Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 15(2), 581-592, 2025
Journal of the Institute of Science and Technology, 15(2), 581-592, 2025

ISSN: 2146-0574, eISSN: 2536-4618

Civil Engineering **DOI: 10.21597/jist.1594983**

Research Article

***Received*: 02.12.2024** ***Accepted*: 07.01.2025**

To Cite: Baş, F. İ. (2025). Prediction of Pedestrian Traffic Accident Severity and Evaluation of Driver and Pedestrian Faults. *Journal of the Institute of Science and Technology*, 15(2), 581-592.

Prediction of Pedestrian Traffic Accident Severity and Evaluation of Driver and Pedestrian Faults

Fatih İrfan BAŞ*

Highlights:
- Predicting the number of injured pedestrians in traffic accidents
- Characteristics of pedestrian accidents
- Driver and pedestrian faults

Keywords:
- Accident
- Pedestrian
- Machine learning
- Faults

ABSTRACT:

This study investigated pedestrians involved traffic accidents with the aim of classifying the severity of accidents based on the number of injured pedestrians using machine learning algorithms, including AdaBoost, Gradient Boosting, XGBoost, K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Classifier, Decision Tree, and Random Forest. The Random Forest model was identified as the best model for classifying pedestrian-involved traffic accidents, achieving high predictive accuracy of 95%, an F1 score of 0.95, and demonstrating low error metrics. The research analyzed both driver and pedestrian faults, alongside factors such as the presence of pedestrian crossings, intersection type, driver age, time of day, month and seasonal variations. The results revealed that accidents at locations without intersections were primarily caused by driver faults, such as speeding, while pedestrian faults, such as crossing at unintended locations, also significantly contribute to the overall accident rate. The findings offered valuable insights into the characteristics of pedestrian accidents to improve traffic safety and reduce pedestrian injuries and fatalities.

[1*] Fatih İrfan BAŞ (**Orcid ID:** 0000-0002-0845-060X), Erzincan Binali Yıldırım University, Engineering and Architecture Faculty, Department of Civil Engineering, Erzincan, Türkiye

***Sorumlu Yazar/*Corresponding Author**: Fatih İrfan BAŞ, e-mail: fibas@erzincan.edu.tr

INTRODUCTION

The first recorded pedestrian accident in history occurred 120 years ago when a test car, traveling at a maximum speed of 13 km/h, struck a female pedestrian (Tiwari, 2020). Since then, the rapid increase in the number of automobiles and other motorized vehicles has contributed significantly to a global rise in road fatalities. In 2021, approximately 1.2 million fatalities occurred worldwide due to road accidents, with 92 out of every 100 accidents taking place in low and middle income countries (WHO, 2023a). Notably, for young individuals between 5 and 29 years old, road injuries caused the highest number of deaths, and they had the 12th highest cause of death for all ages (WHO, 2022).

Driver faults are the primary cause of traffic accidents, accounting for more than 90% (Petridou and Moustaki, 2000). Driver faults encompass not only traffic rule violations that jeopardize road safety but also behaviors such as driving alone, driving for extended periods without rest, disregarding fatigue and sleepiness despite being aware of them and driving under the influence of alcohol (Gökdağ and Baş, 2019). These driver behaviors increase the vulnerability of pedestrians to vehicle collisions and pose significant challenges for policymakers in ensuring pedestrian safety.

Traffic accidents worldwide resulted in approximately 1.35 million deaths, with pedestrians accounting for 23% of the fatalities in 2016. In the African region, the pedestrian mortality rate was the highest, at 40%, while the Southeast Asia region had the lowest rate, at 14% (WHO, 2023b). In Türkiye, in 2023, a total of 53,570 traffic accidents occurred, including 2,893 fatal accidents and 50,677 injuries. As a result, 6,548 people lost their lives, with 46% being drivers, 32% passengers, and 22% pedestrians. Of these accidents, 89% were caused by driver faults, while 9% were due to pedestrian faults (KGM, 2024).

Ensuring pedestrian safety and eliminating fatalities and injuries begins with a comprehensive understanding of the causes and circumstances that lead to accidents. Many researchers have focused on the key factors that cause traffic accidents, such as environmental conditions and driver behaviors. Recently, machine learning algorithms have begun to be implemented in traffic accident analyses, particularly using complex and big data. These algorithms help to reveal the relations between the factors contributing to traffic accidents.

