AUT Journal of Modeling and Simulation

AUT J. Model. Simul., 55(2) (2023) 275-282
DOI: 10.22060/miscj.2024.22687.5340

Analysis of the gasoline consumption on an international scale: A data mining approach

Ali Tavakoli Kashani[1]*, Zahra Sartibi[1], Seyed Ali Ziaee[2], Mahdi Khorasani[3]

[1] School of Civil Engineering, Iran University of Science and Technology, Tehran, Iran.
[2] Department of Civil Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.
[3] Amirkabir University of Technology, Tehran, Iran.

ABSTRACT: The transportation sector accounts for a significant portion of global energy consumption and, gasoline is a major fuel consumed in road transport. On the other hand, the excessive consumption of gasoline can lead to an increase in unnecessary trips and road accidents. This study aims to determine the impact of macroscale factors on gasoline consumption. In this regard, we investigated the effect of gasoline price, oil reserves, and income level on gasoline consumption per capita in about 90 countries, including Iran, over a period of 17 years. Also, Rail and Air travel per capita and membership in the Organization for Economic Co-operation and Development (OECD) were considered as control variables. For this purpose, one of the classification techniques utilized in the area of data mining, Classification, and Regression Tree (CART), was employed. The variable importance measure (VIM) was calculated to quantify the association of each independent variable with the target variable. The results indicated that oil reserves, gasoline prices, and average income have a normalized significance of 100, 58.5, and 30.3 % respectively. Other variables do not have significant importance. So, higher per-capita gasoline consumption is exclusively involved in oil-rich countries such as Iran. Therefore, considering national oil reserves should be prioritized when comparing fuel consumption across world countries. Also, more expensive gasoline would relatively diminish its use. However, this effect of the gasoline price is mostly confirmed for countries with lower national oil reserves which often have higher prices than their counterparts.

Review History:

Received: Sep. 22, 2023
Revised: Mar. 15, 2024
Accepted: Apr. 25, 2024
Available Online: May, 30, 2024

Keywords:

Gasoline Consumption
Gasoline Price
Oil Reserves
Classification and Regression Tree (CART)
Variable Importance Measure (VIM)

1- Introduction

The transportation sector contributes to a sizable portion of global energy consumption, accounting for about 25-30% in Iran and other countries [1]. Furthermore, only about 10% of gasoline resources are allocated to uses other than fuel for the transportation sector. Excessive use of gasoline and other fuels has not only reduced national reserves but has also accelerated the rate of greenhouse gas production [2]. Also, the results of research revealed that increased gasoline consumption results in a greater number of traffic fatalities [3]. In this regard, Identifying the factors that cause irregular fuel consumption is an effective strategy for tackling the issue.

According to the United States Energy Information Administration data, gasoline consumption per capita varies significantly across countries. And, gasoline consumption in Iran is roughly nine times that of Turkey whereas the populations of these countries are nearly equal [4].

In the review of related research, the effect of gasoline prices, oil reserves, and income have been mentioned. Scholars have argued that higher gasoline prices result in greater gasoline usage [5–8]. With rising gasoline prices, individuals may avoid unnecessary long-distance trips, slow their acceleration, or switch to public transportation such as the Rail [9]. However, additional factors must also be considered. Burke and Nishitateno discussed the presence of oil reserves [10]. In addition, income has a positive effect on gasoline consumption [11].

Using data collected from about 90 countries, including Iran, from 2000 to 2016[1], gasoline consumption per capita was investigated in relation to price per liter of gasoline, income level, having oil reserves, rail and air travel per capita, and membership in the Organization for Economic Cooperation and Development (OECD) during the years.

Classification and Regression Tree (CART), a non-parametric data mining technique, was used to unearth hidden relationships between independent and dependent variables, as well as the relative importance of each. To the best of our knowledge, the data mining strategy was not utilized as a method in the relevant publications. For the purpose of developing a CART model, gasoline consumption per capita, the dependent variable, was categorized as either low or high. Noteworthy, in contradiction to crash severity studies, this

*Corresponding author's email: alitavakoli@iust.ac.ir

1 2016 is the last year that gasoline price data are available for the countries.

