ARTIFICIAL INTELLIGENCE STUDIES
2022, 5(2):35-46
www.aistudies.org

Gazi AKADEMİK YAYINCILIK

Makine Öğrenmesi Teknikleri ve Python Araçları Kullanılarak Türkiye'deki Trafik Kazalarının Sayısının Tahmin Edilmesi

Mustafa AL-ASADI*,[a], Sakir TASDEMİR[a], Humar KAHRAMANLI ÖRNEK[a]

[a] *Selçuk Üniversitesi Bilgisayar Mühendisliği Bölümü, Konya, 42130, TÜRKİYE*

MAKALE BİLGİSİ

Alınma: 20.09.2022
Kabul: 18.11.2022

Anahtar Kelimeler
Makine öğrenmesi, Trafik Kazaları Tahmini, Ortalama Mutlak Hata, Python.

***Sorumlu Yazar:**
e-posta: masadi@lisansustu.selcuk.edu.tr

ÖZET

Trafik kazaları son yıllarda dünya genelinde hızlı bir büyüme göstererek büyük can ve mal kayıplarına neden olmaktadır. Bu nedenle trafik kazalarının önceden tahmin edilmesi, ulaşımın ve kamu güvenliğinin iyileştirilmesi için çok önemlidir. Makine öğrenmesi, verilerden bilgi çıkarabilen ve değerleri tahmin etmek için istatistiksel yöntemler kullanabilen yapay zekanın bir dalıdır. Bu çalışmada, Türkiye'de 2029 yılına kadar trafik kazalarındaki ölü veya yaralı sayısını tahmin etmek için üç makine öğrenme tekniği uygulanmıştır:lineer regresyon (LR), karar ağaçları (DT) ve rastgele orman (RF). Bu teknikler TÜİK web adresinden elde edilen gerçek bir veri seti kullanılarak test edildi. Sonuçlarda lineer regresyon'un (LR) en iyi performansa sahip olduğu görülmüştür. Bu sonuç, yaklaşımın yol kazalarını tahmin etmedeki üstünlüğünü göstermektedir. Sonuç olarak bu çalışma, karayolu taşımacılığı ve sigorta acentelerinin karayolu güvenliği stratejileri geliştirmelerine yardımcı olacaktır.

DOI: 10.30855/AIS.2022.05.02.01

Predict the Number of Traffic Accidents in Turkey by Using Machine Learning Techniques and Python Tools

ARTICLE INFO

Received: 20.09.2022
Accepted: 18.11.2022

Keywords
Machine learning, Traffic Accidents Prediction, Mean Absolute Error, Python.

***Corresponding Author:**
e-mail: masadi@lisansustu.selcuk.edu.tr

ABSTRACT

Traffic accidents have grown rapidly throughout the world in recent years, causing great loss of life and property. Therefore, predicting traffic accidents is very important for improving transportation and public safety. Machine learning (ML) is a subfield of artificial intelligence that can extract information from dataset and use statistical approaches to predict values. In this study, three ML techniques were applied to predict the number of dead or injured in traffic accidents in Turkey until 2029: linear regression, decision trees, and random forest. These techniques were tested using a real dataset obtained from the TUIK website. In the results, it was seen that linear regression (LR) had the best performance. This result shows the superiority of the approach in predicting road accidents. Ultimately, this study will help road transport and insurance agencies develop road safety strategies.

DOI: 10.30855/AIS.2022.05.02.01

1. INTRODUCTION (*GİRİŞ*)

In general, traffic accidents have not one reason, but they occur due to the interaction of many factors like driver, environment, car, etc. They modelled a simple neural network with electrical circuits [1]. Traffic accidents are a major issue in our societies all over the world. In 2010, the World Health Organization (WHO) estimated that road traffic injuries killed 1.25 million people [2]. Moreover, statistics show that between 2020 and 2022, 50 road deaths were recorded in Europe every minute. Therefore, Traffic accidents are considered a significant issue to public safety. In Turkey, as of the end of June 2021, the total number of registered vehicles was 24.7 million [3] on average, more than one million cars are added to traffic each year [4]. Developed countries have created a traffic accident information database to analyze and estimate traffic accidents and their consequences. The Traffic Insurance Information Center (TRAMER) was established in Turkey in 2003. TRAMER conducts research on data storage and the compilation of statistical results for traffic accidents. In one of their studies, TRAMER discovered that 983.808 road accidents occurred in the country overall in 2020 [5, 6].

In the last two decades, one of the research areas in road safety has been the severity of RTA. On the road accident severity classification-based models, researchers used novel methodologies. The authors looked at how to develop models using a typical statistical technique. These methods aid in gaining insight into and identifying the underlying causes of vehicle accidents and other issues that affect road safety. Machine learning now outperforms traditional statistical-based models in forecasting the model due to the large volume of data available [7].