Chen et al. (2015) conducted a study to estimate the severity of rear-end collisions using a Bayesian Network and reported that the prediction accuracy of the algorithm was between 60% and 70%. Chen et al. (2016) examined driver injuries in rollover accidents using support vector machine algorithm. Sameen and Pradhan (2017) proposed a Recurrent Neural Network algorithm to estimate the injury severity in traffic accidents. Iranitalab and Khattak (2017) estimated the severity of traffic accidents with several machine learning algorithms and stated that the Nearest Neighbor Classification algorithm demonstrated the best performance. Fountas et al. (2018) investigated the correlation of the features that affect the severity of single vehicle accidents with probit algorithm. Jamal et al. (2021) studied with a few machine learning algorithms and conducted a comparative analysis. They reported that eXtreme Gradient Boosting algorithm had the best performance for prediction of accident severity. Ma et al. (2021) utilized a stacked sparse autoencoder algorithm in their study to identify the factors effecting the severity of traffic accidents. Kuşkapan et al. (2021) conducted an analysis on speed violations of heavy vehicles using Naive Bayes, K-Nearest Neighbors, and support vector machine algorithms. Yang et al. (2022) used a multi task deep neural network to estimate the severity of both fatal and injury-related accidents. Kuşkapan et al. (2022a) performed a study on pedestrian safety at signalized intersections with artificial neural network algorithm. Obasi and Benson (2023) conducted a study to find the parameters that contribute to the severity of accidents. They stated that the Random Forest algorithm had the highest predictions. Çeven and Albayrak (2024) performed a study on accident severity prediction. They reported that the Random Forest algorithm showed better performance than AdaBoost and Multilayer Neural Networks algorithms.

In this study, the severity and the parameters contributing to traffic accidents involving pedestrians were investigated. AdaBoost, Gradient Boosting, XGBoost, K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Classifier, Decision Tree, and Random Forest machine learning algorithms were used in this study. The study identified key patterns and factors influencing pedestrian injuries, such as the presence of pedestrian crossings, intersection type, time of day, weekday, month, season and driver age, with a particular focus on understanding both driver and pedestrian faults that contribute to traffic accidents. The main aim was to take preventive measures and reduce pedestrian accidents, thus improving road safety based on these findings.

MATERIALS AND METHODS

This study examined a dataset of 137 pedestrian accidents that took place in the Diyarbakır city center between 2013 and 2019, sourced from the police department. The accidents resulted in 154 pedestrians being injured, with one pedestrian being killed. Since there was only one pedestrian fatality, the analysis focused solely on the pedestrians who were injured. The accident locations in the city center are shown in Figure 1. The accidents marked with red points represent incidents with one pedestrian injured, yellow points indicate accidents with more than one injured pedestrian, and the green point shows the location of a pedestrian fatality.

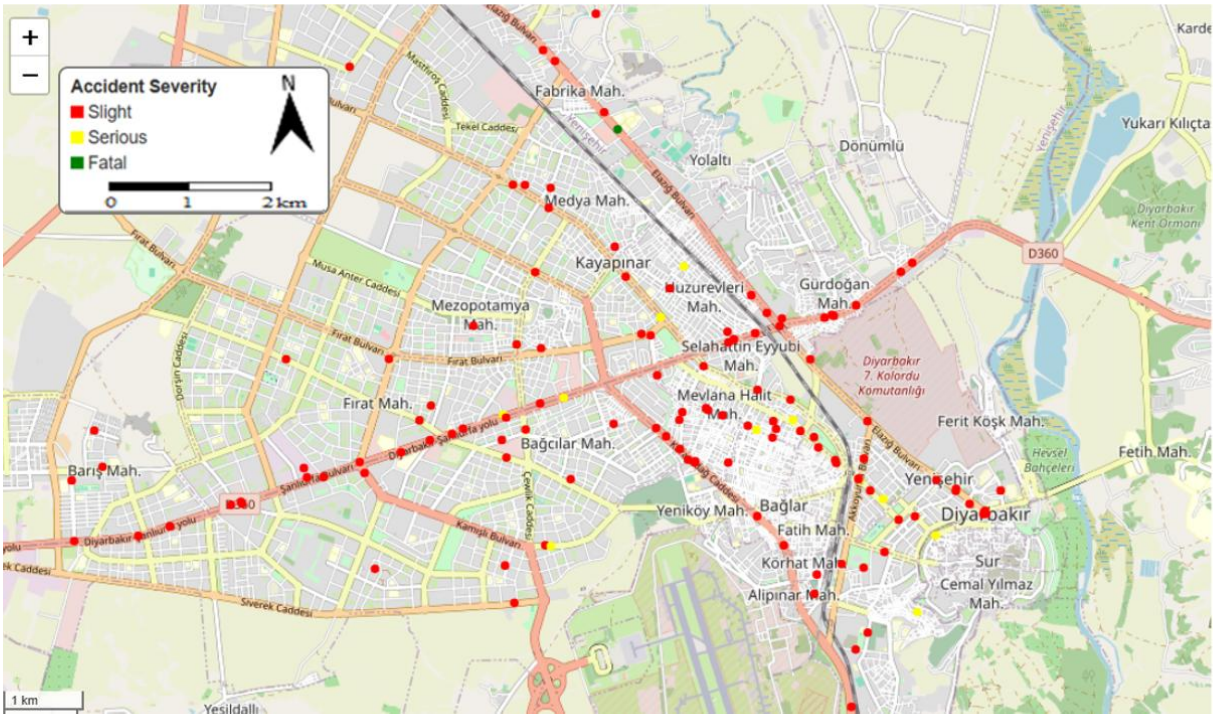

Figure 1. Accident locations in the city center

Each accident was recorded with several parameters, including accident time, date, longitude and latitude, road surface type, road width, intersection details, pedestrian crossing, and other relevant factors. The accidents involving pedestrian injuries were classified into three severity groups: accidents with three injured pedestrians were categorized as severe, accidents with two injured pedestrians as moderate, and accidents with one injured pedestrian as slight. The distributions of accidents were analyzed concerning the presence of pedestrian crossings across different times of day, weekdays, months of the year, and seasons. Furthermore, the distributions of accidents attributed to driver faults were examined across various types of intersections and driver age groups.

