ISSN (Online) 2456-1290

International Journal of Engineering Research in Mechanical and Civil Engineering (IJERMCE)

Volume 11, Issue 7, July 2024

Transportation Master Plan Analysis for Trip Generation Attributes Through Association Rules: The Case of City of Sakarya, Türkiye

[1] Mine POLAT ALPAN*, [2] Zeliha Çağla KUYUMCU, [3] Mustafa TANIŞ, [4] Hakan ASLAN

[1] Research Assistant, Faculty of Engineering, Department of Civil Engineering Zonguldak Bülent Ecevit University, Türkiye
[2] Research Assistant, Faculty of Engineering, Department of Civil Engineering Sakarya University, Türkiye
[3] Assistant Professor, Faculty of Engineering, Department of Civil Engineering Zonguldak Bülent Ecevit University, Türkiye
[4] Assistant Professor, Faculty of Engineering, Department of Civil Engineering Sakarya University, Türkiye

Email: [1] minepolat@beun.edu.tr, [2] caglacaglar@sakarya.edu.tr, [3] tanis@beun.edu.tr, [4] haslan@sakarya.edu.tr

Abstract— Data mining can be defined as the extraction of unclear and previously unknown but potentially usable information and patterns from a large dataset. This technique allows for uncovering relationships among existing data and making predictions for the future when necessary. This study aims to analyse the data, by utilizing related information of individuals, obtained from transportation master plan survey studies of city of Sakarya, Türkiye. Through the apriori algorithm from data mining program WEKA, the relationships and rules were obtained for the events that are likely to occur together. In this regard, the daily transportation and trip related data of 9876 individuals were identified and listed. The results revealed the impact of the variables such as age group, gender, education, employment status, driver's license ownership, household income, and type of vehicle on the number of trips done by the people in the study.

As being one of the data mining techniques, association rule analysis was applied to the data from the survey conducted in order to obtain the related rules and probabilities generated. Such studies are likely to be further developed to lead more accurate and widespread assessments in decisions related to transportation planning.

Index Terms— Data mining, Transportation master plan, Association rules, Apriori algorithms

I. INTRODUCTION

The complex transportation systems of today's cities constantly require more efficient and sustainable solutions due to increasing population and advancing technologies. In this context, urban transportation master plans play a vital role in evaluating the existing transportation infrastructure and services, as well as in developing strategic plans for the future. The transportation master plan prepared for the city of Sakarya, in this sense, involves a comprehensive analysis and planning process.

Data mining methods enable the extraction of meaningful information and patterns from large data sets. This technique significantly contributes to make critical decisions in transportation planning and forecasting future trends. In this study, stated survey data were analysed by using the apriori algorithm from WEKA to derive various relationships and rules related to the attributes of the applicants and their travel pattern habits.

II. DATA MINING

The primary focus of data mining is to extract implicit, previously unknown but potentially usable information and patterns from data set [1]. Data mining is a multidisciplinary field acting as a bridge between many technical areas; including database technology, statistics, artificial intelligence, machine learning, pattern recognition, and data visualization [2].

Association rules, one of the data mining methods, were first introduced by Agrawal et al., within the concept of market basket analysis and have been further used and developed by other studies [3]. Quantitative association rules which can be used with specific modifications to detect possible erroneous data items are particularly interested in this research [4,5].

The environments where data obtained and stored from transactions in e-businesses, operational systems, and many other different ways are called "*data warehouses*." These quantitative or qualitative data reflecting facts about people, objects, transactions, applications, and events are then transformed into meaningful information by employing data mining techniques to use them in the strategic decision-making process [6].

Information may be regarded as processed data [7]. After its collection, data is classified and summarised with regard to mathematical methods and presented graphically in terms of statistical measures such as mean, mode, standard deviation to determine if there are relationships among various variables. Inferential knowledge is produced from the meaningful information obtained from the data, with the aim of making future predictions and transforming them into actions. Figure 1 presents the visual functioning of data mining.

