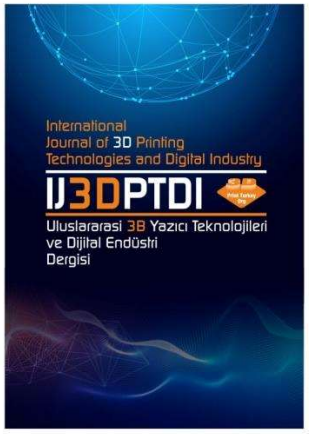

ULUSLARARASI 3B YAZICI TEKNOLOJİLERİ
VE DİJİTAL ENDÜSTRİ DERGİSİ

INTERNATIONAL JOURNAL OF 3D PRINTING
TECHNOLOGIES AND DIGITAL INDUSTRY

ISSN:2602-3350 (Online)

URL: https://dergipark.org.tr/ij3dptdi

DATA MINING AND MACHINE LEARNING APPROACHES IN DATA SCIENCE: PREDICTIVE MODELING OF TRAFFIC ACCIDENT CAUSES

Yazarlar (Authors): Taner Ersöz *, Filiz Ersöz

Bu makaleye şu şekilde atıfta bulunabilirsiniz (To cite to this article): *Ersöz T., Ersöz F., "Data Mining and Machine Learning Approaches In Data Science: Predictive Modeling of Traffic Accident Causes" Int. J. of 3D Printing Tech. Dig. Ind., 6(3): 530-539, (2022).*

DOI: 10.46519/ij3dptdi.1199614

Araştırma Makale/ Research Article

Erişim Linki: (To link to this article): https://dergipark.org.tr/en/pub/ij3dptdi/archive

DATA MINING AND MACHINE LEARNING APPROACHES IN DATA SCIENCE: PREDICTIVE MODELING OF TRAFFIC ACCIDENT CAUSES

Taner Ersöz[a]*, Filiz Ersöz[b]

[a]Karabük University, Business Faculty, Department of Actuary and Risk Management, KARABÜK
[b]Karabük University, Engineering Faculty, Department of Industrial Engineering, KARABÜK

* *Corresponding Author:* tanerersoz@karabuk.edu.tr

(**Received:** 04.11.2022; **Revised:** 07.12.2022; **Accepted:** 28.12.2022)

ABSTRACT

Today, the increase in the number of vehicles causes an increase in traffic accidents, an increase in loss of life and property, and potential risks. Analytical models are presented to investigate the socio-economic, demographic and temporal effects of the factors affecting the level of injury resulting from traffic accidents. By examining the data of various traffic accidents and developing a model, the factors and hazards affecting traffic accidents can be determined by data mining and machine learning approaches. The aim of this study is to determine which classification techniques are important for analyzing traffic accidents and to find out the factor that affects traffic accidents among the variables used in the research. The "Random Forest" algorithm, which gives the best model result among the techniques used in the research, was found. Weather conditions were found to be the most important factor among the factors that lead to traffic accidents, followed by the age and education of the driver. This study presents a traceable approach in terms of revealing the differences between data mining and machine learning under the umbrella of data science and following the processes with an application related to traffic accidents.

Keywords: Traffic Accident, Data Mining, Machine Learning, Classification Algorithms, Random Forest, Naive Bayes, Gradient Boosted, Tree Ensemble, Knime.

1. INTRODUCTION

The increase in global technologies and the amount of data in the world allows us to obtain meaningful and valuable information from the data. International Data Corporation (IDC) and Statista estimate that around 74 zettabytes of data will be produced in 2021 [1]. People, institutions and organizations can keep the data they produce in data storage systems on low costs. They can easily access the data they produce and easily disseminate them over internet networks. However, these data do not make sense alone. It is beneficial when the data is processed and it gives out valuable information.

Today, a lot of knowledge is gained by using data-based methods (Information engineering, data science, data mining, business analytics, etc.) such as identification, comparison, revealing relationships and modeling. This information is now under analysis according to the concepts of "Data science and data analytics". Data science is the discipline of explaining the past and predicting the future through data analysis. This purpose is valid for many disciplines. However, it is necessary to point out the differences. The relationship of data science with statistics, data mining and machine learning is briefly described under the following subtitles. In addition, a comparison between the disciplines of data mining and machine learning is also available.