Machine learning algorithms were implemented using Python 3.9.12 in a Jupyter Notebook environment. Most data sets in practical life are unbalanced, and one of the most critical factors that increase the prediction success of the machine learning algorithms is that the input classes in a balanced data set are close to each other. Conditional Generative Adversarial Networks (cGANs), the Adaptive Synthetic Sampling Approach (ADASYN), K-Nearest Neighbor Oversampling (KNNOR), and the Synthetic Minority Oversampling Technique (SMOTE) are employed to address imbalanced datasets. These oversampling methods generate synthetic data that closely resemble real data to improve the balance and quality of the dataset. In this study, the dataset used for modeling was imbalanced. To mitigate this issue, SMOTE was applied to generate a balanced dataset. The dataset was then split, and % 80 were used for training, while %20 were used for testing. Kappa statistic, mean absolute error (MAE), and root mean squared error (RMSE) were used to evaluate the error rates of the machine learning algorithms. F1 score, accuracy, precision, and recall metrics were used to evaluate the performance of the algorithms. The calculation of the performance metrics was done in comparison process using the Equations 1-4 given below. The standard confusion matrix has four elements: true positive (TP) which is correct, false positive (FP) which is incorrect, false negative (FN) which is incorrect, and true negative (TN) which is correct.

$$Precision = \frac{True_{positive}}{True_{positive}+False_{positive}} \tag{1}$$

$$Recall = \frac{True_{positive}}{True_{positive}+False_{negative}} \tag{2}$$

$$Accuracy = \frac{True_{positive}+True_{negative}}{True_{positive}+False_{postive}+False_{negative}+True_{negative}} \tag{3}$$

$$F1_{score} = \frac{2\ x\ Precision\ x\ Recall}{Precision+Recall} \tag{4}$$

Machine Learning Algorithms

This study analyzed pedestrian traffic accidents as a classification problem. AdaBoost, Gradient Boosting, XGBoost, K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Classifier, Decision Tree, and Random Forest machine learning algorithms were used in order to predict the number of injured pedestrians in traffic accidents.

AdaBoost

AdaBoost is a machine learning algorithm that enhances weak regression models and establishes stronger models by assigning performance weights. AdaBoost analyzes large data sets quickly and effectively (Xiao et al., 2018). AdaBoost iteratively trains low-performing classifiers and then combines their predictions to increase the classification accuracy (Gamil et al., 2024).

Decision Tree

The Decision tree (DT) is a machine learning algorithm that partitions a dataset into subsets (Acito, 2023; Costa and Pedreira, 2023). Data preprocessing and validation steps should be done properly in order to reduce the overfitting tendency of the algorithm (Pachouly et al., 2022). The Decision Tree algorithm divides each leaf into branches to minimize the Gini value during training (Itzkin et al., 2025).

Gradient Boosting

The Gradient Boosting algorithm uses nonparametric models to reduce the errors in the models that it has previously created. Each parameter used in the prediction corrects the error of the previous one (Otchere et al., 2022). The main purpose of the algorithm is to enhance the performance and accuracy of the final model by combining many weak learners (Yoon, 2021; Kuşkapan et al., 2022b).

K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm assumes that the classification of a case is very similar to its neighbors in the vector space. KNN does not use the previous probabilities and takes the nearest neighboring data k value as the basis for the prediction model in the entire data set (Liao and Vemuri, 2002). In this algorithm, data distribution is analyzed using graphical representations. The algorithm's performance depends on the number of k nearest neighbors, the similarity measurement, and the threshold value (Campisi et al., 2024)

Logistic Regression

Logistic Regression algorithm builds a model to describe the relationship between the dependent and independent variables. LR uses the possible fewest features to achieve the best fit of the prediction model (Coşkun et al., 2004). In this algorithm, the primary focus is on classifying data into distinct groups rather than making precise numerical predictions (Babu Nuthalapati and Nuthalapati, 2024).

Random Forest

Random forest (RF) is a machine learning algorithm that can be applied on classification and regression of the datasets. It is one of the ensemble methods and establishes multiple decision trees using randomly selected data subsets (Sekulić et al., 2020). Random Forest algorithm is widely used for its effectiveness in managing datasets that are imbalanced or have missing values (Ullah et al., 2019).

Support Vector Classifier

The support vector classifier algorithm builds a mathematical model to determine the relationship between dependent and independent variables in a dataset. It identifies the best hyperplane accurately splitting the data points and reducing the errors at the boundaries (Izonin et al., 2021). Support vector classifier is a robust algorithm that can be effectively applied to both classification and binary datasets (Gamil et al., 2024).