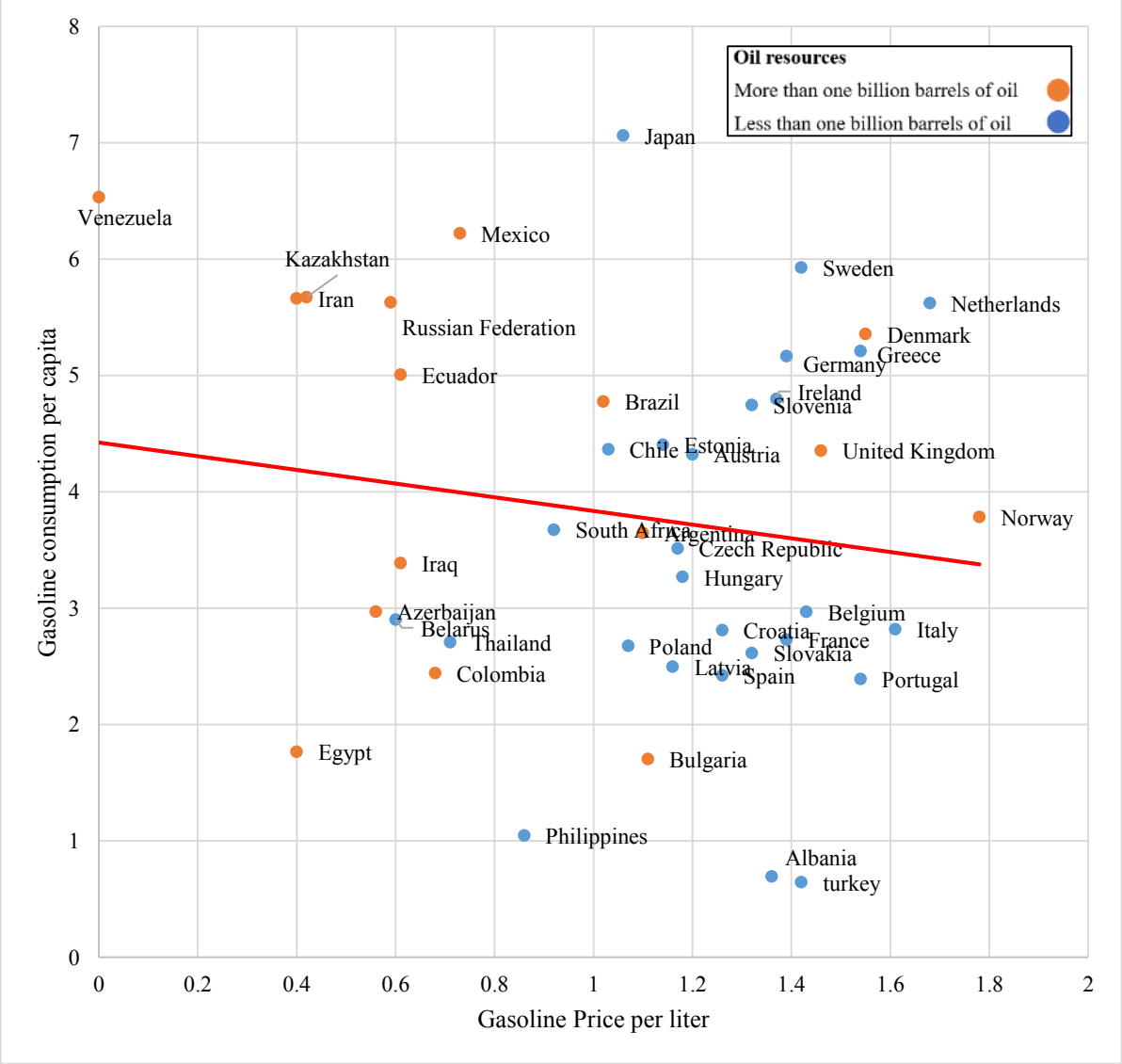

Fig. 1. Gasoline consumption per capita and Gasoline price per liter.

Notes: The data in this figure are for 41 countries in 2016 [4, 15].

approach was not commonly applied in this field of study [11–14].

Figure 1, which is based on 2016 statistics, depicts the negative relationship between the price and gasoline consumption. According to this graph, countries having oil reserves, such as Iran, Venezuela, and Russia, have high consumption and relatively lower gasoline prices.

The primary objective of this study is to determine the impact of macroscale factors on gasoline consumption per capita. Taking into consideration the negative effects of excessive oil consumption as a nonrenewable finite resource, a data mining technique is provided in this work to investigate the critical indicators of fuel use. Consequently, the purpose of this study is to evaluate the impact of gasoline price per liter on gasoline use. In addition, we will examine the reliability of the population as the sole indication of fuel usage. The findings of this study will contribute to the formulation of fuel consumption regulation policies.