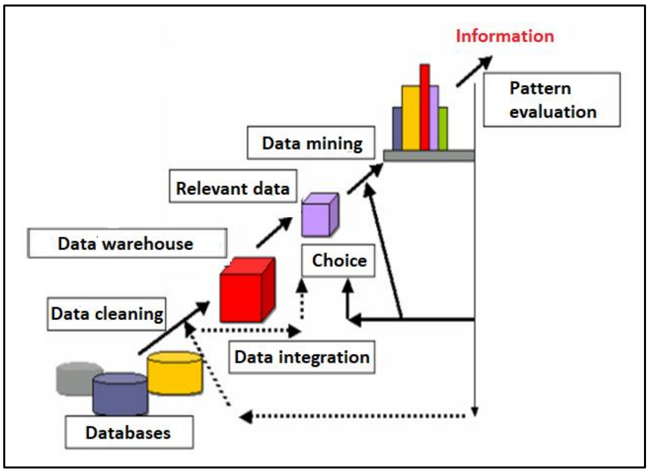


Figure 1. Data Mining Process [8].

Various techniques are used in data mining:

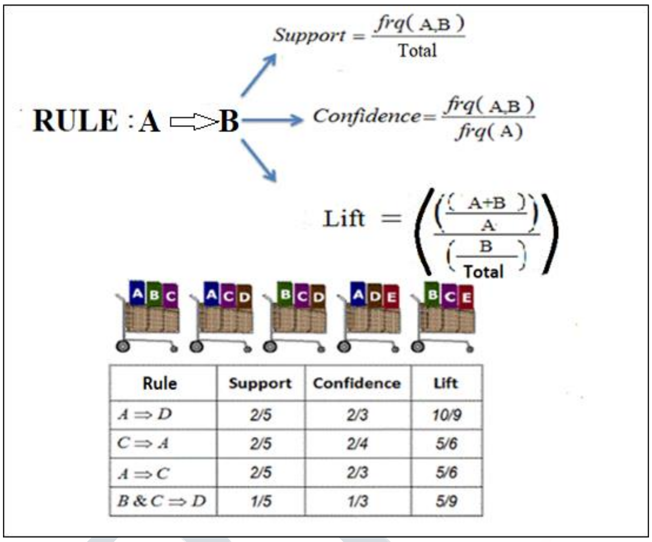
- Classification and Regression
- Clustering
- Association Rules
- Memory-Based Methods
- Artificial Neural Networks (ANN)
- Decision Trees [9].

Among these, the association rules have been considered within the scope of this study. The aim here is to identify rules for events that are likely to occur together. To generate these rules, support and confidence values are utilized aiming to identify frequent associations above the minimum support and minimum confidence values specified by the user.

III. METHOD

In this study, transportation data based on a household survey conducted in the city of Sakarya has been examined. The relationships between variables such as age group, gender, education level, employment status, driver's license ownership, household income level, travel frequency, and mode of transportation were evaluated using three different analysis methods. The findings provide valuable insights for transportation planning and contribute to determine future transportation strategies more effectively.

Association rule analysis, one of the commonly used data mining methods, involves discovering rules, relationships, and associations among the data investigated. Interesting relationships and simultaneous occurrences among data objects are subject to be revealed in this way [10]. Association rules are generated to satisfy user-defined minimum support and confidence threshold values. Figure 2 illustrates the rules and explanations related to the analysis.



$$Support = \frac{frq(A,B)}{Total}$$

$$Confidence = \frac{frq(A,B)}{frq(A)}$$

$$Lift = \left(\frac{\left(\frac{A+B}{A}\right)}{\left(\frac{B}{Total}\right)}\right)$$

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Figure 2. Association Analysis Rules and Explanations [11].

As shown in Figure 2, if the association rule between item sets A and B is represented as "A→B":

- Support: (A → B) = (number of rows containing both A and B) / (total number of rows),
- Confidence: The confidence value of the AB association rule is the percentage of transactions containing A and B together based on known A transaction.
- Lift: It is one of the most important concepts used to evaluate the benefit provided by a model.

3.1. Used Data Set

The data set, acquired from the household survey consisting of 7051 rows and 8300 columns, was evaluated, and the parameters to be used within the scope of this study were identified. During the data preparation phase, the data were filtered and refined regarding the parameters to be used in the study. Missing, excessive or unnecessary data, and the data deemed unsuitable for the analysis in the study were removed from the data set. After this rigorous data processing, transportation data for 9786 individuals was determined and listed for use in the study. The following parameters are planned to be used in the data set for the study:

- Age group (values ranging from 12 to 100)
- Gender (Male and Female)
- Education (Illiterate, Primary School, Middle School, High School, University, Master, Doctorate)
- Employment status (Employed, Unemployed)
- Driver's license ownership (Yes, No)
- Household income level (Various income brackets)
- Number of trips (Values ranging from 1 to 12 trips per day)
- Mode of transportation (Different modes of transportation)