XGBoost

XGBoost utilizes a gradient boosting decision tree as its main algorithm and minimizes the errors from previous models. For this reason, it produces highly accurate results (Davagdorj et al., 2020). The hyperparameter settings are generally simple, and its overfitting tendency is lower (Wang and Ni, 2019). Each newly added tree is trained based on the previous trees and recalculates the residual prediction values (Esmaeili-Falak and Benemaran, 2024).

RESULTS AND DISCUSSION

The classification of pedestrian accidents that occurred in the presence of pedestrian crossings according to time of day, weekday, month of year, and season of year is shown in Figure 2. 107 accidents occurred at locations with marked pedestrian crossings, while 30 accidents occurred at locations where no pedestrian crossings were present.

In terms of the time of day, the highest number of accidents occurred during the lunch rush, afternoon, and evening, while fewer accidents were recorded during the morning rush, daytime, and late evening. Regarding weekdays, the number of accidents at marked pedestrian crossings was similar on

Monday, Tuesday, Wednesday, Thursday, and Friday, all of which were higher than the weekend. Sunday had the fewest pedestrian accidents. In terms of the season, fall experienced the highest number of pedestrian accidents, followed by spring and winter, with summer recording the fewest accidents.

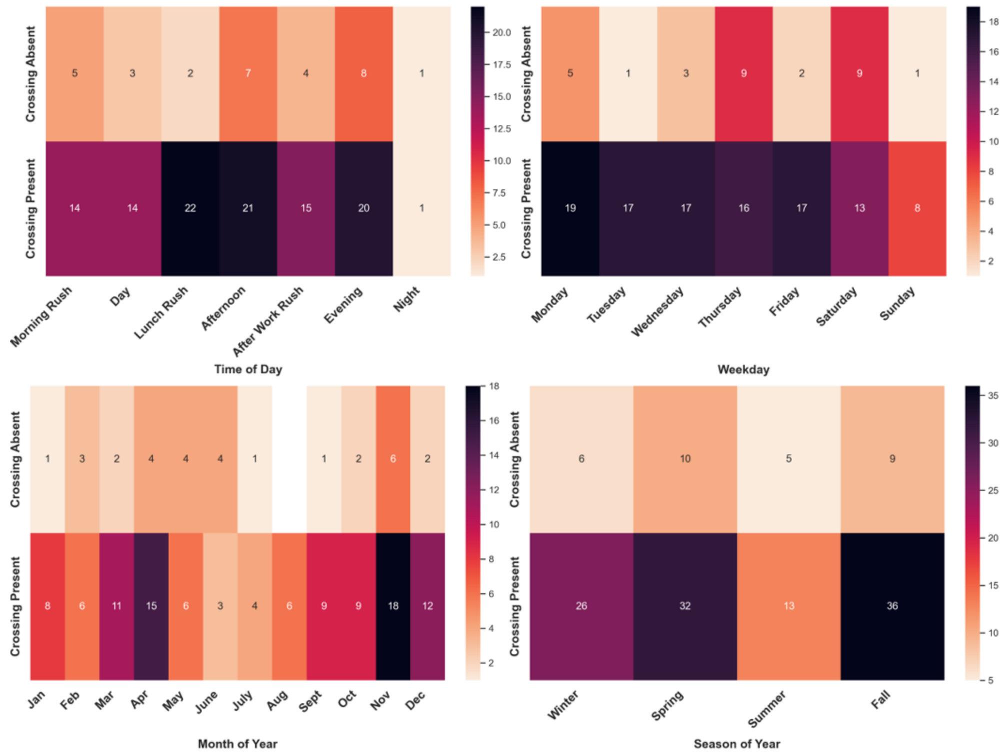

Figure 2. Accident distribution based on presence of pedestrian crossing across time of day, weekday, month of year and season of year

The lunch rush, afternoon, and evening corresponded to periods with higher foot traffic and vehicle volume. During lunchtime, people were more likely to be out walking to have their meals. In the afternoon and evening, pedestrians finished work or school and headed home or daily activities. While the morning rush is characterized by peak vehicle traffic, there were generally fewer pedestrians compared to lunchtime or evening periods. Weekdays saw more pedestrian movement due to people commuting to work, school, or other destinations, resulting in more accidents. In contrast, weekends involved fewer pedestrians, as people were more likely to be shopping, having social activities, or staying at home, which reduced pedestrian mobility. Pedestrian movement was higher in the fall and spring seasons, causing more accidents, while winter season had fewer accidents due to cold weather, which discouraged walking. The summer season had the least number of accidents as people often took holidays, reducing foot traffic.

The distribution of pedestrian accidents based on driver faults across different types of intersections is shown in Figure 3. The highest number of accidents occurred at locations with no intersection, where 64 accidents were recorded. In these cases, the drivers were faultless, and the pedestrians were at fault. The highest number of accidents, 109 in total, occurred at locations without intersections. Of these, 64 accidents were caused by pedestrians at fault, 18 were attributed to drivers failing to adjust their speed appropriately, and 27 resulted from other driver faults. This situation indicates that pedestrian accidents at locations without intersections were primarily the result of driver faults, such as speeding or other faults, rather than pedestrian faults. However, pedestrians were also at fault in many cases, as they attempted to cross at the location without pedestrian crossing or did not follow the road crossing rules.