2- Method

The dependent variable, gasoline consumption per capita, was classified as either high or low usage. Regarding the categorical character of the dependent variable in the present study, Classification and Regression Trees (CART), a decision tree approach, were utilized to estimate the relative relevance of each variable in relation to gasoline consumption per capita.

Extending the modeling strategy, a decision tree is

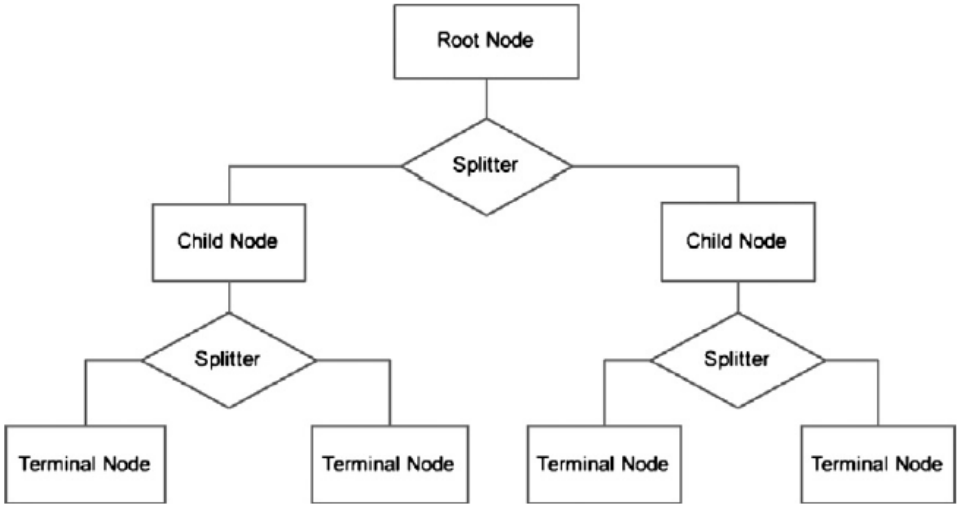

Fig. 2. Decision tree structure

a supervised learning technique with a tree structure at each node of which a yes-or-no question is addressed. Consequently, a decision tree is a decision- or rule-based algorithm, as opposed to probabilistic classification methods. Decision tree models may discover and clearly explain the complex patterns associated with crash risk and do not need to provide a functional form [11]. Decision trees, along with neutral networks, Bayesian networks, support vector machines, and instance-based categorization, are classified as predictive techniques within the data mining paradigms [16].

Classification and Regression Trees (CART) can be considered a standard data mining methodology. CART is a nonparametric model in which there are no predetermined relationships between the target variable and predictors. CART can utilize both categorical and continuous variables as model inputs or outputs [12].

In comparison to other methodologies such as ANNs and SVMs, DTs are capable of graphically exhibiting the internal structures, making them easy to deploy and the results comparatively clear [17]. A crucial step, pruning, is included in the development of DTs to prevent the model from being overfitting. Pruning is the process of removing branches that have a negligible effect on the performance of DT prediction models [18].

As depicted in Figure 2, the CART method is used to develop the classification tree based on the CART principle. To begin with, all the data is centralized at the tree's root. This so-called "root node" is then divided into two child nodes based on a predictor variable (splitter) that optimizes the homogeneity (purity) of the two child nodes. This procedure is repeated for each child node until each node contains the most homogeneous set of data possible. In other words, the underlying principle of tree growth is to recursively partition the target variable to minimize "impurity" in the terminal nodes. Gini's criterion is the most prevalent measure of node impurity [12].

The Gini index is used to quantify homogeneity by calculating the proportion of data that belongs to a class. The Gini criteria are used to assess homogeneity by calculating the proportion of data belonging to a certain class. The Gini index is formulated as follows

$$g(t) = \sum_{i \neq j} p(j|t)p(i|t) \tag{1}$$

Where the two dependent variable categories are characterized as i, and j.

$$p(j|t) = \frac{p(j,t)}{p(t)} \tag{2}$$

$$p(j,t) = \frac{\pi(j)N_j(t)}{N_j} \tag{3}$$

$$p(t) = \sum_j p(j,t) \tag{4}$$

$\pi(j)$ represents the prior probability value for category j, $N_j(t)$ is the number of observations in the j^{th} category for node t, and N_j is the number of category j records for the root node. In the calculation of $N_j(t)$ and N_j, only records with valid

