Accident Severity Prediction in Big Data Using Auto-Machine Learning

Tahsin Baykal[1*], Fatih Ergezer[2], Ekinhan Eriskin[3], Serdal Terzi[2]

[1] *Suleyman Demirel University, Graduate School of Natural and Applied Sciences, 32260, Isparta, Turkey, 0000-0001-6218-0826, tahsinbaykal@gmail.com*

[2] *Suleyman Demirel University, Engineering Faculty, Department of Civil Engineering, 32260, Isparta, Turkey, 0000-0001-8034-5743, 0000-0002-4776-824X, fatihergezer@sdu.edu.tr, serdalterzi@sdu.edu.tr*

[3] *Suleyman Demirel University, Property Protection and Security Department, 32260, Isparta, Turkey,* 0000-0002-0087-0933, *ekinhaneriskin@sdu.edu.tr*

Abstract

Estimating the severity of a traffic accident is a problem in motor vehicle traffic because it affects saving human life. If the severity value can be predicted before the accident, all the emergency teams needed could be sent to the area to provide faster first aid. With this aim, we studied a big data set for accidents in the USA between 2016 and 2020, almost 2.25×10^6 rows long. First, the data is preprocessed by removing unnecessary variables. Then rows with blank cells are removed. Finally, about 1.7×10^6 rows length data are left for the prediction process. A machine learning algorithm has been used to classify the severity based on 16 input parameters. Binary-to-decimal count conversation has been used as a novel preprocessing method. As a result, the model has been built with a total accuracy of 0.816. The test results are also validated with precision, recall, and f1-score values. An auto-machine learning model has been developed and trained to predict the severity of a possible traffic accident based on the weather and road conditions. Thus it will be possible to direct emergency units to areas with high accident severity, and preventing a fatality.

Keywords: big data, machine learning, classification, traffic accident, severity

**Corresponding author: tahsinbaykal@gmail.com, +902462113808*

1. Introduction

Accidents are unplanned and uncontrolled events caused by people, situations, environmental factors, or combinations [1]. Road traffic accidents have become the primary source of property damage, health problems, permanent disability, and death. Traffic accidents are a noteworthy public health drawback worldwide, causing 1.35 million deaths and between 20 and 50 million injuries annually [2]. Every day there are 3,000 deaths from traffic accidents worldwide. The number of injured is almost ten times the number of dead, with around 240,000 cases per year. These high rates and high passenger volume imply the need for a comprehensive passenger safety study [3]. Traffic accidents are among the causes of death and injury worldwide [4]. Therefore, it is imperative to predict traffic accidents accurately. Studies on predicting the severity of traffic accidents utilizing artificial intelligence can be found in the literature. Some of these use genetic algorithms (GA) [5], some with artificial neural networks [6-8], some with the Bayesian network [9], some with machine learning (ML) [10-15], some with deep learning (DL) methods [16-17] estimated traffic accident severity.

Angarita-Zapataa et al. (2021) used the Auto- Machine Learning (Auto-ML) model to estimate traffic accident severity using data covering injury accidents and property damage accidents from approximately 220,000 data between 2014 and 2018 in Medellin, Colombia [18]. Alnami et al. (2021) used data mining and machine learning algorithms using the 633372 traffic accident dataset to predict traffic accident severity in Florida [19]. Theofilatos et al. (2019) Compared ML and Deep Learning DL methods to predict real-time accident occurrence [20]. Hashmienejad et al. (2017) estimated traffic accident severity in the Tehran province of Iran over five years (2008-2013) with the data set containing 14211 accidents and the Non-Dominated Sorting Genetic Algorithm (NSGA-II) method. When they evaluated the estimation results, they stated that the NSGA-II method gave an accuracy of 88.2% [5].

Alkheder et al. (2017) used an artificial neural network (ANN) to evaluate the injury severity of traffic accidents based on 5973 traffic accident records between 2008 and 2013 in Abu Dhabi. For each accident record, they collected 48 different characteristics at the time of the accident, and after data preprocessing, the data were reduced to 16 factors and four injury severity classes. They said that ANN classifiers could accurately forecast accident severity [6]. Sameen and Pradhan (2017) developed a DL model using the Recurrent Neural Network (RNN) to calculate the injury severity of traffic accidents based on 1130 accident records between 2009 and 2015 on the North-South Highway (NSE) in Malaysia to estimate [8].

In addition, there are studies on big data and risk management in the literature.

Zhang et al. (2022) used the big data analytics method to assess the risk of ship grounding in operational conditions [21]. Zhang et al. (2021) are used extensively in evaluating ship values [22]. Kaffash et al. (2021) aimed to provide a comprehensive review of the Intelligent Transportation System (ITS) implementation and a review of the best-known models with big data used in ITS. For this, 586 articles were examined between 1997-2019 [23]. Terzi and Erten (2020) examined sustainable transportation and big data examples. They made inferences about how these data will be used in the transportation sector [24].

Based on the literature, although there are accident severity analysis studies based on artificial intelligence methods, as explained above, accident severity estimation based on AutoML using big data is limited. This study proposes an AutoML framework that can be used as a decision support system to direct emergency units before traffic accidents occur. Data on 2.25 million accidents between February 2016 and December 2020 in 49 United States (USA) states were examined for this. These data are State, Temperature, Wind Chill, Humidity, Pressure, Visibility, Wind Speed, Precipitation, Weather Condition, Year, Month, point of interest (POI – amenity, bump, crossing, give way, Junction, no exit, railway, roundabout, traffic calming, traffic signal, and turning loop), Sunrise Number (Sunrise Sunset, Civil Twilight, Nautical Twilight, and Astronomical Twilight), Weekday and hour. Data of missing details are eliminated. Thus, 1.7 million accident data are used to develop the AutoML model. POIs have been converted to numbers for simplification and ease of model analysis. POI numbers were assigned using the binary count system, and AutoML models were developed. No studies have been found in the literature on assigning POI numbers to data with similar characteristics to the binary count system. Model results were validated with accuracy, acuity, recall, and F1-score.