Because of multiple contributing factors (such as road engineering, human behaviour, environment, etc.), accident prediction classifier or (AMP) development one of the furthermost important and complex factors of road safety design. A successful application of APM can foretell an accident in a specific area and communicate this information to the neighboring road users so they can take preventative action. The majority of previous studies dealt with accident prediction as a classification or regression issue. Moreover, Poisson regression is frequently used to model the riskiness of traffic accidents. The output of a Poisson regression, which uses the Poisson distribution for estimating error, is the response variable's natural logarithm. In [8-10], a Poisson regression variant is applied by relaxing the case of mean equals to variance.

Due to the sparse nature of road accident occurrences and the large number of zeros observed, the classic Poisson regression model suffers significantly during the modelling process. The zero-inflated Poisson regression and zero-inflated negative binomial are suggested in [10], [11], and [12], respectively, to reduce this spurious over-dispersion. The small sample size bias and low sample mean, i.e., the long-term mean of zero, have a significant impact on the suggested model's [11] accuracy. Conway-Maxwell Poisson regression model outcomes were also observed to be similar [8]. Furthermore, although data on traffic accidents is multivariate by nature, it is frequently modeled as univariate data. A novel multivariate Poisson-Lognormal (MPL) model that can handle both over-dispersion and general correlation construction was put forth by Park and Lord in 2007 [13]. The aforementioned regression techniques rely on the assumption that predictors and target variables have a pre-defined underlying relationship. As a result, when these relationships are violated, the prediction accuracy is quite erratic. Chang and Chen in [12] construct a CART model (classification and regression tree) to prevent such performance concerns and demonstrate the actual relationships between predictors (such as road design, environmental factors, and traffic characteristics) and target variables (e.g. road accident). A variety of ML techniques have been investigated in [14-16] in light of the accessibility of vast volumes of data on traffic accidents. In order to predict highway traffic accidents, Lv et al. [13] used the K-NN approach on real-time traffic flow. SVM-based models for multidimensional road accident data were created by the authors in [15], and correlation-based feature selection (FS) is utilized to assess the impact of accident precourses. Another method used in [16] to model large accident data and apply LR and k-means to predict an accident is MapReduce programming. However, as the number of features increases, the prediction's false-positive rate rises dramatically. Decision rules-based DT can be used to identify the underlying behavior and relationships between traits or factors that contribute to auto accidents [12], [14]. As a result of a tree structure constraining feature extraction and feature correlation, the prediction accuracy is considerably low. Applying ensemble machine learning techniques can reduce this issue. One of the ensemble modes, random forest integrates various base models (such as decision trees) to enhance the prediction of traffic accidents [17]. Further advancement can be made by investigating the connection between real-time traffic characteristics and traffic accidents using boosting ensemble models as Adaboost [18] and XGboost [19].

Researchers in the field have paid close attention to predicting the number of accidents. Thus, most countries are keenly interested in the annual number of people killed and injured in traffic accidents [20]. Brüde [21, 22] addressed the problem of predicting the number of road traffic deaths in Sweden using data from 1977 to 1991. Dadashova et al. [22] examined monthly data on the number of fatal accidents in Spain from 2000 to 2011. However, there are few models in Turkey that can predict the number of traffic accidents. In this topic, "Doan and ANgüngör" predicted the number of accidents, deaths, and injuries in Turkey using neural network and nonlinear regression methods [23]. Furthermore, "Doan and ANgüngör" used neural networks to develop models predicting the number of accidents deaths and injuries in Krkkale city [24]. "Doan" compared machine learning techniques such as neural networks and genetic algorithms for Turkey and selected metropolises and concluded that machine learning models performed better [25]. Using neural networks, "Hüseyin Ceylan" developed models for predicting the number of traffic accidents deaths and injuries in Turkey [26].

Previous literature has indicated that there are many methods, ranging from statistical methods to machine learning, that have been used to analyze and predict road accidents. However, road accident prediction research is still under development because of multiple contributing factors such as road engineering, the environment, human behaviour, vehicles and others. Also, we observed the superiority of machine learning algorithms over statistical methods. In this study, we seek to test a previously unused algorithms to predict the number of traffic accidents (such as RF and DT). Also, we are trying to uncover the real extent of Turkey's problems for the next ten years to draw a perspective for safety decision-makers.

The paper is organized as follows: The methods and machine learning techniques used in this study are presented in Sections 2 and 3. Evaluation metrics of each model mentioned above are done in Section 4. The results are presented in Section 5. Lastly, we give some conclusions in Section 6.