2. DATA SCIENCE, STATISTICS AND DATA MINING

Data science or data scientist today involves the study and use of data in the production or service operation to make informed business decisions. Data science is related to a number of

disciplines such as statistics, mathematics, data mining, machine learning, and many others. Today, a data scientist must have expertise in data mining, data analysis, computer programming, statistics, machine learning, data visualization, and big data analytics [2].

Within scope of the data science, data mining and statistics is about extracting information from the data, discovering the deeper form in the data, identifying, and defining the relationships. Although both sciences are similar, data mining, unlike statistics, makes use of big data and databases in revealing valuable information, and the techniques it uses as a tool are very different from the discipline of statistics. It discovers knowledge by the way of collecting, tabulating and analyzing the statistical data. It has a more effective use to represent the population, especially with the help of samples. However, data mining is defined as a process in revealing valid, new, and potentially useful trends and patterns in data [3] and as a discipline that allows revealing meaningful relationships/rules from big and meaningless data [4].

2.1. Machine Learning and Data Mining

In analytical forecasting applications, statistics, data mining and machine learning are related concepts. Machine learning, as a term, is referred as statistical machine learning in some studies [4-6] and is an important tool in the development of modern software. However, machine learning is a subgenre of artificial intelligence that provides predictive results without human intervention.

Machine learning and data mining are concepts that are inspired by each other and although they have common points, they have some differences, as well. These differences are given in Table 1 [7].

Table 1. Differences between data mining and machine learning

Data Mining	Machine Learning
Data mining is learning from big data, discovering information and finding hidden relationships and patterns. Big data, which includes both structured and unstructured data types, is the data that is analyzed and classified, transformed into a meaningful and actionable form.	Machine learning works on a concept that machines develop from existing data and self-learning. Past experiences and big data are important. The algorithms used are based on mathematics and programming languages and use data mining methods and algorithms to create predictive models.
Data mining is about extracting rules from data. Data mining is based on databases and mostly statistical methods.	Machine learning includes an algorithm that learns from data and improves it automatically.
Data mining uses many machine learning techniques, but often has logically different goals.	Machine learning uses data mining techniques such as the preprocessing step to improve learner accuracy.
Data mining uses many other techniques besides machine learning algorithms. Data mining can use a machine learning algorithm as a tool, but data mining also uses statistics as another tool to extract something from raw data.	Machine learning algorithms can be used in the data mining process. In general, the classification algorithms are mostly the same. These algorithms can be used in both data mining and machine learning.
Data mining uses machine learning tools to uncover meaningful and valuable information.	Machine learning uses computational methods used in data mining tasks such as clustering and classification.
Some algorithms of machine learning such as clustering and classification are used in data mining tasks.	Machine learning predicts analytical results with Artificial Intelligence (AI). It uses algorithms automatically.
Data mining delves deep into data to extract useful and meaningful information to the user.	Machine learning brings the machine closer to perfection by developing complex algorithms and working iteratively with a trained dataset.
Data mining needs human factor while discovering information from data. Data mining produces predictive model results with human-defined data that does not come automatically. But when the variable	Data mining serves as an input to machine learning. Machine learning algorithms work continuously and automatically to improve model performance. It can also analyze when any error might occur. Human labor is very

changes, the model changes and cannot be used forever.	low. Once applied, it can be used forever.
In order to increase the model performance in the data mining process, the size of the data and the cleaning of the data and making it ready for the model are very important. Outliers in the data reduce the accuracy of the model or increase model errors. In addition, in data mining analysis, taking the relevant variables into the model and establishing the right model suitable for the data and the problem increases the reliability of the results.	In machine learning, it uses self-learning algorithms to improve model performance, and machine learning is result-oriented. They are self-training systems that make meaningful determinations by simulating the data and parameters you have presented. To find the optimum value for each parameter and adjust these parameters to increase the accuracy of the model, you need to have a good understanding of their individual effects on the model.
Data mining is not capable of self-learning. It follows predefined guidelines. You need to follow the steps in data analysis and it responds to a specific problem.	Machine learning algorithms can self-identify, change their rules according to different situations, and solve the problem automatically.
Data mining uses an existing dataset (such as a data warehouse) to find meaningful relationships and patterns from big data.	Machine learning is trained on a training dataset that teaches the computer how to make sense of the data and then make predictions about new datasets.
Data mining extracts rules from existing data.	Machine learning teaches the computer how to learn and understand these rules.
Although the accuracy rate in data mining is not high, information can be discovered from small data as well as big data.	A machine learning algorithm needs a standardized data stream and requires big data for accurate results.