Table 1. Variable description

Variable	Defined Levels
Gasoline consumption per capita	1- High (10-31), 2-low (0-10)
Gasoline price	1- High (1.64-2.54), 2- middle-high (1.12-1.64), 3- middle-low (0.63-1.12), 4- low (0-0.63)
Rail travel per capita	1- High (1427-2405), 2- middle (514-1427), 3- low (0-514)
Air travel per capita	1- High (3-30), 2- middle (0.9-3), 3- low (0-0.9)
Oil reserves	1- Have more than one billion barrels of oil, 2- Have less than one billion barrels of oil
Income level	1- High, 2- middle-high, 3- middle-low, 4- low
OECD membership	1- Yes, 2- No

values for the split predictor in node t and the root node are taken into account, respectively [11,19].

The prior probability indicates the proportion of each class in the population, but if their prior probabilities are additionally changed based on the proportion of each class in the training data, the final model will predict all of the data in the dominant class, hence improving the model's overall accuracy. Due to the imbalanced frequency of data relating to a class, the forecast accuracy of the less frequent level will diminish. In circumstances where levels of target variables have an imbalanced percentage but the same significance in terms of prediction accuracy, it has been proposed to assign equal prior probabilities so that those with a lesser proportion can also be included in predictions. Although the model's overall accuracy declines, the forecast accuracy of the data with the smallest proportion rises, which is in most circumstances more significant to decision-makers [16].

The following equation defines the relative significance of each variable X (with h levels) in the model.

$$VIM(X) = \sum_{i=1}^{h} \frac{nx_i}{n} (I(C|X = x_i) - (c)) \quad (5)$$

In relation to Equation (1), "I" is the Gini index. Also, n shows the total number of observations. Additionally, C and nx_i represent the class of the dependent variable (gasoline consumption per capita) and the number of observations where the considered class is x_i, respectively [11,20].

3- Data

The target variable is Gasoline consumption per capita. National motor gasoline consumption statistics were obtained from the United States Energy Information Administration (EIA) [21]. To determine gasoline consumption per capita, daily motor gasoline consumption (thousand barrels) was divided by a country's total population (converted to million people). Addressing the suggested primary independent variable, gasoline price, we obtained data from the German Agency for International Cooperation (GIZ) regarding gasoline pump prices per liter (in US dollars) [15].

We additionally integrated road transportation competitors into the modeling procedure. Consequently, we considered the rail travel per capita (passenger-km divided by population) and air travel per capita (total passengers divided by population) as reported by the International Civil Aviation Organization Statistics of the World (ICAO) and the International Union of Railways (UIC), respectively [22-24]. In accordance with the most recent categorization by the WHO[1] (based on 2010 GDP numbers translated to billions of US dollars, 2018), the influence of income level on per capita was also examined. As previously noted, it is hypothesized that oil reserves (at least one billion barrels) and OECD[2] membership both influence gasoline use per capita [25]. Notable is that the bulk of OECD countries is developed European, North American, and Asian nations. The data is restricted to the years 2000 to 2016, as gasoline prices are unavailable in the majority of countries beginning in 2017.

Due to the characteristics of the Classification and Regression Tree modeling technique, the k-means clustering algorithm was used to classify the variables involved in this research. Table 1 describes the classification of independent variables. It should be noted that 70% and 30% of observations were included in the training and test datasets, respectively.

1 World Health Organization

2 Organization for Economic Co-operation and Development

Table 2. CART model assessment result

Sample	Observed	Predicted		
		high	low	Percent Correct
Training	high	83	37	69.2%
	low	41	938	95.8%
	Overall Percentage	11.3%	88.7%	92.9%
Test	high	27	13	67.5%
	low	19	355	94.9%
	Overall Percentage	11.1%	88.9%	92.3%

Table 3. Importance of variables (VIM)

Independent Variable	Importance	Normalized Importance
Oil reserves	.079	100.0%
Gasoline price	.046	58.5%
Income level	.024	30.3%
Air travel per capita	.006	7.3%
OECD membership	.004	4.9%
Rail travel per capita	.003	3.9%

4- Results

The Gini index was utilized as the tree growth method in this investigation. According to the suggested classification, 11% of countries have high gasoline consumption per capita, whereas 89% have low gasoline consumption per capita. Due to the disproportionate number of observations in the two classes, as advised by a number of earlier research, the prior probabilities were set to be equal [11,26]. Despite the fact that this action decreases the overall precision, it increases the predictive ability of the class with a significantly lower proportion of observations [27].