The following section is about the theory of the model, where the AutoML and binary count system, which has been used to convert the POI parameters into numbers, have been specified. Afterward, detailed information about the dataset, such as the source and the content, is given as the practical application. In the fourth section, the analysis findings have been given and discussed. In the last section, a summary of the study has been presented.

2. Theory of the Model

2.1. Auto Machine Learning (Auto-ML)

The ML technique uses various probabilistic, statistical, and optimization algorithms to learn from experience and recognizes valuable patterns of large, unstructured, and complex datasets [25]. ML is a subclass of artificial intelligence that creates a mathematical model based on sample data (training data).ML learns first with the training set and then evaluates the performance of the regressor or classifier with the test set. It then aims to create a regressor or classifier [26]. The problem with Auto-ML is to make a trained process line $\mathrm{M:XY}$, which is hyper parameterized and minimizes the generalization error by automatically generating estimates for samples taken from

$$\mathrm{Pd:GE}\left(\mathrm{M}_{\lambda}\right)=\int\mathcal{L}\left(\mathrm{M}_{\lambda}\left(\mathrm{x}\right),\mathrm{y}\right)\mathrm{P}_{\mathrm{d}}\left(\mathrm{x},\mathrm{y}\right)\mathrm{dxdy} \tag{1}$$

Since a dataset can only be described through a set of n independent observations $D_d = \{(x_1, y_1), ..., (x_n, y_n)\} \sim P_d$, we can only empirically approximate the generalization error experimentally using example data:

$$GE(M_\lambda, D_d) = \frac{1}{|D_d|} \sum_{(x_i, y_i) \in D_d} \mathcal{L}(M_\lambda(x_i), y_i) \tag{2}$$

Auto-ML systems automatically search for the best $M\lambda*$:

$$M_\lambda * \in \underset{\lambda \in \Lambda}{\arg\min}\, GE(M_\lambda, D_{train}) \tag{3}$$

Moreover, estimate *GE*, e.g., by k-fold cross-validation:

$$GE_{CV}(M_\lambda, D_{train}) = \frac{1}{k} \sum_{i=1}^{k} GE\left(M_\lambda^{D_{train}^{(train,i)}}, D_{train}^{(val,i)} \right) \tag{4}$$

Where $M_\lambda^{D_{train}^{(train,i)}}$ mean that M_λ the training was carried out on the ith training fold $D_{train}^{(train,i)}$ [27].

Auto-learn: It uses Bayesian optimization, ensemble choice, and meta-learning to find a promising ML pipeline consisting of an AutoML method, preprocessing techniques, and an ML classifier [18].

2.2. Binary Count System

The point of interest (POI) and daylight status parameters are considered binary numbers in this study. If any POI or daylight status is present on the accident data, the corresponding number is 1, and 0, if not present. The binary number system only uses the values 0 and 1. In the binary number system, each parameter is represented as a power of 2. On the other hand, the decimal number system is used daily and shown using the digits 0 to 9 [28]. By converting the binary to the decimal system, the count of the parameters is decreased. Thus, the computational time needed to run the program can be reduced.

Conversion between binary and decimal counts has been done by multiplying the binary count with two raised to the power of the position of the POI. Here, the right-most POI variable's position starts with 0, and the position number increase by one for each variable standing left.

3. Practical Application

A traffic accident dataset has been applied practically to apply the above-explained theoretical method. The accident dataset comprises approximately $2.25x10^6$ traffic accidents in 49 states of the USA between February 2016 to December 2020 [29]. Figure 1 shows the frequency distribution of traffic accidents in the USA up to December 2020.

The following classifiers were applied in this study: Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting (GrB), Multilayer Perceptrons (MLP), k-Nearest Neighbors (kNN), Gaussian Naive Bayes (GB), Extra Trees, Adaptive Boosting, Latent Dirichlet Allocation (LDA) and Passive-Aggressive Classifiers.

The POI number has been calculated using a binary count system by combining 13 variables: Amenity, Bump, Crossing, Give Way, Junction, No Exit, Railway, Roundabout, Station, Stop, Traffic Calming, Traffic Signal, and Turning Loop (Figure 2). Also, the Sunrise Number has been calculated by combining four variables – Sunrise Sunset, Civil Twilight, Nautical Twilight, and Astronomical Twilight – using a binary count system.

4. Results and Discussion

$2.25x10^6$ data covering 49 states of the USA were filtered to 1.7x10^6 data to estimate traffic accident severity with AutoML. The data used are presented in Table 1. Here the severity levels could be thought of as "No effect accident," "Property Damage Accident," "Injury Accident," and "Fatal Accident" for levels 1 to 4, respectively.

State, Temperature(F), Wind Chill(F), Humidity (%), Pressure(in), Visibility(mi), Wind Speed(mph), Precipitation(in), Weather Condition, Year, Month, Day, POI Number, Sunrise Number, Weekday and Hour parameters were used as inputs in the Auto-ML model and severity was estimated. Statistical information about the data set is given in Table 2. The histograms of the dataset are presented in Figure 3.

In addition, a correlation matrix was created to examine the relationship between the variables used in the AutoML model. The correlation matrix is given in Figure 4.

When Figure 4 is examined, it is seen that the relationship between the input parameters and the output parameter is relatively low.