3. TRAFFIC ACCIDENTS AND THE AFFECTING FACTORS

Traffic accidents occur every day and every time in our world, and according to statistics, deaths and serious injuries constitute an important part of it. It is often difficult to determine how accidents have occurred and what specific circumstances have led to such incidents. This makes it difficult for local law enforcement officers to explain the accident situation.

The number of land vehicles according to the 2021 statistics of the Ministry of Transport and Infrastructure of Turkey; from 2003 to the end of September 2021, it has increased 181% in the last 29 years and the number of vehicles has reached 25 022 960 [8]. In parallel with the increasing number of vehicles, traffic accidents are also increasing.

Human error, vehicle, and environmental conditions play an important role in the occurrence of traffic accidents. It is possible to identify these important factors from the data obtained from the accidents. Measures should be developed to prevent the accident and/or the behavior leading it in identifying the factors that cause traffic accidents and detecting patterns. It is claimed that traffic accidents occur randomly [9-12]. It is very important to identify the potential risk factors that contribute to traffic accidents and their severity.

3.1. Literature Review

In this section, academic studies on traffic accidents using machine learning and data mining processes are examined. Reference studies examining the effects of data mining and machine learning on traffic accidents and social science were reviewed, and their contribution to the literature was sought for.

Shakil et al. [13], in their study, estimated the severity and affecting qualities of those who had traffic accidents. Using the traffic accident data set recorded from 1979 to 2015 in England, the authors made use of and applied different machine learning algorithms such as decision tree (J48), OneR, ZeroR, and Naive Bayes. The research ended with model performance comparisons and inferences on the model, which predicted the severity of the casualty with an accuracy of over 96%.

In the study of Dağlı et al. [14], information about the traffic accident situation was estimated by using the traffic data center, published as open access by the Istanbul Metropolitan Municipality of Turkey. Decision tree methods, artificial neural network and k-

nearest neighbor algorithm were used in the research over the 136.964 data and features related to 12 traffic accidents. Model achievement in the classifier analysis was as follows: The success of the ANN model was 92.1%, the kNN model was 91% and the DT was 89.8%. In addition, the estimation capabilities of the models were determined with the ROC curves and AUC values.

In the study conducted by Çelik and Sevli [15], traffic accidents that occurred in Texas (Austin, Dallas and San Antonio) between 2011-2021 were examined. Factors affecting traffic accidents were determined from the traffic accident data. In the application study, XGBoost, Logistic Regression, SVM, Random Forest and KNN techniques were applied. It was determined that the technique with the highest machine learning model success was logistic regression with 89% model success. As a result of the research, it has been determined that the most important variables affecting traffic accidents are speeding, careless driving and inability to follow the road.

In the Zhu study [16], traffic accidents of vehicles and pedestrians and the factors causing accidents were determined by using three-year traffic accident data from Hong Kong. CRT, Rondom forest, ANN, SVM and Gradient Boosting algorithms were used to reveal the predictive model. As a result of the research, it was determined that the model with the best performance was ANN. In addition, it was concluded that the probability of fatal and severe vehicle-pedestrian collisions increases in light rain and uncontrolled intersections.

Sharef et al. investigated [17] the accident injury levels of the city of Abu Dhabi between the years 2008-2013 by data mining. In the study, among the data mining techniques, Lineer Support Vector Machine, Decision Tree and Bayesian Network were used. The results of the research revealed that elderly and male drivers and front seat occupants suffered serious or fatal injuries.

The study by Acı and Özden [18] examined the factors affecting the traffic accident outcomes in Adana, Turkey and classified the same using machine learning techniques. The research data was obtained from the weather conditions between the years 2005-2015 provided by the Regional Directorate of Meteorology, along with the 25015 data obtained from the reports kept by the Adana Regional Traffic Branch. Matlab and IBM SPSS Statistic were used as machine learning. In the study, Support Vector Machine, k-NN, Logistic Regression, Naive Bayes, Multilayer Perceptron and Decision Tree techniques were applied. As to the techniques with the highest model success; it has been determined that they are k-NN, Multilayer Perceptron and Decision Tree techniques. As a result of the research, it has been determined that the weather conditions are cloudy and the ground temperature is important in traffic accidents.