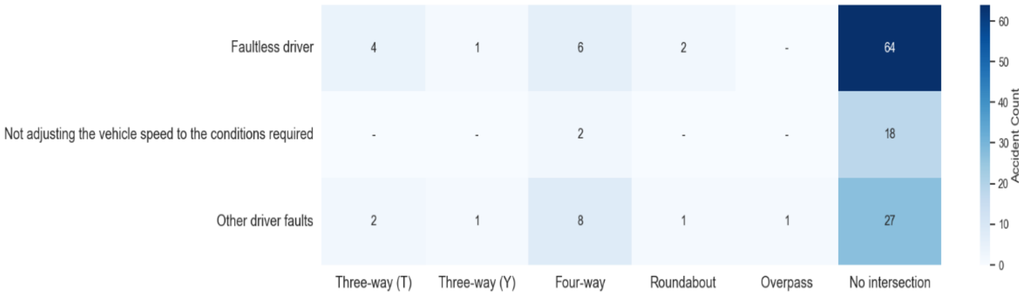

Figure 3. Accident distribution based on driver faults across different types of intersections

The distribution of pedestrian accidents based on driver faults across driver age intersections is given in Figure 4. The heatmap revealed that drivers aged 30-39 were responsible for the highest number of accidents with 35 accidents in the faultless driver category and 17 accidents in the other driver faults category. This suggested that middle-aged drivers were frequently involved in pedestrian accidents, either due to their driving behavior or external factors. Younger drivers (20-29) exhibited a relatively high frequency of accidents caused by failing to adjust their speed to conditions, with eight accidents showing more aggressive driving behavior. On the other hand, older drivers (60-65) had fewer accidents, with only three accidents in the faultless driver category and one accident in the other driver faults category. This situation probably reflects the more careful driving habits that come with age, but it may also result from fewer drivers in the traffic of this age group.

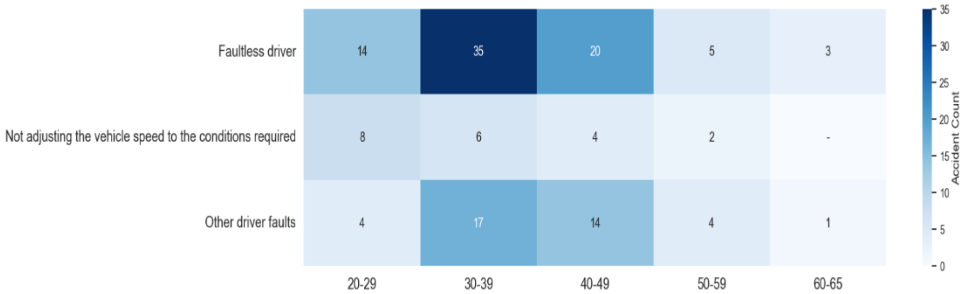

Figure 4. Accident distribution based on driver faults across driver age

The confusion matrices of Gradient Boosting and KNN machine learning algorithms are presented in Figure 5. In the classification scheme, Class 1 represents accidents involving only one injured pedestrian, Class 2 represents accidents with two injured pedestrians, and Class 3 represents accidents with three injured pedestrians. When the Gradient Boosting Confusion matrix was examined, it was seen that the model correctly classified 24 accidents in cases where one pedestrian was injured, 21 accidents in cases where two pedestrians were injured, and 21 accidents in cases where three pedestrians were injured.

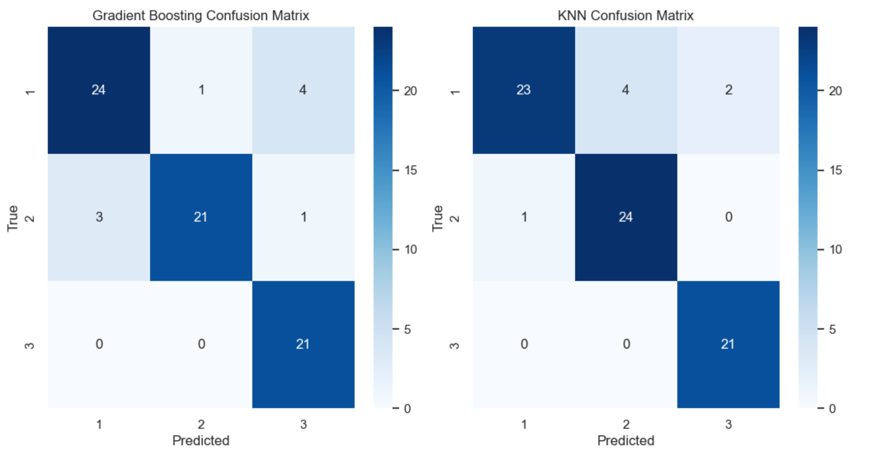

Figure 5. Confusion matrices for Gradient Boosting and KNN algorithms

Learning curves are commonly used to find the algorithm hyper parameters and identify potential issues such as overfitting (Meek et al., 2002). These curves illustrate the relationship between predictive performance and learning effort, show the convergence of the model and signs of overfitting. The curves for Gradient Boosting and KNN algorithms are given in Figure 6. It was observed that both the training and validation accuracy lines increased and converged similarly. The small gaps between the training and validation accuracy lines show that the models did not overfit.