Auto-ML model uses RF, SVM, GrB, MLP, kNN, GB, Extra Trees, Adaptive Boosting, LDA, and Passive-Aggressive Classifiers algorithms. In order to develop Auto-ML models, the data set is divided into 75% training and 25% test set. The detail of the Auto-ML model is given in Figure 5. RF, with an accuracy value of 0.816, was selected as the most suitable AutoML model. The hyperparameters found as a result of the Auto-ML model are shown in Figure 6.

Accuracy, precision, recall, and f1 score were used to assess the model's performance developed usin precision, recall, and f1 value. The confusion matrix of the model is shown in Figure 7.

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{5}$$

$$\mathrm{Pr\,ecision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

$$\mathrm{Re\,call} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

$$\mathrm{F1Score} = 2 * \frac{\left(\mathrm{Re\,call} * \mathrm{Pr\,ecision}\right)}{\left(\mathrm{Re\,call} + \mathrm{Pr\,ecision}\right)} \tag{8}$$

Where; TP stands for True Positive, TN stands for True Negative, FP stands for False Positive, and FN stands for False Negative.

Figure 7 correctly predicts 3015 of 6955 accidents for severity 1, 317644 of 333055 accidents for severity 2, 24802 of 71625 accidents for severity 3, and 3545 of 15999 accidents for severity 4.

The Accuracy, Precision, Recall, and F1 scores calculated for each severity class are given in Table 3.

When Table 3 is examined, it is seen to give Severity 1 with 0.99 accuracies. According to the F1 score results, Severity 2 gave the best result with 0.89 and severity 1,3,4 with the order of other severity classes.

5. Conclusion

Between 2016 and 2020, more than 2.25 million traffic accidents occurred in the USA. In this study, real traffic and accident data for 49 states of the USA were used to estimate the severity of traffic accidents. Similar variables are combined with the Binary Count System. Auto-ML model was developed for traffic accident severity estimation. The correlation matrix determined the relations of each variable with the others. When the correlation matrix was examined, it was seen that each variable gave a low correlation between severity. The accuracy value was used to determine the best classifier. It has been determined that the best algorithm is RF, with an accuracy value of 0.816. Based on the literature studies, it has been seen that such a comprehensive study has yet to be carried out in almost whole states in the USA. In addition, the severity estimation made with approximately 1.7 million data yielded outstanding results. The severity estimation performed with such extensive data will be a source of inspiration for future studies. The developed Auto-ML model is suggested to be used as a decision support system to quickly direct the emergency units to the scene within a reasonable time according to the traffic accident severity.

Acknowledgment

The author Baykal, T. was supported by the Council of Higher Education's, Türkiye 100/2000 doctoral scholarship.

References

[1] Colling, D. A. "Industrial safety: management and technology", *Prentice Hall* (1990).

[2] International Traffic Safety Data and Analysis Group (Irtad) "Road Safety Annual Report 2020", *International Transport Forum* (2020).

[3] Kashani, A.T. and Mohaymany, A.S. "Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models", *Safety Sci*. 49(10), pp. 1314–1320 (2011).

[4] AlMamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R., et al. "Comparison of machine learning algorithms for predicting traffic accident severity", *In 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, 272-276 (2019, April).

[5] Hashmienejad, S. H. A. and Hasheminejad, S. M. H. "Traffic accident severity prediction using a novel multi-objective genetic algorithm", *International journal of crashworthiness*, 22(4), 425-440 (2017).

[6] Alkheder, S., Taamneh, M. and Taamneh, S. "Severity prediction of traffic accident using an artificial neural network", *Journal of Forecasting*, 36(1), 100-108, (2017).

[7] Rezaie Moghaddam, F., Afandizadeh, S. and Ziyadi, M. "Prediction of accident severity using artificial neural networks" *International Journal of Civil Engineering,* 9(1), 41-48 (2011).

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

[9] Zong, F., Xu, H. and Zhang, H. “Prediction for traffic accident severity: comparing the Bayesian network and regression models”, *Mathematical Problems in Engineering* (2013).

[10] Rasaizadi, A., Sherafat, E. and Seyedabrishami, S. E. “Short-term prediction of traffic state: Statistical approach versus machine learning approach”, *Scientia Iranica*, 29(3), 1095-1106 (2022).

[11] Mohanta, B.K., Jena, D., Mohapatra, N. , et al. “Machine learning-based accident prediction in secure iot enable transportation system”, *Journal of Intelligent & Fuzzy Systems*, (Preprint), 1-13 (2021).

[12] Labib, M.F., Rifat, A.S., Hossain, M.M., et al. “Road accident analysis and prediction of accident severity by using machine learning in Bangladesh”, *In 2019 7th International Conference on Smart Computing & Communications (ICSCC)*, 1-5 (2019, June).

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

[14] Nassiri, H. and Mohamadian Amiri, A. “Prediction of roadway accident frequencies: Count regressions versus machine learning models”, *Scientia Iranica*, 21(2), 263-275 (2014).

[15] Nassiri, H. and Edrissi, A. “Modeling Truck Accident Severity on Two-Lane Rural Highways”, *Scientia Iranica*, 13(2), 193-200 (2006).

[16] Vaiyapuri, T. and Gupta, M. “Traffic accident severity prediction and cognitive analysis using deep learning”, *Soft Computing*, 1-13 (2021).

[17] Yang, Z., Zhang, W. and Feng, J. “Predicting multiple types of traffic accident severity with explanations: A multi-task deep learning framework” *Safety Science*, 146, 105522 (2022).

[18] Angarita-Zapataa, J. S., Maestre-Gongorab, G. and Calderínc, J. F. “A Case Study of AutoML for Supervised Crash Severity Prediction” *In 19th World Congress of the International Fuzzy Systems Association (IFSA), 12th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), and 11th International Summer School on Aggregation Operators (AGOP)*, pp. 187-194. Atlantis Press (2021, August).