2. METHODS *(YÖNTEMLER)*

In this study, traffic accident estimation models were developed using three techniques (linear regression (LR), decision trees (DT), and random forest (RF)) to predict the number of traffic accidents (with death or injury) in Turkey until the year 2029. Our choice of these models is that the decision trees have not been used for this problem previously. Where previous studies have focused on the use of linear regression and neural networks for this purpose. The selected models were tested using real data are given in Table 1 that include numbers of accidents (with death or injury), materially damaged accidents, deaths and injuries resulting from traffic accidents covering the years 1998–2017 are obtained from the website: http://www.tuik.gov.tr, under the section of Ministry of Transport and Infrastructure. These statistics were prepared by the Turkish Statistical Institute (TurkStat) in cooperation with the Gendarmerie General Command and General Directorate of Security in Turkey.

Statistics on traffic accidents involving killed or injured people and material loss that occur on Turkey's road network are produced in order to direct traffic actions, provide a safer traffic environment, and identify deficiencies in traffic rules. All accidents that occurred on Turkey's road network in a police or gendarmerie zone and were reported to TRAMER (Motor TPL Insurance Information Center) with official reports are included.

Until 2015, the number of people killed in road traffic accidents was based solely on deaths at the scene. As a result of the work completed by the coordination of the General Directorate of Public Security, General Command of Gendarmerie, Ministry of Health, and Turkish Statistical Institute to harmonise the statistics of persons killed in road traffic accidents with international definitions, statistics on people injured in road traffic accidents and sent to health facilities who died within 30 days of the accident due to related accident and its consequences are now available. The death figures published as the number of people killed in traffic accidents differ from the Causes of Death Statistics published in terms of coverage and method.

In the models' development, the number of traffic accidents (with death or injury) were used as a model parameter with data between 1998 and 2017. The model's output was compared to actual values, and it was discovered that it is appropriate for this purpose.

This study's methodology is based on supervised ML. To put it another way, the algorithm "learns" by using data samples to infer a model, which is then tested with samples that were not used to develop the model. These test samples enable us to make comparisons the predicted values of the model to the actual values, allowing us to assess the model's accuracy in predicting real-world samples. The predicted values in this experiment are the number of traffic accidents. The steps in the used method are as follows:

- **Step 1: General investigation:** We reviewed all studies associated with the analysis and prediction of traffic accidents and the algorithms that were used.
- **Step 2: Data splitting:** Following the preparation of the dataset, 80 percent of the data was randomly selected to train the model, with the remaining 20 percent used for testing.

- **Step 3: Building models:** To predict traffic accident numbers, three different ML models are built. These models are trained on data between 1998 and 2017, and then forecasts are made for traffic accident numbers from 2018 to 2029.
- **Step 4: Evaluation models:** In addition to the Train and Test Split (Step 2), in which the performance of the classifier was evaluated using the training set, several metrics were calculated to evaluate the models' performance using the testing data, including mean absolute errors, root mean square errors, and coefficient of determination. Machine learning models are built using the scikit-learn Python module.

3. MACHINE LEARNING TECHNIQUES *(MAKINE ÖĞRENMESI TEKNIKLERI)*

3.1 Linear Regression

The linear regression equation represents the relation between the dependent variable and one or more independent variables. Simple linear regression is used when there is only one independent variable (SLR). MLR is the model's name (multiple linear Regression) [27].

Table 1. The annual distribution of traffic accident data in Turkey from 1998 to 2017.

n	Year	An accident with Death or Injury	Material Damaged Accident	Dead	injured
0	1998	5960	12552	1148	11241
1	1999	8754	18823	1534	15687
2	2000	10702	23577	1625	20529
3	2001	11318	22235	1432	21705
4	2002	11104	21751	1269	21820
5	2003	11002	22363	1148	21944
6	2004	13415	29118	1346	26548
7	2005	15079	35685	1310	30109
8	2006	16951	47265	1268	33326
9	2007	20047	56080	1545	39243
10	2008	19781	31888	1288	39305
11	2009	19392	16014	1331	39661
12	2010	19391	12932	1307	40021
13	2011	21042	12714	1253	43925
14	2012	23195	13567	1195	46994
15	2013	25273	12704	1292	50542
16	2014	26140	12119	1228	51723

17	2015	27810	13797	1276	54618
18	2016	28440	13731	1227	54762
19	2017	28559	13826	1235	54830

A simple linear regression (as shown in Figure 1) is a straight line that passes through a set of points with the goal of keeping the sum of the remaining squares of the model as low as possible. This indicates that regression is one of the most basic statistical methods, with the slope of the line representing the relationship between y and x corrected by the standard deviations of these variables. Ordinary least squares (OLS) is commonly assumed to minimize residuals (vertical distances among fitted lines & the data points). The sum of squared residuals (SSL) is used to calculate line accuracy through the sample points, and the goal is to keep this sum as small as possible.