According to Table 2, the model's prediction accuracy is acceptable for both the training and test sets of observations. In addition, more than 67% of all courses were accurately predicted.

Figure 3 and Table 3 detail the normalized significance of the independent variables. Also, figure 4 displays the output of the calibrated decision tree for this research.

The results indicate that having oil reserves has the greatest influence on gasoline consumption per capita, roughly doubling the significance of the second most influential variable, gasoline price. This result is consistent with the findings of a number of other researchers who have also concluded that the gasoline price is not the most significant indicator of gasoline consumption [3]. Scholars have also argued that any increase in the price of gasoline reduces its demand [6]. Lastly, the factors of air travel per capita, rail travel per capita, and OECD membership are of lesser significance. It appears that the increase in travel via competing modes will reduce the proportion of road transport, resulting in a decrease in fuel consumption.

5- Conclusion

Gasoline is one of the most precious resources on earth, and only a few countries have access to it. Nevertheless, excessive fuel usage is a significant contributor to air pollution, traffic congestion, and automobile accidents. Fewer articles in the transportation area have examined the factors influencing gasoline use on a global scale.

The CART technique may clearly illustrate the links between the independent factors and the dependent variable due to the presentation of visually appealing findings.

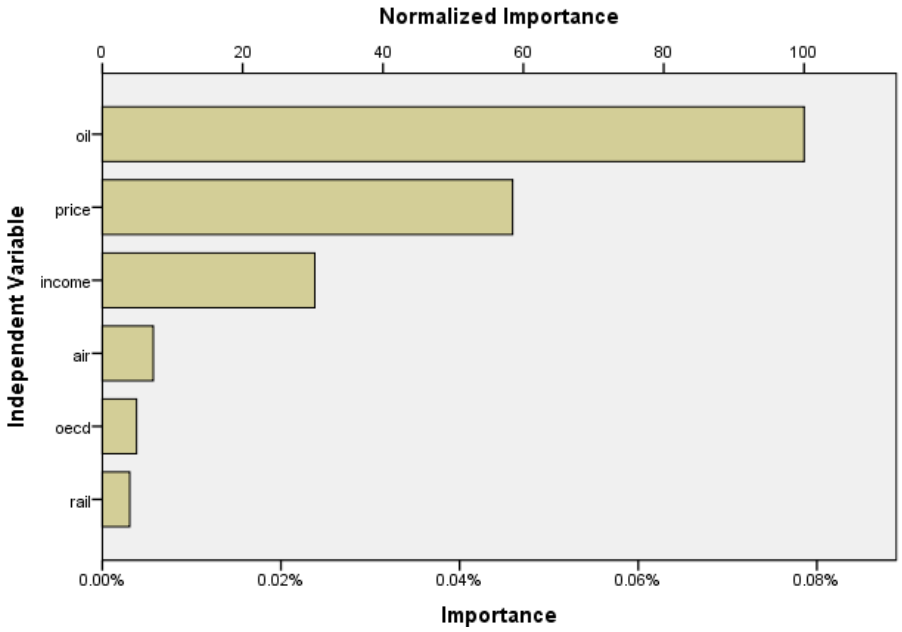

Fig. 3. Normalized importance of Variables

Additionally, the relative significance of variables is a helpful outcome of this approach which was used in this research.

Based on the results of the method used, oil reserves, gasoline prices, and average income have a normalized importance of 100, 58.5, and 30.3 percent respectively. While the normalized importance of other variables is less than 10 percent. Therefore, the exitance of national oil reserves is a very influential factor, and comparing countries based on this factor should precede comparisons based on gasoline pricing or average income in terms of gasoline consumption.

In countries such as Iran that own large oil reserves are accompanied by higher gasoline consumption than in other countries. On the other hand, in countries with lower national oil reserves which already have higher prices than their counterparts, more expensive gasoline can decrease its consumption.

Any increase in rail and air travel as alternatives to road travel would decrease the proportion of road traffic and gasoline usage. However, our results did not validate this notion, since any growth in the use of alternative means of transportation, particularly in developing countries, it can also indicate an increase in the number of travelers.