[19] Alnami, H. M., Mahgoub, I. and Al–Najada, H. “Highway Accident Severity Prediction for Optimal Resource Allocation of Emergency Vehicles and Personnel”, *In 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, , (pp. 1231-1238) (2021, January).

[20] Theofilatos, A., Chen, C. and Antoniou, C. “Comparing machine learning and deep learning methods for real-time crash prediction”, *Transportation research record*, 2673(8) 169-178 (2019).

[21] Zhang, M., Kujala, P. and Hirdaris, S. “A machine learning method for the evaluation of ship grounding risk in real operational conditions”, *Reliability Engineering & System Safety*, 226, 108697 (2022).

[22] Zhang, M., Montewka, J., Manderbacka, T., et al. “A big data analytics method for the evaluation of ship-ship collision risk reflecting hydrometeorological conditions”, *Reliability Engineering & System Safety*, 213, 107674, (2021).

[23] Kaffash, S., Nguyen, A. T. and Zhu, J. “Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis”, *International Journal of Production Economics*, 231, 107868 (2021).

[24] Terzi S. and Erten K. M. “The effect of big data analysis for sustainable transportation”, *Journal of Innovative Transportation*, 1(1), 1102 (2020).

[25] Mitchell, T. M. “Machine learning WCB”, McGraw-Hill Boston, MA, (1997).

[26] Zhang, X. D. “Machine learning”, *In A Matrix Algebra Approach to Artificial Intelligence (pp. 223-440),* Springer, Singapore (2020).

[27] Feurer, M., Eggensperger, K., Falkner, S., et al. “Auto-sklearn 2.0: The next generation”, *arXiv preprint arXiv:2007*, 04074 (2020).

[28] Mahat, M. S. S. “Number System Conversion for Beginners (Decimal to Binary, Octal and Hexadecimal Conversion)”, *Turkish Journal of Computer and Mathematics Education (TURCOMAT),* 12(14), pp 1445-1458 (2021).

[29] Moosavi, S., Samavatian, M. H., Parthasarathy, S., et al. “Accident risk prediction based on heterogeneous sparse data: New dataset and insights”, *In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 33-42 (2019, November).

Figures and Tables Captions

Figure 1. Frequency Distribution of US-Accidents (2016-2020)

Figure 2. Schematic representation of the calculation of the POI number (*1 if true; 0 if false)

Figure 3. The histograms of the dataset.

Figure 4. Correlation matrix of the dataset

Figure 5. The detail of the Auto-ML model

Figure 6. Auto-ML model hyperparameters

Figure 7. Confusion matrix of the model

Table 1. Data used in accident analysis (These data were compiled by Moosavi et al., [27])

Table 2. The statistical information of the dataset

Table 3. Data on severity classes

Figures

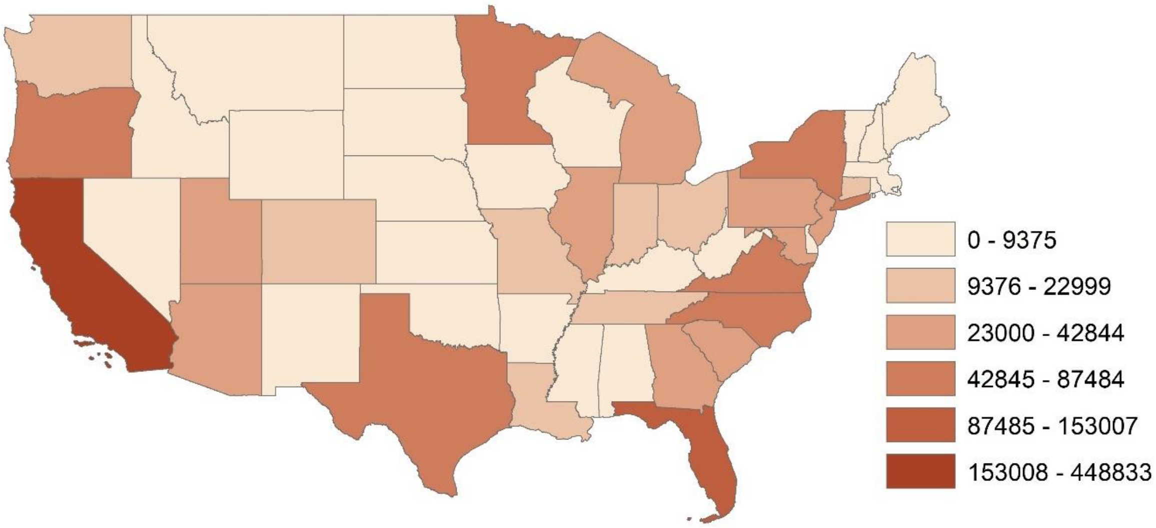


Figure 1. Frequency Distribution of US-Accidents (2016-2020)