4. EXPERIMENTAL FINDINGS

Knime (Konstanz Information Miner) [19] can be defined as open-source enterprise software for data mining processes and machine learning. The Knime analytics tool is a low-code tool used in data science that solves business problems without the need for coding. With the help of every node in the Knime workflow; accessing data from different databases (.xls, .arff, .csv etc.), merging and collecting data from different databases, performing basic statistical analysis by reading the data, preparing the data for the model from the basic stages of data mining (Cleaning, normalization, etc.), and automatic training of the data are possible. It provides optimization process by applying data mining and machine learning algorithms. It exposes process parameters via GUI dialog on each node in the Knime workflow. Also, the Knime analytics tool is a good visualization tool and can provide data science solutions with workflow. Knime is a machine learning software for automation, management and deployment that provides workflows for data science and data mining solutions. In fact, it includes components of data mining process and automating machine learning, and we can collect it under the heading of "*Data Science*".

CRISP-DM data mining methodology was used in the research. CRISP-DM is to be used judiciously in solving the research problem.

4.1. Data Mining Process

Data mining and machine learning application process stages and related concepts are analyzed and explained by considering the research data.

4.1.1. Business understanding

The first step in this study, in which machine learning and data mining are evaluated together, is to understand the job and its requirements. The business understanding in the research is to determine the factors that cause traffic accidents and to explain which is the best among the classification techniques used in the research.

4.1.2. Understanding traffic accident data

At this stage, data on traffic accidents were collected. A questionnaire form was used to collect information in the research. In the study carried out countrywide in Turkey, it was responded by 528 individuals residing in Turkey who hold a driving license and a traffic insurance policy in April and May 2021. The questionnaire was given to individuals obtained by random sampling method on the internet (Digital) environment due to the Covid-19 pandemic. Google Forms was used to create the online survey form. The effect of the pandemic has prevented the increase in the amount of data.

4.1.3. Data preparation

This is the stage of preparing the data for the model. Data collection, integration, clearing of outliers and extremes, normalization, transformation and feature reduction.

4.1.3.1. Data collection

Traffic data set was obtained by means of a questionnaire. There is no integration in the data. The data relate to traffic accidents. The variables used in the research were the traffic accident type, tendency to have traffic insurance, gender, education, age, whether the accident occurred within or out of the city, the month of the accident and the weather on the day of the accident.

4.1.3.2. Data cleaning and outliers

It covers the work done to improve data quality. It is the removal of outliers, outliers, and missing values in the data. In this study, traffic accident data were also cleaned. In addition, missing data were detected in Knime and were not included in the model.

4.1.3.3. Data transformation and reduction

Classification techniques were used in the study and some features were found to be insignificant in reaching the goal of the study. Features that were found to be insignificant for the research, such questions were removed from the data set. It is necessary to select the features from the research data in order to have high model accuracy and to work in the model.

4.1.4. Modeling

After the study data was prepared for the model, the features were selected and the modeling was started. The techniques to be used in the research were determined. In the training phase, a new model was created with the real value of the class variable. At this stage, the estimated and actual values for model performance success are compared and the results are interpreted. Also, there was no need to transform the data in the study. The type of traffic accident was chosen as the target variable in the study. In the research, the tendency to have traffic insurance, gender, education, age, whether the accident was in the city or outside the city, the month of the accident and the weather on the day of the accident were taken as independent variables.

4.1.4.1. Classification techniques

Among the classification techniques in the research, ***Random Forest, Naive Bayes, Gradient Boosted*** and ***Tree Ensemble Algorithms*** were used. The accuracy rates of the algorithms used in the research were compared. The algorithm that gives the highest model accuracy is the Random Forest algorithm. The factors affecting traffic accidents were determined and interpreted according to this algorithm.

Random Forest [20] algorithm is a supervised learning algorithm that creates multiple trees trained on a randomly selected subset of the training set. Trees in different subsets generalize their classification in complementary ways. Because the technique is a combination of learning patterns, it outperforms other techniques. In the Random Forest algorithm, n random records are taken from the data set that contains k records. It generates an output by creating individual decision trees for each random sample [21]. Generating a Random Forest algorithm is given in Figure 1.