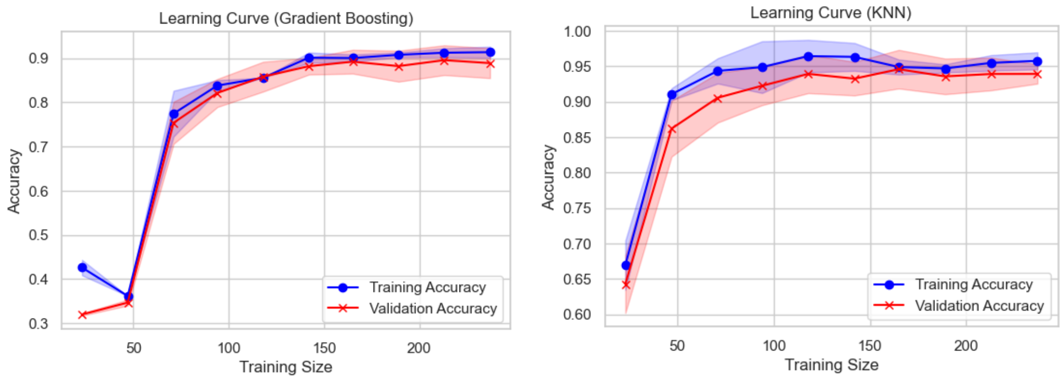

Figure 6. Learning curves of Gradient Boosting and KNN algorithms

The number of injured pedestrians was successfully predicted with low error rates and accuracies between 77% and 95%. The performance of the utilized machine learning algorithms in predicting the number of injured pedestrians was evaluated with accuracy, precision, recall, F1 score, Kappa statistic, MAE, and RMSE metrics and given in Table 1.

As shown in Figure 7, Random Forest algorithm had the highest accuracy with 94.7% and an F1 score of 0.947. The lowest MAE and RMSE values were acquired as 0.067 and 0.306, respectively, with the Random Forest algorithm. AdaBoost algorithm showed the lowest accuracy with 80% and a higher MAE with 0.24. The Logistic Regression algorithm had the lowest performance with 77.3% accuracy, 0.773 F1 score, and higher MAE and RMSE values.

Table 1. Performance metrics for each algorithm

Algorithm	Accuracy	Precision	Recall	F1 Score	Kappa Statistic	MAE	RMSE
KNN	0.907	0.912	0.907	0.909	0.860	0.120	0.416
Decision Tree	0.867	0.875	0.867	0.871	0.801	0.173	0.503
SVC	0.893	0.905	0.893	0.899	0.838	0.133	0.432
Random Forest	0.947	0.947	0.947	0.947	0.919	0.067	0.306
AdaBoost	0.800	0.808	0.800	0.804	0.695	0.240	0.566
XGBoost	0.870	0.874	0.870	0.872	0.799	0.172	0.502
Logistic Regression	0.773	0.773	0.773	0.773	0.657	0.267	0.589
Gradient Boosting	0.880	0.888	0.880	0.884	0.820	0.173	0.529

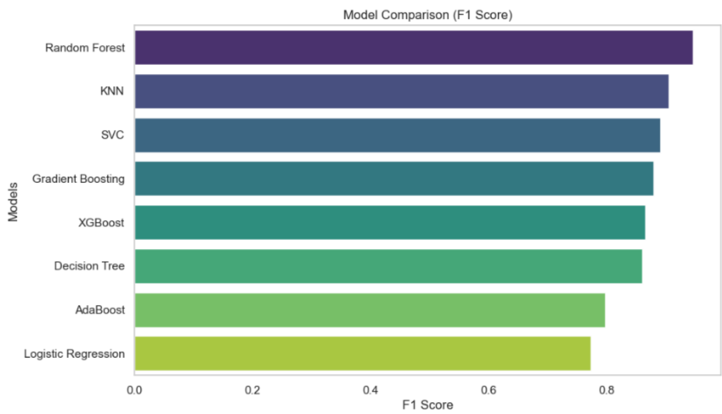

Figure 7. Comparison of F1 scores across models

The Random Forest and KNN algorithms demonstrated the highest accuracies as 95% and 91%, respectively. These findings are consistent with the study carried out by Choi et al. (2020). Random Forest algorithm was used to identify the fatality risk of employees in construction areas and had the best accuracy with %92 among Decision Tree, Logistic Regression, and AdaBoost algorithms. In the study conducted by Raman et al. (2020), Random Forest had the highest accuracy with 79% to predict the severity of industrial accidents. Kang and Ryu (2019) investigated accident types at construction sites. They stated that Random Forest model achieved the highest prediction accuracy with 71%. Sarkar et al. (2017) conducted a study on occupational accident data. They reported that Random Forest algorithm achieved 71% accuracy to predict the injuries and fatalities.