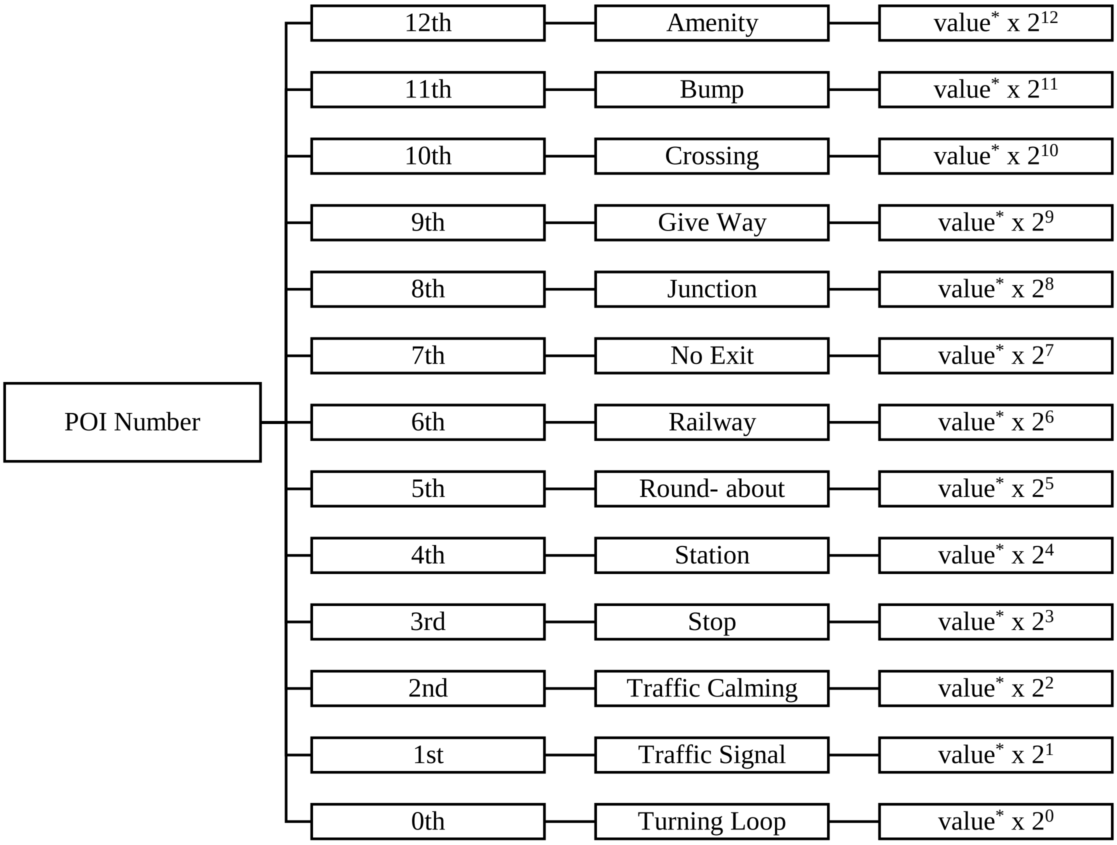


Figure 2. Schematic representation of the calculation of the POI number (*1 if true; 0 if false)