IV. DATA ANALYSIS AND FINDINGS

In the apriori algorithm, rules are generated and ranked based on the co-occurrences of parameters in household surveys. A support value of 15% was determined as the threshold value among a total of 9786 records. After the minimum lift ratio was set to 1, the algorithm was applied. For instance, in an application aimed at predicting traffic accidents, injuries occurred in 35 out of 100 accidents identified by the model, and if out of a randomly selected 100 accidents, injuries occurred in only 5, then the lift ratio was calculated as 7 [12]. Three different analyses were conducted to generate association rules using the apriori algorithm through the parameters used in these analyses as shown in Table 1.

Table 1. Parameters Used in the Analyses.

	Analysis 1	Analysis 2	Analysis 3
Parameters	Trip count	Trip count (without 2)	Trip count
	Age	Age	Age
	Gender	Gender	Gender
	Education Level	Education Level	Education Level
	Occupational status	Occupational status	Driver's licence ownership
	Driver's licence ownership	Driver's licence ownership	Household income
	Household income	Household income	
	Mode of travel	Mode of travel	

As indicated in Table 1, different parameters were evaluated for each analysis. Descriptions of the conducted analyses are provided sequentially.

4.1. Analysis 1

The first analysis presented and interpreted in the study involves the inferences obtained from different association scenarios.

- Employment status "unemployed" and mode of transportation "pedestrian" → gender "female" (79%, lift 2.05).

It was found that 79% of the individuals who stated in the household survey that they do not work and generally prefer walking as the mode of transportation are female, with a lift value of 2.05.

- Gender "female" and household income "1250 TL and below" → employment status "unemployed" and driver's license ownership "none" (81%, lift 1.95).

It was determined that the females with a household income of 1250 TL and below not possessing a driver's license and being unemployed represent 81%, with a lift value of 1.95.

- Trip count "2" and gender "female" → employment status "unemployed" and driver's license ownership "none" (74%, lift 1.77).

It was found that females having 2 trips per day, being unemployed and not having a driver's license signify 74%, with a lift value of 1.77.

- Gender "female", employment status "unemployed", and household income "1250 TL and below" → driver's license ownership "none" (91%, lift 1.73).

Unemployed females with a household income of 1250 TL and below have a 91% probability of not possessing a driver's license, with a lift value of 1.73.

- Employment status "employed" and driver's license ownership "yes" → gender "male" (92%, lift 1.49).

It was determined that employed individuals possessing a driver's license are 92% likely to be male, with a lift value of 1.49.

4.2. Analysis 2

The outcomes obtained through the second analysis conducted to present the inferences obtained from different association scenarios are given below.

- Trip count "4" and gender "male" → driver's license ownership "yes" and employment status "employed" (73%, lift 1.64).

The likelihood of males who make 4 trips a day possessing a driver's license and being employed is 73%, with a lift value of 1.64.

- Trip count "4" and employment status "unemployed" → driver's license ownership "none" (70%, lift value 1.52).

Individuals making 4 trips a day but stating they are unemployed have a 70% probability of not possessing a driver's license, with a lift value of 1.52.

- Trip count "4" and employment status "employed" and driver's license ownership "yes" → gender "male" (94%).

Individuals making 4 trips a day, being employed, and having a driver's license are 94% likely to be male.

- Gender "female" → employment status "unemployed", and driver's license ownership "none" (71%, lift 1.96).

The probability of females being unemployed and not possessing a driver's license is 71%, with a lift value of 1.96.

- Age range "40-64" → education "primary school" and household income "1250 TL and below" (38%, lift 1.34).

Participants aged between 40 and 64 years with household income of 1250 TL or below have a 38% probability of primary school education, with a lift ratio of 1.34.

4.3. Analysis 3

The third analysis conducted presents and illustrates the inferences obtained from different association scenarios.

- Age range "18-39" and gender "male" → driver's license ownership "yes" (74%, lift 1.58).

It was determined that males aged between 18 and 39 have a 74% probability of possessing a driver's license, with a lift value of 1.58.

- Mode of transportation "pedestrian" → driver's license ownership "none" (82%, lift 1.54).

Participants who stated they travel daily as pedestrians have an 82% probability of not possessing a driver's license, with a lift value of 1.54.