The Random Forest algorithm was found to be particularly effective and reliable in capturing complex patterns and relationships within the data as a valuable tool for predictive analysis in this study.

CONCLUSION

The distribution and causes of pedestrian accidents that occurred in Diyarbakır city center between 2013 and 2019 were investigated in this study. The presence of pedestrian crossings, time of day, weekday, month of year, season of year, driver and pedestrian faults, and driver age factors were evaluated. Severity of accidents based on the number of injured pedestrians was predicted using machine learning algorithms.

The best algorithm for classifying pedestrian traffic accidents was found to be the Random Forest algorithm with high prediction accuracy and minimum error.

The analyses revealed that more accidents occurred at pedestrian crossings during lunch breaks, afternoons, and evenings when foot traffic was heavy. Pedestrian accidents were found to be higher during the fall and spring seasons because of the high pedestrian movement. The number of pedestrian accidents was lower in the winter season because of the reduced pedestrian movement due to the cold weather. Additionally, pedestrian accidents were the least during the summer season as people were on holiday.

Driver faults played a significant role in accidents at locations without intersections. However, pedestrians were also found to be at fault in crossing the street at unexpected points. This situation highlights the importance of raising awareness about pedestrians to avoid crossing streets in areas without pedestrian crossings.

Middle-aged drivers (30-39 years) were found to be responsible for the highest number of accidents, both in faultless and faulty driving. Younger drivers (20-29 years) were found to be more prone to accidents caused by failing to adjust their speed appropriately to the conditions. Older drivers (60-65 years) were involved in fewer accidents, possibly due to their careful driving.

These findings reveal that pedestrian and driver behaviors should be addressed to reduce pedestrian accidents. Increased driver awareness about pedestrians will reduce pedestrian accidents, especially during peak traffic times. Reduced vehicle speed limits in the city center and bumps at pedestrian crossings are recommended to decrease the number of pedestrian injuries and deaths.

Conflict of Interest

The article author declares that there is no conflict of interest.

Author’s Contributions

The sole author made substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data; was involved in drafting the manuscript or revising it critically for important intellectual content; and gave final approval of the version to be published.

REFERENCES

Acito, F. (2023). Classification and Regression Trees. In *Predictive Analytics with KNIME: Analytics for Citizen Data Scientists* (pp. 169-191): Springer.

Babu Nuthalapati, S., and Nuthalapati, A. (2024). Accurate weather forecasting with dominant gradient boosting using machine learning. *Int. J. Sci. Res. Arch*, 12(2), 408-422.

Campisi, T., Kuşkapan, E., Çodur, M. Y., and Dissanayake, D. (2024). Exploring the influence of socio-economic aspects on the use of electric scooters using machine learning applications: A case study in the city of Palermo. *Research in Transportation Business & Management*, 56, 101172.

Chen, C., Zhang, G., Qian, Z., Tarefder, R. A. and Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis Prevention, 90*, 128-139.

Chen, C., Zhang, G., Tarefder, R., Ma, J., Wei, H. and Guan, H. (2015). A multinomial logit model-Bayesian network hybrid approach for driver injury severity analyses in rear-end crashes. *Accident Analysis Prevention, 80*, 76-88.

Choi, J., Gu, B., Chin, S. and Lee, J.-S. (2020). Machine learning predictive model based on national data for fatal accidents of construction workers. *Automation in Construction, 110*, 102974.

Costa, V. G. and Pedreira, C. E. (2023). Recent advances in decision trees: An updated survey. *Artificial Intelligence Review, 56*(5), 4765-4800.

Coşkun, S., Kartal, M., Coşkun, A. and Bircan, H. (2004). Lojistik regresyon analizinin incelenmesi ve diş hekimliğinde bir uygulaması. *Cumhuriyet Üniversitesi Diş Hekimliği Fakültesi Dergisi, 7*(1), 42-50.

Çeven, S. and Albayrak, A. (2024). Traffic accident severity prediction with ensemble learning methods. *Computers Electrical Engineering, 114*, 109101.

Davagdorj, K., Pham, V. H., Theera-Umpon, N. and Ryu, K. H. (2020). XGBoost-based framework for smoking-induced noncommunicable disease prediction. *International journal of environmental research public health, 17*(18), 6513.

Esmaeili-Falak, M. and Benemaran, R. S. (2024). Ensemble extreme gradient boosting based models to predict the bearing capacity of micropile group. *Applied Ocean Research*, 151, 104149.

Fountas, G., Anastasopoulos, P. C. and Abdel-Aty, M. (2018). Analysis of accident injury-severities using a correlated random parameters ordered probit approach with time variant covariates. *Analytic methods in accident research, 18*, 57-68.

Gamil, S., Zeng, F., Alrifaey, M., Asim, M. and Ahmad, N. (2024). An Efficient AdaBoost Algorithm for Enhancing Skin Cancer Detection and Classification. *Algorithms*, 17(8), 353.

Gökdağ, M. and Baş, F. İ. (2019). The Effect of Fatigue and Sleepiness upon Driver Behaviors. *Erzincan University Journal of Science Technology, 12*(2), 850-862.