3.2 Decision Trees

Decision trees are a type of prediction classifier that is used in a variety of fields, including artificial intelligence and economics. Diagrams of logical constructions are generated given a set of data, which are comparable to rule-based prediction systems in that they represent and categorize a series of conditions that occur sequentially in order to solve a problem [28, 29]. Decision trees (DT) are widely used in operations research, particularly in decision analysis, for help determine the most likely strategy for achieving a goal. Nonetheless, it is a standard tool in automated learning. A DT is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the test's outcome, and each leaf node represents a class label. Classification rules are represented by the paths from the Root to the Leaf. Simple decision trees (shown in Figure 2) are made up of nodes, number vectors, arrows, and labels [30].

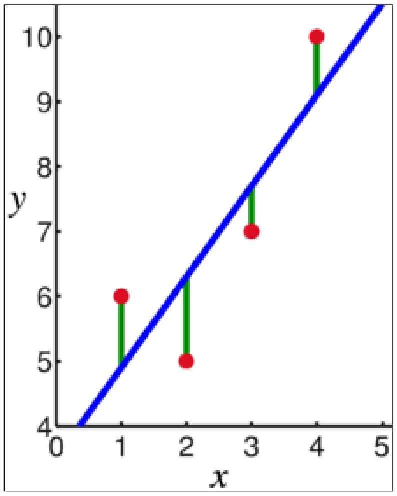

Figure 1. Depicts the link between the dependent and independent variables (x) (y)

1) Each node can be defined as when a decision has to be made among several possible ones, that is means that as the number of nodes increases, the number of possible endings that the individual can reach increases. This creates a tree with many nodes complicated to draw by hand and analyse due to numerous paths that can be followed.

2) The vectors of numbers would be the final solution that is reached depending on the various available possibilities, giving the profits in that solution.

3) The arrows are the connections between one node and another and represent each different action.

4) The labels are found on each node and each arrow and name each action

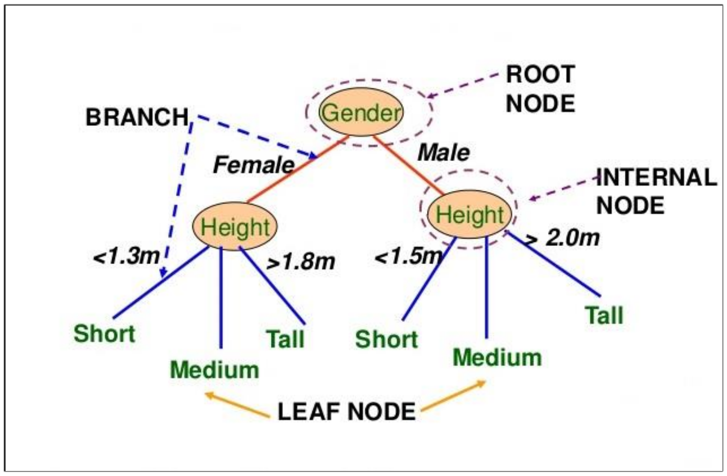

Figure 2. Traditional decision trees

3.3 Random Forest

Random Forest (RF) creates the Classification and Regression Tree (CART) method by combining bootstrapping and random feature selection. RF is made up of a number of unrelated decision trees. An RF generates a set of DT from a random selected subset of training data for a classification operation. It then collects votes from various DTs to determine the final class of the target. Figure 3 depicts the overall architecture of the RF. Random Forest was coined by Leo Breiman in 1999 [31]. He investigated various methods for randomizing decision trees, such as bagging and boosting. Tin Kam Ho's 1995 research foreshadowed his work [32]. It is beneficial to learn more about how to calculate different options in order to understand and use them. The majority of options are based on two data objects generated by random forests. The following are the main characteristics of RF:

1) bootstrap sampling (bagging) - a random sample size with replacement.
2) Random FS- selecting only a small number of m examples in each node's split at random.
3) full-depth tree development
4) Out-of-bag error - calculating error on samples not chosen during bootstrap sampling [33].