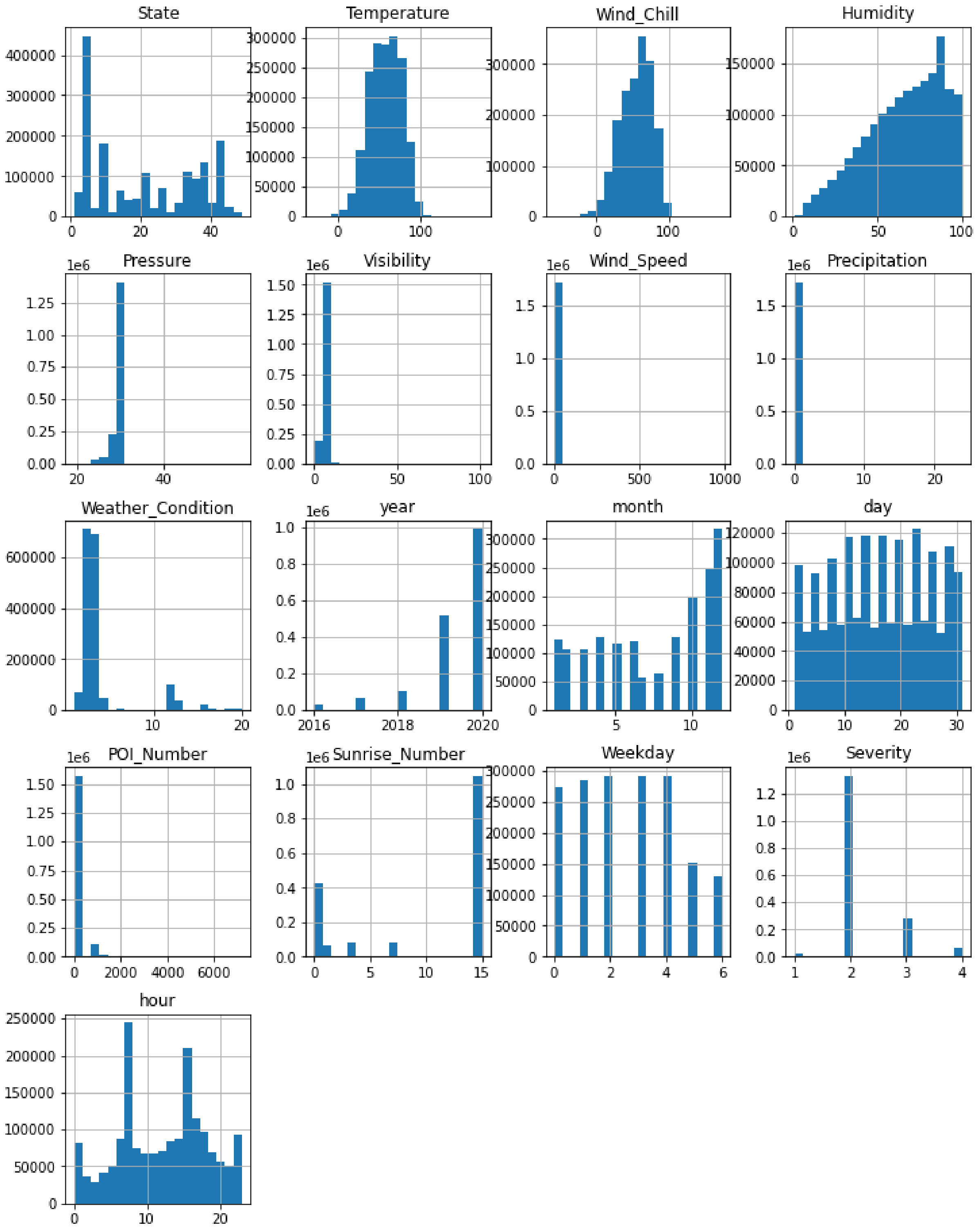
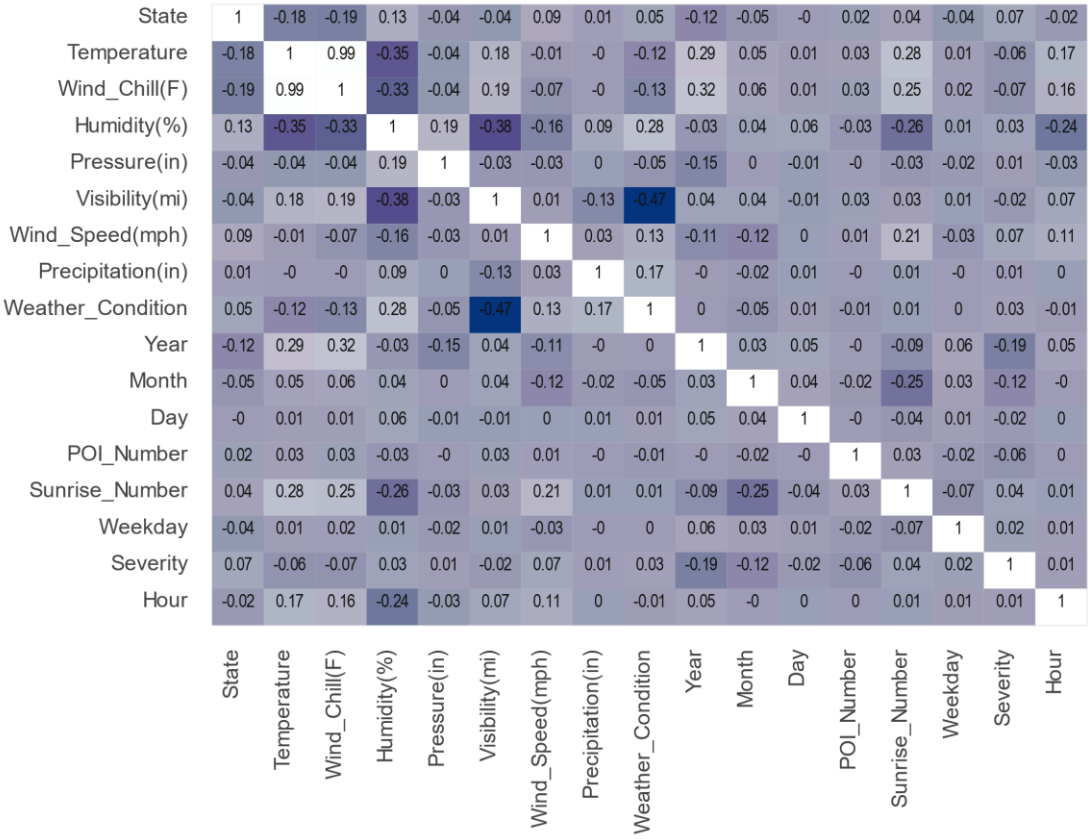


Figure 3. The histograms of the dataset.



	State	Temperature	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Speed(mph)	Precipitation(in)	Weather_Condition	Year	Month	Day	POI_Number	Sunrise_Number	Weekday	Severity	Hour
State	1	-0.18	-0.19	0.13	-0.04	-0.04	0.09	0.01	0.05	-0.12	-0.05	-0	0.02	0.04	-0.04	0.07	-0.02
Temperature	-0.18	1	0.99	-0.35	-0.04	0.18	-0.01	-0	-0.12	0.29	0.05	0.01	0.03	0.28	0.01	-0.06	0.17
Wind_Chill(F)	-0.19	0.99	1	-0.33	-0.04	0.19	-0.07	-0	-0.13	0.32	0.06	0.01	0.03	0.25	0.02	-0.07	0.16
Humidity(%)	0.13	-0.35	-0.33	1	0.19	-0.38	-0.16	0.09	0.28	-0.03	0.04	0.06	-0.03	-0.26	0.01	0.03	-0.24
Pressure(in)	-0.04	-0.04	-0.04	0.19	1	-0.03	-0.03	0	-0.05	-0.15	0	-0.01	-0	-0.03	-0.02	0.01	-0.03
Visibility(mi)	-0.04	0.18	0.19	-0.38	-0.03	1	0.01	-0.13	-0.47	0.04	0.04	-0.01	0.03	0.03	0.01	-0.02	0.07
Wind_Speed(mph)	0.09	-0.01	-0.07	-0.16	-0.03	0.01	1	0.03	0.13	-0.11	-0.12	0	0.01	0.21	-0.03	0.07	0.11
Precipitation(in)	0.01	-0	-0	0.09	0	-0.13	0.03	1	0.17	-0	-0.02	0.01	-0	0.01	-0	0.01	0
Weather_Condition	0.05	-0.12	-0.13	0.28	-0.05	-0.47	0.13	0.17	1	0	-0.05	0.01	-0.01	0.01	0	0.03	-0.01
Year	-0.12	0.29	0.32	-0.03	-0.15	0.04	-0.11	-0	0	1	0.03	0.05	-0	-0.09	0.06	-0.19	0.05
Month	-0.05	0.05	0.06	0.04	0	0.04	-0.12	-0.02	-0.05	0.03	1	0.04	-0.02	-0.25	0.03	-0.12	-0
Day	-0	0.01	0.01	0.06	-0.01	-0.01	0	0.01	0.01	0.05	0.04	1	-0	-0.04	0.01	-0.02	0
POI_Number	0.02	0.03	0.03	-0.03	-0	0.03	0.01	-0	-0.01	-0	-0.02	-0	1	0.03	-0.02	-0.06	0
Sunrise_Number	0.04	0.28	0.25	-0.26	-0.03	0.03	0.21	0.01	0.01	-0.09	-0.25	-0.04	0.03	1	-0.07	0.04	0.01
Weekday	-0.04	0.01	0.02	0.01	-0.02	0.01	-0.03	-0	0	0.06	0.03	0.01	-0.02	-0.07	1	0.02	0.01
Severity	0.07	-0.06	-0.07	0.03	0.01	-0.02	0.07	0.01	0.03	-0.19	-0.12	-0.02	-0.06	0.04	0.02	1	0.01
Hour	-0.02	0.17	0.16	-0.24	-0.03	0.07	0.11	0	-0.01	0.05	-0	0	0	0.01	0.01	0.01	1