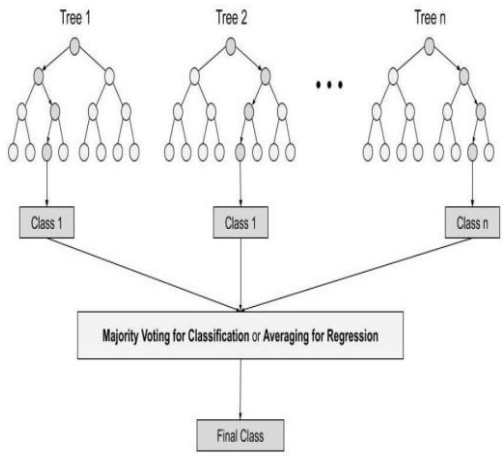

Figure 1. Random Forest algorithm [20]

Naive Bayes proposes a machine learning model based on probability principles, using Bayesian theory. Gradient Boosted Algorithm is an Ensemble Algorithm to make better predictions with few trees. The classification model is formed by correcting the errors from the previous tree. Tree Ensemble algorithm uses the regression method to learn from the dataset. Unlike the decision tree learner, it takes an approach. It uses the regression method to learn from the dataset. It has multiple regression trees for forecasting [22].

The aim of this study is to determine which classification techniques are important for analyzing traffic accidents and to determine which factor affects traffic accidents among the variables used in the research. The screenshot below in Figure 2 shows the Knime workflow for building and running data mining and machine learning algorithms.

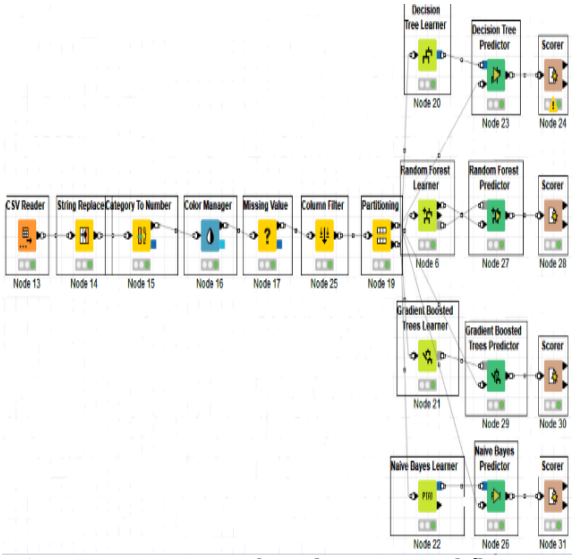

Figure 2. Knime machine learning workflow using classification techniques

4.1.4.2. Training the model

Training the model is the main difference between machine learning and data mining processes. By training the algorithm, the data is fed and the learning process is completed. In the estimation of the classifier algorithms, 70% of the data was used as the training data set and the remaining 30% as the test data.

"Gini" and "Knowledge Gain" as a quality measure in the model were compared in terms of performance. As a result of the research, the model accuracy rate of the "Knowledge Gain" quality measurement was higher.

4.1.4.3. Parameter setting

Model performance can be increased with the help of advanced parameter performance settings [23]. The hyperparameter optimization value is a parameter value and is used to control the process. A predefined loss over data function is minimized and a model that provides the optimal model hyperparameter value is called for. It is required to find out the most reliable model among the ones created, along with the one with the highest degree of accuracy.

One of the evaluation methods is k-fold cross validation. The K-fold cross validation method is a technique used to validate the performance of the model, it uses different parts of the data set as the validation set and evaluates the model [24].

4.1.4.4. Model prediction

This stage is the performance preparation stage of the model [25]. In this phase, the established model is now free from the human element, what means that the machine makes predictions on its own. It helps in determining how well the model that best represents the selected traffic accident data will work.

4.1.5. Model performance evaluation

In the success and high accuracy of the classifier model, the real and predicted classes of the observation are compared. As model performance measures, accuracy, precision, true negative, TP, recall, precision, FN, specificity, and F measures are estimated.

At this stage, it will work upon the classification algorithms that can be used to train the dataset from the Knime machine learning workflow. In measuring model performances, various evaluation criteria were used among the classifiers.

The following Table 3 and measurements were used in the model performance evaluation.

Confusion Matrix: It is known as a good performance measure in classification in machine learning. This matrix describes the actual and estimated values of the data in four different combinations. Evaluation of machine learning with confusion matrix is given in Table 2.

Table 2. Evaluation of machine learning by confusion matrix

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative(0)	FN	TN

TP (True Positive): The positive true value and the predicted value are the same.

FN (False Negative): It is the wrong prediction of the obtained model. Indicates misclassification.

TN (True Negative): Since the predicted and actual values are the same, it has made a correct classification.

FP (False Positive): It has made an incorrect classification. Model predicted it positively while the true value was negative.