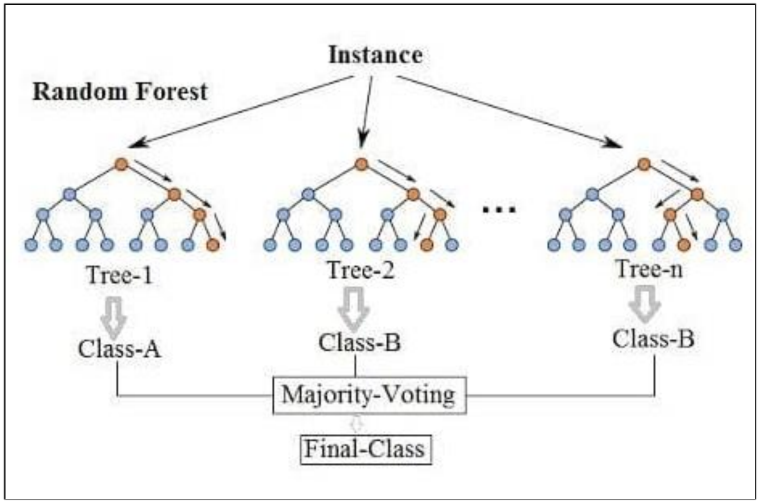

Figure 3. Implementation and evaluation of results

4. EVALUATION METRICS

Several metrics were calculated to assess the performance of the classifiers. Regression methods for example Train and Test Split (holdout), cross-validation (CV), and the bootstrap, in general, can be used with predictive classifiers to estimate classifier performance using the training set [34]. The most important metrics for evaluating the regression model using testing data are mean absolute errors (MAE), mean square errors (MSE), besides, root mean square errors (RMSE). Traffic accidents can be handled in a variety of ways using machine learning. We can treat it as a regression problem and predict the number of accidents based on the dataset's attributes (for example, date, road engineering, the environment, human behavior, vehicles, and so on). We established three regression models in this study. In developing models, the number of traffic accidents (with death or injury) was used as a parameter (sufficient to build the baseline then compare results).

4.1 Train and Test split

Using different sets of training and testing to evaluate the model's performance is the simplest way to do so. The data is divided into two parts using this technique. The first section trains the model and predictions makes on the second section, which then compares predictions to the expected outcomes. The size of the split data is generally determined by the size of the dataset. It is common practice to use 80% of the training data and 20% of the testing data [35]. In our study, we split data using train and test split. Where the original samples are divided into 80% for training and 20% for testing at random.

4.2 Error Measurements

Each ML model attempts to solve a problem with various objects and data. Mean absolute errors (MAE), Mean square errors (MSE), Root mean square errors (RMSE), and the coefficient of determination (R2) are commonly used in regression problems to evaluate the model, as formalized in Equations (1) to (3).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\breve{y\imath} - yi| \quad (1)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}* (yi - \breve{y\imath}\ )2 \quad (2)$$

$$\mathrm{R}^2 = \frac{1-MSE\ (Model)}{MSE\ (Baseline)} \quad (3)$$

Where y_i is the actual expected result and $\hat{y}_i$ is the model's prediction.

The average error size in a set of predictions, regardless of their direction, is measured by MAE. The average of all individual differences on the test sample for absolute variances between prediction and actual observation. MSE is an abbreviation for the average squared error of predictions. It computes the square variance between the predictions and the target, then averages those values for each point. The greater this value, the poorer the model. It is never negative, but for a ideal model, it would be zero. The square root of MSE is RMSE. The Square Root is used to scale the errors to match the scale of the targets [36]. R^2, which is closely related to the Mean square errors of the model and baseline, is additional metric we can use to evaluate a model. MSE baseline is the simplest model possible. Predicting the average of all samples is always the simplest viable model. A value near 1 indicates a model with close to zero error, while a value near zero indicates a classifier that is close to the baseline [28, 37].

5. RESULTS (BULGULAR)

A baseline is the foundation for a ML model's worst-case acceptable performance on a large dataset. In general, suppose a model performs worse than the baseline. It will be a failure in that case, and we should try a different model or admit that using ML techniques to improve the model is not appropriate for our problem. A baseline result is the simplest possible prediction. In some cases, such as regression, this may be the mean or median, while in others, such as classification problems, it may be the most common prediction [38]. The average of accidents (with death or injury) equal to 18167 accidents can be used as the baseline prediction in our case.

Figures 4, 5, and 6 depict the LR, DT, and RF algorithms used to predict traffic accidents in Turkey. Table 2 displays the MAE, MSE, RMSE, and R^2 for each model.

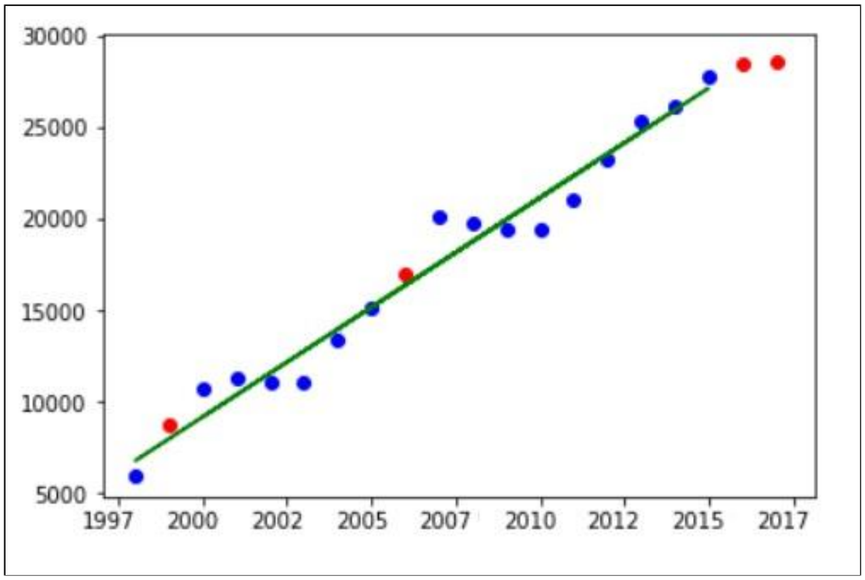

Figure 4. Simple Linear Regression and Scatter Chart (Traffic accidents vs. years).