References

[1] M. González-Torres, L. Pérez-Lombard, J. F. Coronel, I. R. Maestre, and D. Yan, "A review on buildings energy information: Trends, end-uses, fuels and drivers," Energy Reports, vol. 8, pp. 626-637, 2022.

[2] L. Chapman, "Transport and climate change: a review," J Transp Geogr, vol. 15, no. 5, pp. 354–367, 2007.

[3] A. Tavakoli Kashani and zahra Sartibi, "Gasoline price, Gasoline consumption and road fatalities," The 17th international conference on transportation and traffic engineering. Tehran, Iran, 2018. (In Persian)

[4] U. S. E. I. Administration, "Motor Gasoline consumption by country," 2000-2016. 2019.

[5] P. J. Burke and S. Nishitateno, "Gasoline prices, gasoline consumption, and new-vehicle fuel economy: Evidence for a large sample of countries," Energy Econ, vol. 36, pp. 363–370, 2013.

[6] P. J. Burke and S. Nishitateno, "Gasoline prices and road fatalities: International evidence," Econ Inq, vol. 53, no. 3, pp. 1437–1450, 2015.

[7] H. H. Jafari and A. Baratimalayeri, "The crisis of gasoline consumption in the Iran's transportation sector," Energy Policy, vol. 36, no. 7, pp. 2536–2543, 2008.

[8] S. E. West and R. C. Williams III, "The cost of reducing gasoline consumption," American Economic Review, vol. 95, no. 2, pp. 294–299, 2005.

[9] A. Tavakoli Kashani and Z. Sartibi, "Is There a Relationship Between Rail Transport and Road Fatalities?" Iranian Journal of Science and Technology, Transactions of Civil Engineering, pp. 1–10, 2022.

[10] H. Mahmood, N. Maalel, and M. S. Hassan, "Probing the energy-environmental Kuznets curve hypothesis in oil and natural gas consumption models considering urbanization and financial development in Middle East