Based on the definitions given, the formulas for the model performance criteria are given below.

Precision: TP/ (TP + FP)
Accuracy: (TP + TN) / Total (TP + TN + FP + TN)
F-Measure: 2TP/ (2TP + FP + FN)
Recall: TP/ (TP + FN)

Classification algorithms were used in the study and the results are given in Table 3.

Table 3. Prediction results of algorithms

Algorithms	Recall (%)	Precision (%)	Accuracy (%)	F measure (%)
Random Forest	96.3	87.3	85.1	91.6
Gradient Boosted	86.4	87.0	75.2	86.7
Naïve Bayes	86.1	87.8	75.3	86.9
Tree Ensemble	95.6	85.1	81.9	90.0

In terms of overall performance, the "Random Forest" algorithm is the best due to its iterative classification. It is observed that the confusion matrix of the "Random Forest" algorithm has a high accuracy level of 85.1% and other metrics.

When the data mining process is modeled, it can be defined as the stage of distributing valuable and meaningful information obtained based on the model performance results.

4.1.6. Deployment

This is the stage where the data mining process is modeled and presented as a report. The model result of the research obtained by data mining is given. This was followed by "Month in which the accident occurred", "Age", "Weather conditions, "City or not" and "Education of the driver".

Some of the rules found as a result of the predictive model are given below.

- The estimated reason for the driver, who is between the ages of 18-24 and who drives outside the city in March and April and has a traffic accident in snow and sleet weather, is "Disobeying traffic rules".
- The estimated cause of the traffic accident by the driver driving in the same conditions and in the city is "Unconscious use".
- In the same weather conditions and if the age of the driver with a graduate education is between 18-34 years old, the estimated reason for the traffic accident is "Unconscious use".
- In the same weather conditions and if the age of the driver, whose education is at university or below, is more than 34, the estimated cause of the traffic accident is "careless driving".
- In February, April and June, the estimated cause of traffic accidents for drivers over 24 years of age with a university degree or higher is "Overspeed".

The results of the Random Forest algorithm screenshot are given in Figure 3.