Iranitalab, A. and Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis Prevention, 108*, 27-36.

Itzkin, M., Palmsten, M. L., Buckley, M. L., Birchler, J. J. and Torres-Garcia, L. M. (2025). Developing a decision tree model to forecast runup and assess uncertainty in empirical formulations. *Coastal Engineering*, 195, 104641.

Izonin, I., Tkachenko, R., Shakhovska, N. and Lotoshynska, N. (2021). The additive input-doubling method based on the SVR with nonlinear kernels: Small data approach. *Symmetry, 13*(4), 612.

Jamal, A., Zahid, M., Tauhidur Rahman, M., Al-Ahmadi, H. M., Almoshaogeh, M., Farooq, D. and Ahmad, M. (2021). Injury severity prediction of traffic crashes with ensemble machine learning techniques: A comparative study. *nternational journal of injury control safety promotion, 28*(4), 408-427.

Kang, K. and Ryu, H. (2019). Predicting types of occupational accidents at construction sites in Korea using random forest model. *Safety Science, 120*, 226-236.

KGM. (2024). *Trafik Kazaları Özeti 2023*: Karayollaı Genel Müdürlüğü, Trafik Güvenliği Dairesi Başkanlığı.

Kuşkapan, E., Çodur, M. Y. and Atalay, A. (2021). Speed violation analysis of heavy vehicles on highways using spatial analysis and machine learning algorithms. *Accident Analysis Prevention, 155*, 106098.

Kuşkapan, E., Sahraei, M. A., Çodur, M. K. and Çodur, M. Y. (2022a). Pedestrian safety at signalized intersections: Spatial and machine learning approaches. *Journal of Transport Health, 24*, 101322.

Kuşkapan, E., Çodur, M. K. and Çodur, M. Y. (2022b). Türkiye'deki Demiryolu Enerji Tüketiminin Yapay Sinir Ağlari İle Tahmin Edilmesi. *Konya Journal of Engineering Sciences*, 10(1), 72-84.

Liao, Y. and Vemuri, V. R. (2002). *Using text categorization techniques for intrusion detection.* Paper presented at the 11th USENIX Security Symposium (USENIX Security 02).

Ma, Z., Mei, G. and Cuomo, S. (2021). An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors. *Accident Analysis Prevention, 160*, 106322.

Meek, C., Thiesson, B. and Heckerman, D. (2002). The learning-curve sampling method applied to model-based clustering. *Journal of Machine Learning Research, 2*(Feb), 397-418.

Obasi, I. C. and Benson, C. (2023). Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents. *Heliyon, 9*(8).

Otchere, D. A., Ganat, T. O. A., Ojero, J. O., Tackie-Otoo, B. N. and Taki, M. Y. (2022). Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science Engineering, 208*, 109244.

Pachouly, J., Ahirrao, S., Kotecha, K., Selvachandran, G. and Abraham, A. (2022). A systematic literature review on software defect prediction using artificial intelligence: Datasets, Data Validation Methods, Approaches, and Tools. *Engineering Applications of Artificial Intelligence, 111*, 104773.

Petridou, E. and Moustaki, M. (2000). Human factors in the causation of road traffic crashes. *European journal of epidemiology, 16*, 819-826.

Raman, P., Kannan, N., Kumar, S. and Raunak, K. (2020). Analysis and prediction of industrial accidents using machine learning. *International Journal of Advanced Science and Technology, 29*, 4990-5000.

Sameen, M. I. and Pradhan, B. (2017). Severity prediction of traffic accidents with recurrent neural networks. *Applied Sciences, 7*(6), 476.

Sarkar, S., Pateshwari, V. and Maiti, J. (2017). *Predictive model for incident occurrences in steel plant in India.* Paper presented at the 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT).

Sekulić, A., Kilibarda, M., Heuvelink, G. B., Nikolić, M. and Bajat, B. (2020). Random forest spatial interpolation. *Remote Sensing, 12*(10), 1687.

Tiwari, G. (2020). Progress in pedestrian safety research. *International journal of injury control safety promotion, 27*(1), 35-43.

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U. and Kim, S. W. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE access,* 7, 60134-60149.

Wang, Y. and Ni, X. S. (2019). A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization. *arXiv, arXiv:1901.08433*.

WHO. (2022). *Preventing injuries and violence: an overview* (9240047131). Retrieved from https://iris.who.int/handle/10665/361331

WHO. (2023a). *Global status report on road safety 2023: summary*: World Health Organization.

WHO. (2023b). *Pedestrian safety: a road safety manual for decision-makers and practitioners, second edition*. Geneva: World Health Organization.

Xiao, L., Dong, Y. and Dong, Y. (2018). An improved combination approach based on Adaboost algorithm for wind speed time series forecasting. *Energy Conversion Management, 160*, 273-288.

Yang, Y., Wang, K., Yuan, Z. and Liu, D. (2022). Predicting Freeway Traffic Crash Severity Using XGBoost-Bayesian Network Model with Consideration of Features Interaction. *Journal of advanced transportation, 2022*(1), 4257865.

Yoon, J. (2021). Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. *Computational Economics*, 57(1), 247-265.