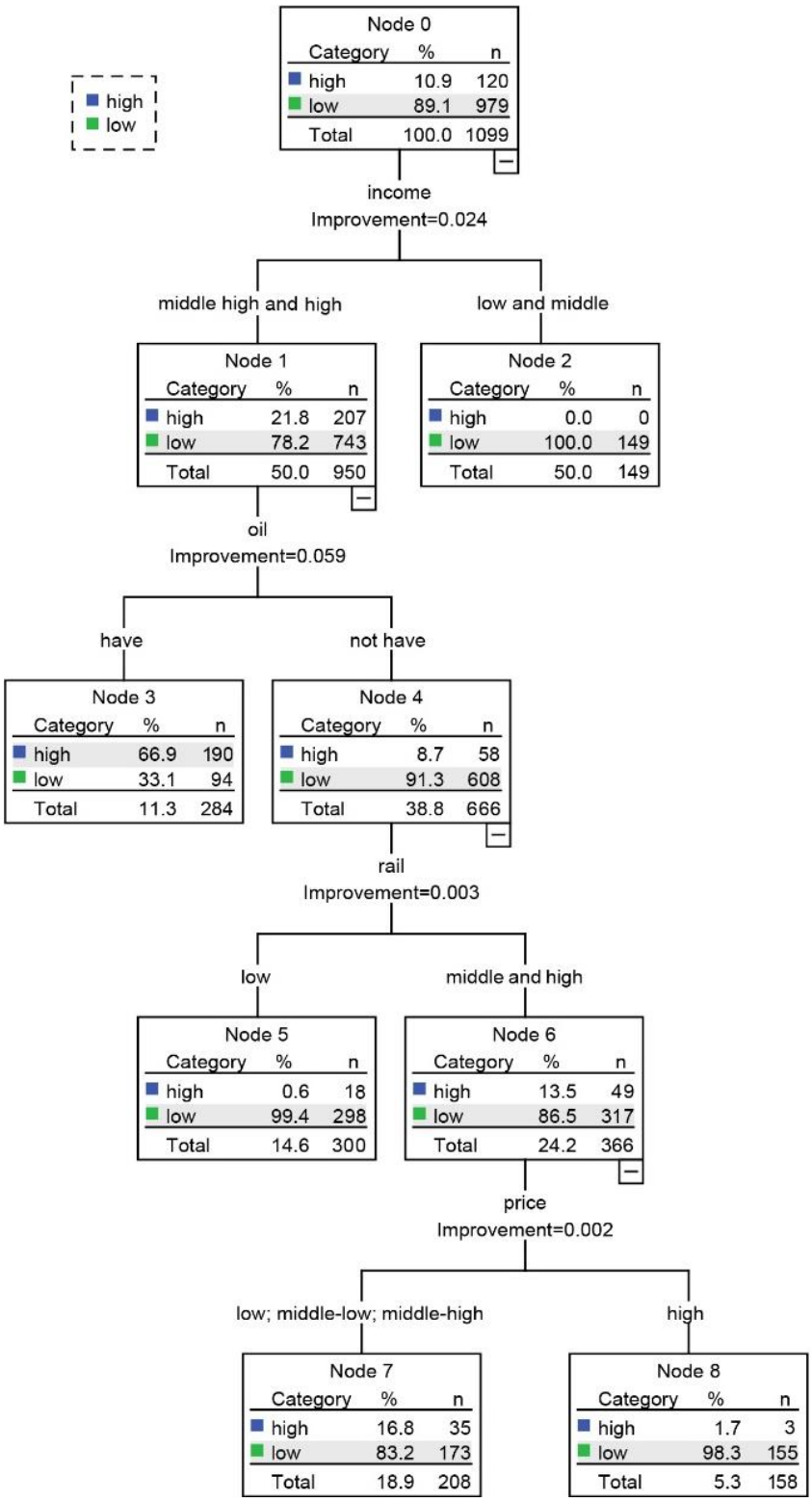

Fig. 4. Decision tree

countries," Energies (Basel), vol. 14, no. 11, p. 3178, 2021.

[11] A. T. Kashani and A. S. Mohaymany, "Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models," Saf Sci, vol. 49, no. 10, pp. 1314–1320, 2011.

[12] A. Tavakoli Kashani, R. Rabieyan, and M. M. Besharati, "A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers," J Safety Res, vol. 51, pp. 93–98, Dec. 2014, doi: 10.1016/j.jsr.2014.09.004.

[13] S. Krishnaveni and M. Hemalatha, "A perspective analysis of traffic accident using data mining techniques," Int J Comput Appl, vol. 23, no. 7, pp. 40–48, 2011.

[14] L. Li, S. Shrestha, and G. Hu, "Analysis of road traffic fatal accidents using data mining techniques," in 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), 2017, pp. 363–370.

[15] German Agency for International Cooperation (GIZ), "International Fuel Prices."

[16] Lior Rokach and Oded Maimon, Data mining with decision trees: theory and applications. 2015.

[17] D. Delen, L. Tomak, K. Topuz, and E. Eryarsoy, "Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods," J Transp Health, vol. 4, pp. 118–131, Mar. 2017, doi: 10.1016/j.jth.2017.01.009.

[18] L.-Y. Chang and H.-W. Wang, "Analysis of traffic injury severity: An application of non-parametric classification tree techniques," Accid Anal Prev, vol. 38, no. 5, pp. 1019–1027, Sep. 2006, doi: 10.1016/j.aap.2006.04.009.

[19] X. Wen, Y. Xie, L. Jiang, Z. Pu, and T. Ge, "Applications of machine learning methods in traffic crash severity modelling: current status and future directions," Transp Rev, vol. 41, no. 6, pp. 855–879, Nov. 2021, doi: 10.1080/01441647.2021.1954108.

[20] P. Treeratpituk and C. L. Giles, "Disambiguating authors in academic publications using random forests," in Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries, Jun. 2009, pp. 39–48. doi: 10.1145/1555400.1555408.

[21] United States Energy Information Administration (EIA), "Open Data."

[22] International Civil Aviation Organization (ICAO), "Statistics."

[23] International Union of Railways (UIC), "Statistics."

[24] World Bank, "World Bank Open Data."

[25] Organisation for Economic Co-operation and Development (OECD), "OECD Data."

[26] D. Steinberg and M. Golovnya, "CART 6.0 user's guide," Salford Systems, San Diego, CA, 2007.

[27] A. V. Veettil and A. K. Mishra, "Quantifying thresholds for advancing impact-based drought assessment using classification and regression tree (CART) models," Journal of Hydrology, vol. 625, p. 129966, 2023.

HOW TO CITE THIS ARTICLE

A. Tavakoli Kashani, Z. Sartibi, S. A. Ziaee, M. Khorasani, Analysis of the gasoline consumption on an international scale: A data mining approach, AUT J. Model. Simul., 55(2) (2023) 275-282.

DOI: 10.22060/miscj.2024.22687.5340