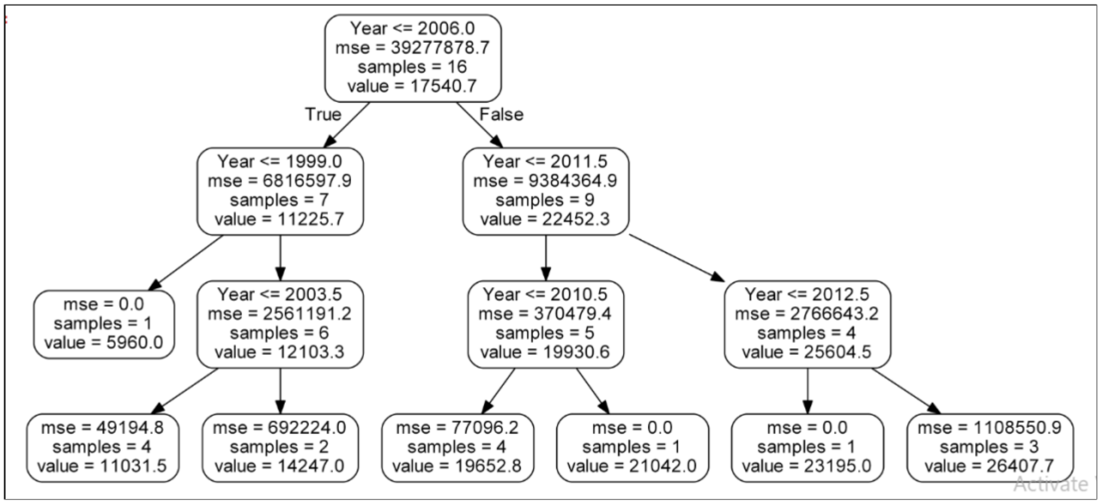

Figure 5. For traffic accident prediction, decision trees with min samples leaf = 1 and max depth = 3 are used.

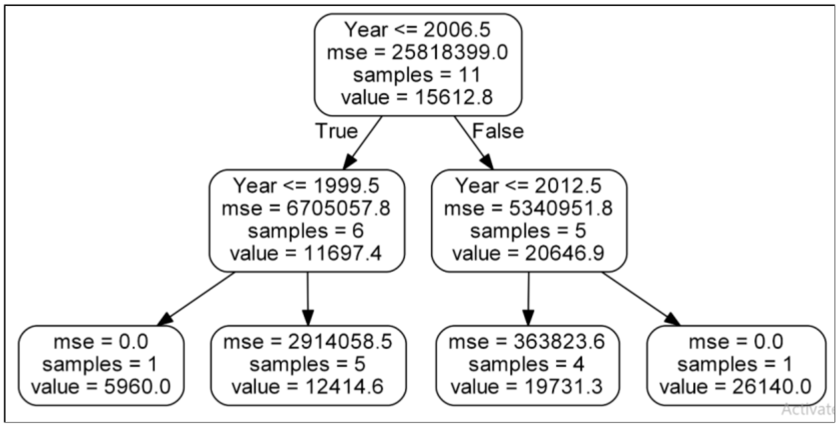

Figure 6. Random Forest was used to predict traffic accidents, with the number of estimators set to 10 and the maximum depth set to 3.

Table 2. MAE, MSE, RMSE and R² for models.

n	Classifier	MAE	MSE	RMSE	R²
1	Linear Regression	618.99	479424.40	692.40	0.993
2	Decision Trees	1511.25	3067180.25	1751.33	0.955
3	Random Forest	1201.89	1602268.33	1265.80	0.976

From the Table 2, Linear Regression had the best performance from other models, which have a minimum value for MAE equal to 618 accidents, which is more than a 29-degree improvement over the baseline (18167 accident). Moreover, it has the highest R2 value (0.993). Therefore, the number of casualties was predicted using the Linear Regression technique in Figure 7.

