise Resource Planning", McGraw-Hill/Irwin, 1 edition, PP: 45.
- [2] Han, J., Kamber, M., 2011, "Data Mining: Concepts and Techniques", San Diego Academic Press, pp: 25
- [3] David J. HAND, 2001, "Data Mining: Statistics and More?" The American Statistician. Vol. 52, No. 2, pp. 112-113
- [4] Herbert A. E. 1996. "Introduction to Data Mining and Knowledge Discovery". Two Crows Corporation, Third Edition, pp. 5-10.
- [5] The CRISP-DM process model, 1999, <http://www.crisp-dm.org/>
- [6] Fayyad, U., Piatetsky-Shapiro, G., Padhraic, S., 2008, "From Data Mining to Knowledge Discovery in Databases." <http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf>.
- [7] Ramageri M. Data mining techniques and applications. Indian Journal of Computer Science and Engineering. 2011; 4(1): 301-5.
- [8] Hu R. Medical data mining based on association rules. Computer and Information Science. 2010; 3(4): 104-8.
- [9] Kremic, E. and A. Subasi (2015). "Performance of Random Forest and SVM in Face Recognition." The International Arab Journal of Information Technology
- [10] C.-S. Li and M.-C. Chen, "A data mining-based approach for travel time prediction in freeway with non-recurrent congestion," Neurocomputing, vol. 133, pp. 74-83, 2014.
- [11] Zhang G Peter. 2003. "Time series forecasting using a hybrid ARIMA and neural network model". Neurocomputing 50. pp 159-175.
- [12] Zhang G Peter, Min Qi. 2005. "Neural network forecasting for seasonal and trend time series". European Journal of Operational Research .Vol 160. pp 501-515
- [13] Lathia, N., Froehlich, J., Capra, L., 2010, "Mining Public Transport Usage for Personalised Intelligent Transport Systems". In IEEE ICDM.
- [14] Adams S.O., Akano R.O., Asemota O, J. 2011. "Forecasting Electricity Generation in Nigeria using Univariate Time Series Models". European journal of Scientific Research. Vol. 58. No. 1. pp. 30-37.
- [15] Feyza Gürbüz, Fatma Turnab. "Rule extraction for tram faults via data mining for safe." Transportation Research Part A. 2018.
- [16] Jithin Raj, Hareesh Bahuleyan, Lelitha Devi Vanajakshi. 2016. "Application of data mining techniques for traffic density estimation and prediction"
- [17] Milad Ghasri, Taha Hossein Rashidi, S. Travis Waller. "Developing a disaggregate travel demand system of models using." Transportation Research Part A. 2017.
- [18] Soyoung Jung, Xiao Qin, Cheol Oh. "Improving strategic policies for pedestrian safety enhancement." Transportation Research Part A. 2016.
- [19] Hana Ševčíková, Adrian E. Raftery, Paul A. Waddell. "Uncertain benefits: Application of Bayesian melding to the Alaskan." Transportation Research Part A. 2011.
- [20] Albert Y. Chen, Ting-Yi Yu. "Network based temporary facility location for the Emergency Medical Services considering the disaster

induced demand and the transportation infrastructure in disaster response." Transportation Research Part B. 2016.

[21] Yuxuan Ji, Nikolas Geroliminis. "On the spatial partitioning of urban transportation networks." Transportation Research Part B. 2012.

[22] Mehrdad Kargari, Mohammad Mehdi Sepehri. 2012. "Store's clustering using a data mining approach for distributing automotive spare-parts to reduce transportation costs"

[23] Venkata Siva Reddy and R. Vasanth Kumar Mehta. 2019 "Study on Computational Intelligence

Approaches and Big Data Analytics in Smart Transportation System"

[24] Wirth, R. and J. Hipp. CRISP-DM: Towards a standard process model for data mining. in Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. 2000. Citeseer

[25] Ismaili, M., 2012, Concepts and techniques of data mining, Niaz Danesh Publications,

[26] Ahmadvand, A.M., Minaei Bidgoli, B., Akhundzadeh, A. "Citizen Satisfaction Analysis Using Data Mining Techniques: A Case Study of Tehran Municipality". Third Data Mining Conference. 2019

[27] Rastegar, N., Alizadeh, S., Ghazanfari, M. "Application of Clustering Method in Selecting Export Strategies of Goods in Iran". The Second Iranian Data Mining Conference

[28] Zahra Mohammad Beigi. 2014. Application of RFID in order to control and reduce urban accidents with data mining approach

[29] Alireza Kianian. 2016. Public Transportation Service Quality Assessment and Improvement Actions Selection using Data Mining and Quality Function Deployment (Case study: Mashhad Bus Transportation System)

Mahdi Yousefi Nejad Attari was born in Bonab Iran. He is an assistant professor in Department of Industrial Engineering at Bonab branch of Azad University Iran.

He received his Ph.D. Degrees in Department of Industrial Engineering from the University of Qazvin branch of Islamic Azad. His research activities include Blood supply chain and data mining techniques.

Babak Ejlaly was born in Gilan Iran in 1983. He received the B.S. degree in Mazandaran University in Industrial Engineering in 2016 and M.S degree in Amir Kabir University of Tehran in The Same Major in 2011.

He worked on management of planning sources and row material waste in Iran Tobacco Company in 2008 to Now. He has also researched the fields of neural network and data mining as a researcher at Koshlar University of Guilan.

Hannane Heidarpour was born in Tabriz, Iran in 1993. She received the B.S. degree in Tabriz University in Industrial Engineering in 2016 and M.S degree in Seraj University of Tabriz in The Same Major in 2020. In 2016 She worked on management of new energy sources along with energy obtained from municipal waste. In 2019 To 2020 She Worked on Data Mining Field. Eventually She implemented Data Mining on Transportation to classify the factors that affect in Transportation to find Out

How Data Mining can help to Optimize the International Transportation Cost a time, Although This is New Science That needs Lots of Researches but hope This Article can help.

Ali Ala, was born in Mazandaran, Iran. He received his M.Sc. in Industrial Engineering from Wuhan University and Technology in China. He graduated with his Ph.D. in Industrial Engineering from Shanghai Jiao Tong University (SJTU), focusing on Healthcare Optimization, Operation Research,

and Supply Chain System.