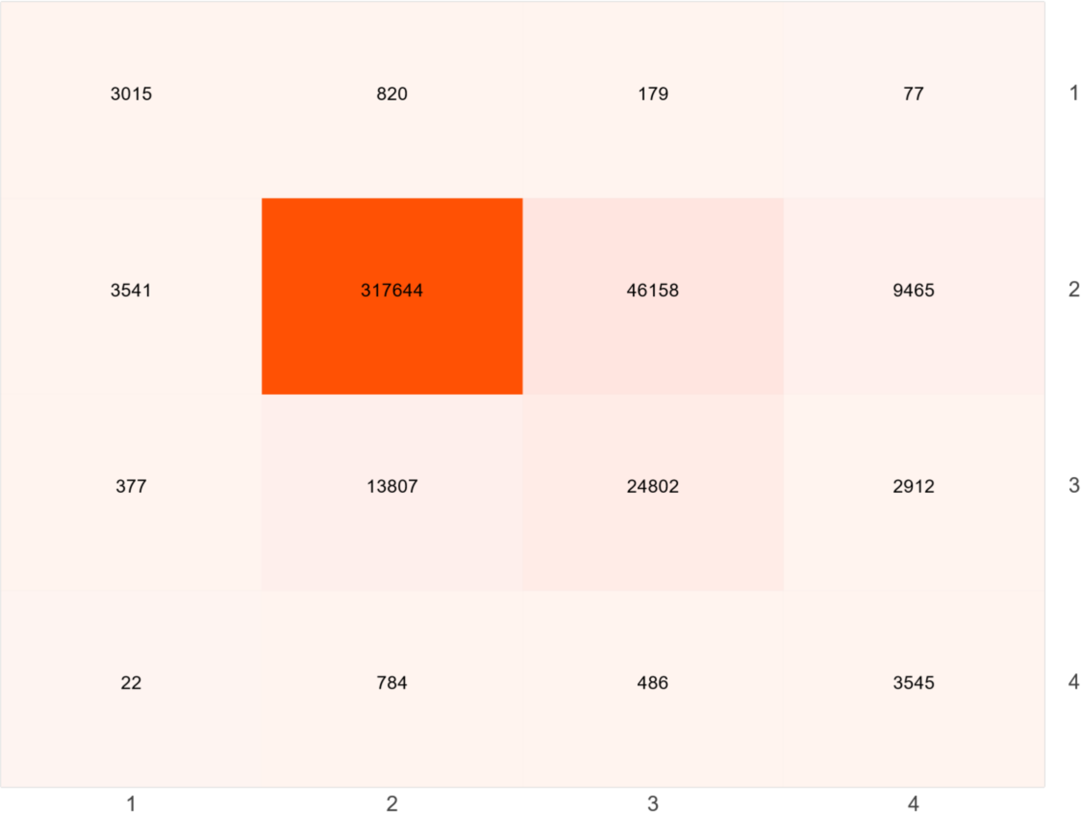
Figure 4. Correlation matrix of the dataset

```
[(0.940000,SimpleClassificationPipeline({'balancing:strategy':'weighting','classifier:__choice__':'sgd','data_preprocessing:categorical_tr
ansformer:categorical_encoding:__choice__':'no_encoding','data_preprocessing:categorical_transformer:category_coalescence:__choice
__':'minority_coalescer','data_preprocessing:numerical_transformer:imputation:strategy':'median','data_preprocessing:numerical_transfo
rmer:rescaling:__choice__':'robust_scaler','feature_preprocessor:__choice__':'extra_trees_preproc_for_classification','classifier:sgd:alph
a':0.00010658622809304751,'classifier:sgd:average':'False','classifier:sgd:fit_intercept':'True','classifier:sgd:learning_rate':'invscaling','cl
assifier:sgd:loss':'perceptron','classifier:sgd:penalty':'l2','classifier:sgd:tol':0.003449144391639749,'data_preprocessing:categorical_trans
former:category_coalescence:minority_coalescer:minimum_fraction':0.007975286334462867,'data_preprocessing:numerical_transform
er:rescaling:robust_scaler:q_max':0.9127392843602231,'data_preprocessing:numerical_transformer:rescaling:robust_scaler:q_min':0.21
922545741477822,'feature_preprocessor:extra_trees_preproc_for_classification:bootstrap':'False','feature_preprocessor:extra_trees_prep
roc_for_classification:criterion':'entropy','feature_preprocessor:extra_trees_preproc_for_classification:max_depth':'None','feature_prepro
cessor:extra_trees_preproc_for_classification:max_features':0.16550428909478176,'feature_preprocessor:extra_trees_preproc_for_class
ification:max_leaf_nodes':'None','feature_preprocessor:extra_trees_preproc_for_classification:min_impurity_decrease':0.0,'feature_prep
rocessor:extra_trees_preproc_for_classification:min_samples_leaf':18,'feature_preprocessor:extra_trees_preproc_for_classification:min
_samples_split':8,'feature_preprocessor:extra_trees_preproc_for_classification:min_weight_fraction_leaf':0.0,'feature_preprocessor:extr
a_trees_preproc_for_classification:n_estimators':100,'classifier:sgd:eta0':8.391437198884603e-06,'classifier:sgd:power_t':
0.6153057870148475},
dataset_properties={
  'task': 2,
  'sparse': False,
  'multilabel': False,
  'multiclass': True,
  'target_type': 'classification',
  'signed': False})),
(0.040000,SimpleClassificationPipeline({'balancing:strategy':'none','classifier:__choice__':'adaboost','data_preprocessing:categorical_tra
nsformer:categorical_encoding:__choice__':'one_hot_encoding','data_preprocessing:categorical_transformer:category_coalescence:__ch
oice__':'no_coalescense','data_preprocessing:numerical_transformer:imputation:strategy':'most_frequent','data_preprocessing:numerical
_transformer:rescaling:__choice__':'quantile_transformer','feature_preprocessor:__choice__':'feature_agglomeration','classifier:adaboost
:algorithm':'SAMME','classifier:adaboost:learning_rate':0.4034077156997028,'classifier:adaboost:max_depth':7,'classifier:adaboost:n_e
stimators':280,'data_preprocessing:numerical_transformer:rescaling:quantile_transformer:n_quantiles':1440,'data_preprocessing:numeri
cal_transformer:rescaling:quantile_transformer:output_distribution':'normal','feature_preprocessor:feature_agglomeration:affinity':'cosin
e','feature_preprocessor:feature_agglomeration:linkage':'average','feature_preprocessor:feature_agglomeration:n_clusters':94,'feature_pr
eprocessor:feature_agglomeration:pooling_func': 'max'},
dataset_properties={
  'task': 2,
  'sparse': False,
  'multilabel': False,
  'multiclass': True,
  'target_type': 'classification',
  'signed': False})),
(0.020000,SimpleClassificationPipeline({'balancing:strategy':'none','classifier:__choice__':'random_forest','data_preprocessing:categoric
al_transformer:categorical_encoding:__choice__':'no_encoding','data_preprocessing:categorical_transformer:category_coalescence:__ch
oice__':'no_coalescense','data_preprocessing:numerical_transformer:imputation:strategy':'mean','data_preprocessing:numerical_transfor
mer:rescaling:__choice__':'quantile_transformer','feature_preprocessor:__choice__':'feature_agglomeration','classifier:random_forest:bo
otstrap':'False','classifier:random_forest:criterion':'entropy','classifier:random_forest:max_depth':'None','classifier:random_forest:max_fe
atures':0.5089615362026388,'classifier:random_forest:max_leaf_nodes':'None','classifier:random_forest:min_impurity_decrease':0.0,'cla
ssifier:random_forest:min_samples_leaf':1,'classifier:random_forest:min_samples_split':11,'classifier:random_forest:min_weight_fractio
n_leaf':0.0,'data_preprocessing:numerical_transformer:rescaling:quantile_transformer:n_quantiles':1422,'data_preprocessing:numerical_
transformer:rescaling:quantile_transformer:output_distribution':'uniform','feature_preprocessor:feature_agglomeration:affinity':'euclidea
n','feature_preprocessor:feature_agglomeration:linkage':'ward','feature_preprocessor:feature_agglomeration:n_clusters':366,'feature_prep
rocessor:feature_agglomeration:pooling_func': 'mean'},
dataset_properties={
  'task': 2,
  'sparse': False,
  'multilabel': False,
  'multiclass': True,
  'target_type': 'classification',
  'signed': False})),
]
```