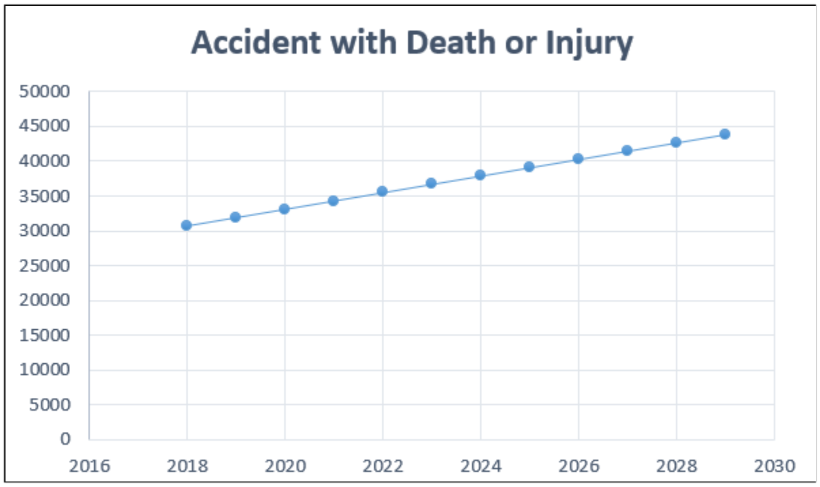

Figure 7. Predictions about the number of traffic accidents produced by the Linear Regression Model for Turkey.

As shown in Figure 7. When the data is examined, it is clear that the estimated traffic accidents are increasing. The number of traffic accidents is expected to reach 43877 in 2029, representing a 65% increase over 2017.

6. DISCUSSION AND CONCLUSIONS *(TARTIŞMA VE SONUÇLAR)*

Researchers have paid considerable attention to predicting the number of traffic accidents. However, there are limited models to predict the number of accidents in Turkey. In this study, three machine-learning models have been developed: LR, DT, and RF, to estimate the number of accidents (with death or injury) in Turkey until the year 2029. During the development of the models, the number of previous accidents (with death or injury) was used as an independent variable in training the models. To determine the model with the best performance, (MAE), (MSE), (RMSE), and the (R2) values are calculated. During comparing the results, linear regression yielded the best performance with the lowest RMSE, with a value of 692. As well, our outcomes show that using decision trees and random forests to estimate the number of traffic accidents is very promising compared to previous work (see Table 3). We achieve much better accuracy. In addition, our model requires fewer inputs (one input) than the [20, 23] models (55 inputs) and (18 inputs), respectively. As a result, it was inspiring that our modeling and analysis strategy performed significantly better.

Table 3. Evaluation of the Literature Studies

Study No.	Algorithm Used	RMSE	No. features
[23]	ANN	32107	5
	Smeed	124080	
	Andrea ssen	161595	
[26]	ANN	2873	5
This work	**Linear Regression**	**692**	1
	Decision Trees	1751	
	Random Forest	1265	

The results of our study indicate that the estimated number of accidents will increase in Turkey by 2029 and will be around 43,877, resulting in deaths and injuries. The rising number of traffic accidents and injuries in the country can be interpreted as a sign of a serious road safety problem. New alternative transportation plans and strategies should be developed to address this issue. In this regard, the road transportation system should be improved, and a significant portion of the country's road transportation should be shifted to the air, railway, and maritime transportation systems. In future work, we seek to use another dataset to improve the performance of models or to test other algorithms not used in the literature.

REFERENCES (KAYNAKLAR)

[1] Al-Radaideh, Q.A. and E.J. Daoud, Data mining methods for traffic accident severity prediction. Int. J. Neural Netw. Adv. Appl, 2018. 5: p. 1-12.

[2] Organization, W.H., Global status report on road safety 2015. 2015: World Health Organization.

[3] Coban, H.H., A. Rehman, and A. Mohamed, Analyzing the Societal Cost of Electric Roads Compared to Batteries and Oil for All Forms of Road Transport. Energies, 2022. 15(5): p. 1925.

[4] Peden, M., et al., World report on road traffic injury prevention. 2004, World Health Organization Geneva.

[5] Daglioglu, N., et al., Determination of phosphatidylethanol (PEth) 16: 0/18: 1 in dried blood samples of drivers involved in traffic accidents: A pilot study. Legal Medicine, 2022: p. 102091.

[6] Ersen, M., A.H. Büyüklü, and S.E. Taşabat, Data Mining as a Method for Comparison of Traffic Accidents in Şişli District of Istanbul. Journal of Contemporary Urban Affairs, 2022. 6(2): p. 113-141.

[7] Sarkar, S., et al., Application of optimized machine learning techniques for prediction of occupational accidents. Computers & Operations Research, 2019. 106: p. 210-224.

[8] Lord, D., S.R. Geedipally, and S.D. Guikema, Extension of the application of Conway-Maxwell-Poisson models: Analyzing traffic crash data exhibiting underdispersion. Risk Analysis: An International Journal, 2010. 30(8): p. 1268-1276.

[9] Geedipally, S.R., D. Lord, and S.S. Dhavala, The negative binomial-Lindley generalized linear model: Characteristics and application using crash data. Accident Analysis & Prevention, 2012. 45: p. 258-265.