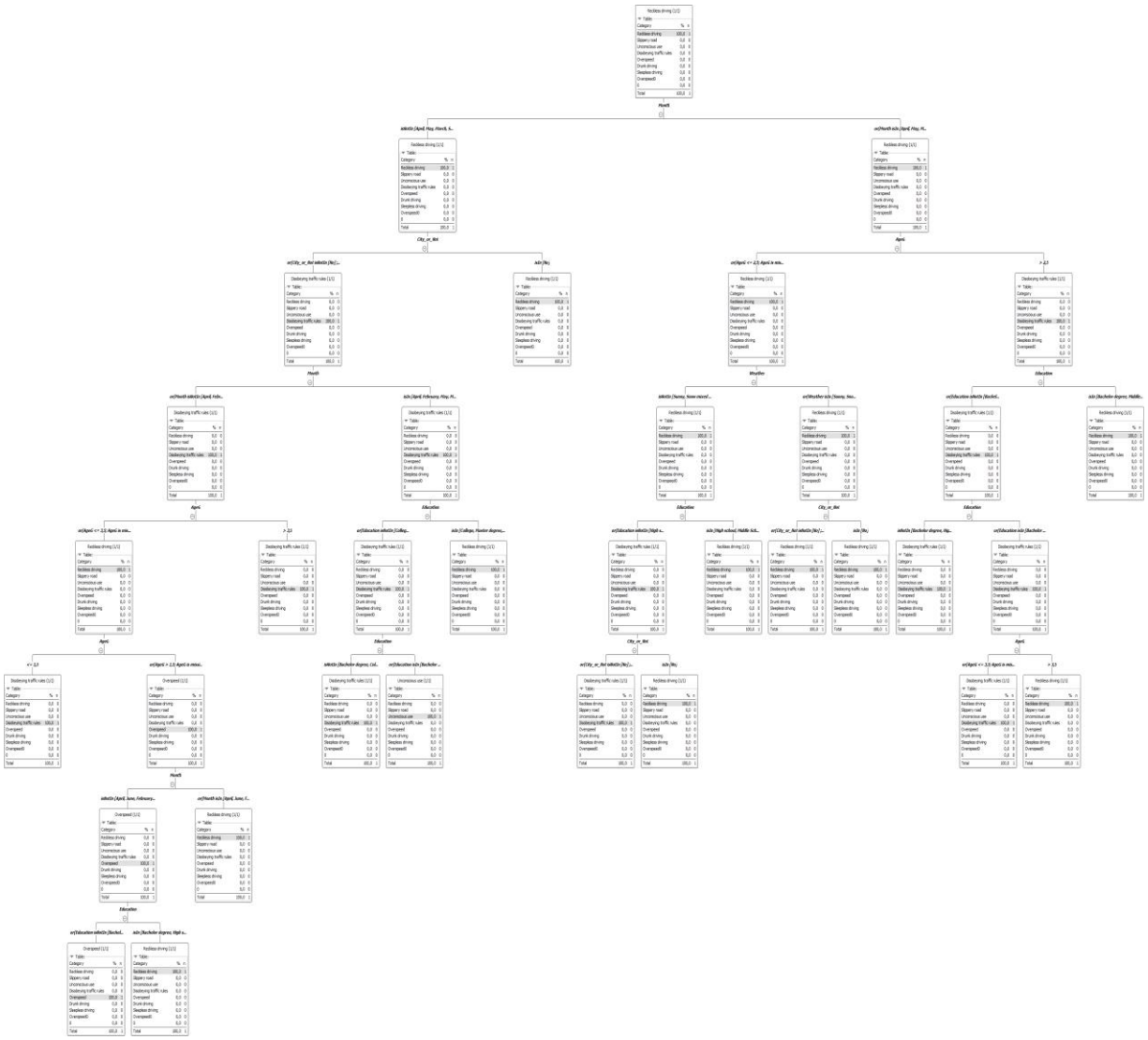

Figure 3. Random Forest algorithm screenshot

5. CONCLUSION

Deaths and injuries occur worldwide due to traffic accidents, and measures should be taken to reduce the effects of accidents. To identify the most effective measures needed to be taken, effective analyzes are required to find out and classify the causes leading to the occurrence of accidents. The purpose of this article is to determine the factors that affect occurrence of the traffic accidents through machine learning and data mining approaches. Data on traffic accidents were obtained from online surveys conducted throughout Turkey. The small sample size in the data set obtained during the Covid-19 pandemic, the lack of detailed information about the characteristics of the vehicle and the driver are the limitations of the research. In the future, more detailed analyzes can be made by meeting these criteria.

Random forest, Gradient Boosted, Naïve Bayes and Tree Ensemble algorithms were used in the study. However, according to the recall, precision, accuracy and F measurement model performance criteria, the Random Forest Algorithm created the best model.

The model result of the research obtained by data mining is given. This was followed by "Month in which the accident occurred", "Age", “Weather conditions”, "City or not” and "Education of the driver". According to the results obtained from the classification models used in the research, it is evaluated that the

proposed models can be used for the prediction of traffic accidents. It is suggested that there are important approaches for data mining to test for its current situation using machine learning, to make foreseeable predictions and to make a correct decision, and it should be used extensively in this area. Analysis findings can help us to take measures to prevent accidents, reduce the number of traffic authorities, and take necessary preventive and corrective actions for a minimum time and cost. The models obtained in this research and also from similar studies in the literature can be used in the creation of a decision support system.

By examining the data from various traffic accidents and developing a model, the factors and hazards that affect traffic accidents can be determined by data mining and machine learning approaches. The resulting model/s and the accident scene and behavioral patterns will be useful in the development of a traffic safety policy and in the strategic roadmap of law enforcement officers. Particularly due to the limited budgets, scientific research has a very high contribution to the determination of the accident size and affecting factors, and to the prevention of death and serious injuries.

REFERENCES

1. IDC&Statista, “Data Created Worldwide 2010-2024”, https://financesonline.com/how-much-data-is-created-every-day/, May 8, 2022.

2. KDnuggets, “Machine Learning Algorithms”, https://www.kdnuggets.com/ 2021/ 01/machine-learning-algorithms-2021. html, June 3, 2022.

3. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P., “From data mining to knowledge discovery in databases”, AI Magazine, Vol 17, Issue 3, Pages 37, 1996.

4. Ersöz, F., “Veri Madenciliği Teknikleri ve Uygulamaları”, Seçkin Yayınevi, Ankara, 2019.

5. Patel, K., Fogarty, J., Landay, J., and Harrison, B., “Investigating statistical machine learning as a tool for software development”, In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08), Association for Computing Machinery, New York, NY, USA, Pages 667–676, 2008.