- Trip count "2" and gender "female" → driver's license ownership "none" (84%, lift 1.59).

Female participants who make 2 trips a day have an 84% probability of not possessing a driver's license, with a lift value of 1.59.

- Gender "female" → driver's license ownership "none" (83%, lift 1.57).

Female participants have an 83% probability of not possessing a driver's license, with a lift value of 1.57.

V. RESULTS AND DISCUSSION

This study utilizes association rules analysis through apriori algorithm on household survey data collected for transportation master plan in order to unveil prevalent patterns aiming to provide realistic insights into city dynamics and optimisation of public welfare so that urban decision-making process is based upon accurate approach. The incorporation of such data, alongside methodologies like data mining, proves pivotal in gauging community dynamics.

Key findings from various analyses reveal that individuals identified as non-working and predominantly walking are mainly female (79%). Additionally, females reporting an income of 1250 TL or less exhibit an 81% probability of being non-license holders and non-workers. Similarly, females undertaking two daily trips exhibit a 74% probability of being non-workers and non-license holders. Moreover, females with income level below 1250 TL and non-working status have a staggering 91% likelihood of being non-license holders. Conversely, those identified as employed and license holders are predominantly male (92%).

Further analyses indicate that males making four daily trips have a 73% likelihood of holding a driver's license and being employed, while those not working exhibit a 70% likelihood of being non-license holders. Moreover, those making four trips daily while employed and holding a license are 94% likely to be male. Females, on the other hand, exhibit a 71% probability of being non-workers and non-license holders, while participants aged 40 to 60 with primary education and incomes below 1250 TL have a 38% probability of being within this category.

Additionally, males aged 18 to 39 have a 74% likelihood of possessing a driver's license, while those traveling as pedestrians exhibit an 82% probability of being non-license holders. Similarly, females undertaking two daily trips have an 84% likelihood of being non-license holders. Furthermore, it should be stated that females in general exhibit an 83% probability of being non-license holders.

As this study examines the number of trips with regard to ages, genders, educational statuses, employment statuses, license ownerships, and household income levels among Sakarya residents, it sheds light on their proportional influences. On the other hand, enhanced iterations to be carried out soon in the future are thought to set up more accurate decision-making process in transportation planning.

REFERENCES

[1]. Fian Tabea, T. (2021). *From black spots to black patterns: Pattern recognition with road traffic accident data. Illustrated with single-vehicle accidents with a single occupation and personal injury that occurred outside the built-up area on the Austrian road network between 2012 and 2019.*

[2]. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques Third Edition.* Elsevier. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

[3]. Rakesh Agrawal, Tomasz Imielinski, & Arun Swami. (1993). *Mining Association Rules between Sets of Items in Large Databases.*

[4]. Srikant, R., & Agrawal, R. (1996). *Mining Generalized Association Rules* (s. 26)

[5]. Korn, F., Labrinidis, A., Kotidis, Y., & Faloutsos, C. (1998). *Ratio Rules: A New Paradigm for Fast, Quantifiable Data Mining.*

[6]. Witten D. (2011). *Data Mining: Practical Machine Learning Tools and Techniques.* Elsevier. https://doi.org/10.1016/C2009-0-19715-5

[7]. Capurro, R., & Hjørland, B. (2003). *The concept of information.* Annual Review of Information Science and Technology, 37(1), 343-411. https://doi.org/10.1002/aris.1440370109

[8]. Fayyad, U. (1996). *From Data Mining to Knowledge Discovery in Databases.*

[9]. Maimon, O., & Rokach, L. (Ed.). (2005). Data mining and knowledge discovery handbook. Springer.

[10]. Yurtay, Y., Yurtay, N., Çelebi, N., Bacınoğlu, N. Z. Ve Ak, G. *Sakarya İline Ait Yangın Verilerinin Veri Madenciliği Yöntemleriyle Değerlendirilmesi.* ISITES 2014.

[11]. Nensi Kansagara. *Concept Description and Association Rule Mining.* https://www.amirajcollege.in/wp-content/uploads/2020/06/2170715-chapter-5-concept-description-and-association-rule-mining.pdf

[12]. Akpınar, H. *Veri Tabanlarında Bilgi Kesfi ve Veri Madenciliği.* İ.Ü. İsletme Fakültesi Dergisi, 2000; 29, (1), s: 1-22.