[10] Malyshkina, N.V. and F.L. Mannering, Zero-state Markov switching count-data models: An empirical assessment. Accident Analysis & Prevention, 2010. 42(1): p. 122-130.

[11] Lee, A.H., et al., Modeling young driver motor vehicle crashes: data with extra zeros. Accident Analysis & Prevention, 2002. 34(4): p. 515-521.

[12] Chang, L.-Y. and W.-C. Chen, Data mining of tree-based models to analyze freeway accident frequency. Journal of safety research, 2005. 36(4): p. 365-375.

[13] Park, E.S. and D. Lord, Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record, 2007. 2019(1): p. 1-6.

[14] Lv, Y., S. Tang, and H. Zhao. Real-time highway traffic accident prediction based on the k-nearest neighbor method. in 2009 international conference on measuring technology and mechatronics automation. 2009. IEEE.

[15] Abellán, J., G. López, and J. De OñA, Analysis of traffic accident severity using decision rules via decision trees. Expert Systems with Applications, 2013. 40(15): p. 6047-6054.

[16] Dong, N., H. Huang, and L. Zheng, Support vector machine in crash prediction at the level of traffic analysis zones: assessing the spatial proximity effects. Accident Analysis & Prevention, 2015. 82: p. 192-198.

[17] Harb, R., et al., Exploring precrash maneuvers using classification trees and random forests. Accident Analysis & Prevention, 2009. 41(1): p. 98-107.

[18] Zhao, H., et al., Vehicle accident risk prediction based on AdaBoost-so in vanets. IEEE Access, 2019. 7: p. 14549-14557.

[19] Parsa, A.B., et al., Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accident Analysis & Prevention, 2020. 136: p. 105405.

[20] Deretić, N., et al., SARIMA modelling approach for forecasting of traffic accidents. Sustainability, 2022. 14(8): p. 4403.

[21] Brüde, U., What is happening to the number of fatalities in road accidents? A model for forecasts and continuous monitoring of development up to the year 2000. Accident Analysis & Prevention, 1995. 27(3): p. 405-410.

[22] Dadashova, B., et al., Methodological development for selection of significant predictors explaining fatal road accidents. Accident Analysis & Prevention, 2016. 90: p. 82-94.

[23] Doğan, A.A.E. and A.P. ANgüngör, Estimating road accidents of Turkey based on regression analysis and artificial neural network approach. Advances in transportation studies, 2008. 16: p. 11Y22.

[24] Doğan, E. and A. Akgüngör. Investigation of traffic accidents and results with artificial neural networks: Kırıkkale Case. in Proceedings of the 8th Transportation Congress.

[25] Dogan, E., Regression analysis and artificial intelligence approach for traffic accident prediction models in Turkey and selected some great provinces, in Institute of Science and Technology. 2007, Kırıkkale University.

[26] Ceylan, H., An artificial neural networks approach to estimate occupational accident: A national perspective for turkey. Mathematical problems in engineering, 2014. 2014.

[27] Seltman, H., Experimental design and analysis. 2015. Mixed models. A flexible approach to correlated data, 2017: p. 357-377.

[28] Al-Asadi, M.A. and S. Tasdemır, Predict the value of football players using FIFA video game data and machine learning techniques. IEEE Access, 2022. 10: p. 22631-22645.

[29] Al-Asadi, M.A. and S. Tasdemír, Empirical comparisons for combining balancing and feature selection strategies for characterizing football players using FIFA video game system. IEEE Access, 2021. 9: p. 149266-149286.

[30] Kamiński, B., M. Jakubczyk, and P. Szufel, A framework for sensitivity analysis of decision trees. Central European journal of operations research, 2018. 26(1): p. 135-159.

[31] Breiman, L., Random forests. Machine learning, 2001. 45(1): p. 5-32.

[32] Ho, T.K. Random decision forests. in Document analysis and recognition, 1995., proceedings of the third international conference on. 1995. IEEE.

[33] Jiang, R., et al., A random forest approach to the detection of epistatic interactions in case-control studies. BMC bioinformatics, 2009. 10(1): p. S65.

[34] Kuhn, M. and K. Johnson, Applied predictive modeling. Vol. 26. 2013: Springer.

[35] Brownlee, J., Machine learning mastery with python. Machine Learning Mastery Pty Ltd, 2016: p. 100-120.

[36] Willmott, C.J. and K. Matsuura, Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 2005. 30(1): p. 79-82.

[37] Dufour, J.-M., Coefficients of determination. McGill University, 2011.

[38] Brownlee, J. How To Get Baseline Results And Why They Matter. 2014 5/8/2019]; Available from: https://machinelearningmastery.com/how-to-get-baseline-results-and-why-they-matter/.