6. Sotirios P. C., Vassilis S, Petropoulos A., Stavroulakis, E., Vlachogiannakis, N., “Forecasting stock market crisis events using deep and statistical machine learning techniques”, Expert Systems with Applications, Vol. 112, Pages 353-371, 2018.

7. Ersöz F.&Çınar, Y., "Veri madenciliği ve makine öğrenimi yaklaşımlarının karşılaştırılması: Tekstil sektöründe bir uygulama", Avrupa Bilim ve Teknoloji Dergisi, Vol. 29, Pages 397-414, 2021.

8. Ulaştırma ve Altyapı Bakanlığı, “Ulaşan ve Erişen Türkiye Raporu”. https://www. uab.gov.tr/uploads/pages/bakanlik-yayinlari/ ulasan-ve-erisen-turkiye-2021.pdf., June 11, 2022.

9. Prato, C., Bekhor, S., Gal-Tzur, A., Mahalel, D., Prashker, J., “Exploring the potential of data mining techniques for the analysis of accident patterns, 12th World Conference on Transportation Research. Lisbon, Portugal, 2010.

10. Elvik, R., “A re-parameterisation of the power model of the relationship between the speed of traffic and the number of accidents and accident victims”, Accident Analysis & Prevention, Vol 50, Pages 854–860, 2013.

11. Martín, L., Baena, L., Garach, L., López, G. and de Oña, J., “Using data mining techniques to road safety improvement in Spanish roads”, Procedia - Social and Behavioral Sciences, Vol. 160, Pages 607 – 614, 2014.

12. Gupta, M., Solanki, V. K. and Smith, V. K., “Analysis of data mining technique for traffic accident severity problem: a review”, Second International Conference on Research in Intelligent and Computing in Engineering, ACSIS, Vol. 10, Pages 197–199, 2017.

13. Shakil, F.A., Hossain, S.M., Hossain, R.&Momen, S., “Prediction of road accidents using data mining techniques” In Proceedings of International Conference on Computational Intelligence and Emerging Power System, Pages 25-35. Springer, Singapore, 2022.

14. Dağlı E., Büber M., Taspınar Y.S., “Detection of accident situation by machine learning methods using traffic announcements: the case of metropol Istanbul”, International Journal of Applied Mathematics Electronics and Computers, Vol. 10, Issue 3, Pages 61-67, 2022.

15. Çelik A.&Sevli O., “Predicting traffic accident severity using machine learning techniques, TJNS, Vol. 11, Issue 3, Pages 79-83, 2022.

16. Zhu, S., “Analyse vehicle–pedestrian crash severity at intersection with data mining techniques”, International Journal of Crashworthiness, Vol. 27, Issue 5, Pages 382, 1374, 2021.

17. Sharaf A., Fahad A.& Ahmad A., “Risk analysis of traffic accidents’ severities: An application of three data mining models”, ISA Transactions, Vol. 106, Issue 2, Pages 213-220, 2020.

18. Özden, C. Acı, Ç., Analysis of injury traffic accidents with machine learning methods: Adana case” Pamukkale Univ. Müh. Bilim Derg., Vol 24, Issue 2, Pages 266-275, 2018.

19. Berthold, M., Cebron, N., Dill, F., Gabriel, T., Kötter, T., Meinl, T., Ohl, P., Thiel, K., Wiswedel, B., "KNIME-the Konstanz information miner", ACM SIGKDD Explorations Newsletter, Vol 11, Issue 1, Page 26, 2009.

20. Ho T.K., “Random decision forests”, In: Proceedings of 3rd International Conference on Document Analysis and Recognition. IEEE, Pages 278–282, 1995.

21. Analytics Vidhya, “Understanding Random Forest”,https://www.analyticsvidhya.com/ blog 2021/06/understanding-random-forest/, June 5, 2022.

22. Knime Developers, https://www.knime. com/developers, September 5, 2022.

23. Martinez, J.C., “The 7 Steps of Machine Learning”,https://livecodestream. dev/post /7-steps-of-machine-learning/, June 2, 2017.

24. Bergstra, J., Ca, J. B., & Ca, Y. B., “Random search for hyper-parameter optimization Yoshua Bengio”, Journal of Machine Learning Research, Vol. 13, Pages 281–305, 2012.

25. Mayo, M, “Frameworks for Approaching the Machine Learning Process”, https://www. kdnuggets.com/2018/05/general-approaches-machine-learning-process.html, October 19, 2022.