Figure 5. The detail of the Auto-ML model

```
SimpleClassificationPipeline({'balancing:strategy': 'none', 'classifier:_choice': 'random_forest',
'data_preprocessing:categorical_transformer:categorical_encoding:__choice': 'no_encoding',
'data_preprocessing:categorical_transformer:category_coalescence:__choice': 'no_coalescense',
'data_preprocessing:numerical_transformer:imputation:strategy': 'mean',
'data_preprocessing:numerical_transformer:rescaling:__choice': 'quantile_transformer',
'feature_preprocessor:__choice_': 'feature_agglomeration', 'classifier:random_forest:bootstrap': 'False',
'classifier:random_forest:criterion': 'entropy', 'classifier:random_forest:max_depth': 'None',
'classifier:random_forest:max_features': 0.5089615362026388, 'classifier:random_forest:max_leaf_nodes':
'None', 'classifier:random_forest:min_impurity_decrease': 0.0, 'classifier:random_forest:min_samples_leaf':
1, 'classifier:random_forest:min_samples_split': 11, 'classifier:random_forest:min_weight_fraction_leaf': 0.0,
'data_preprocessing:numerical_transformer:rescaling:quantile_transformer:n_quantiles': 1422,
'data_preprocessing:numerical_transformer:rescaling:quantile_transformer:output_distribution': 'uniform',
'feature_preprocessor:feature_agglomeration:affinity': 'euclidean',
'feature_preprocessor:feature_agglomeration:linkage': 'ward',
'feature_preprocessor:feature_agglomeration:n_clusters': 366,
'feature_preprocessor:feature_agglomeration:pooling_func': 'mean'},

dataset_properties={

  'task': 2,

  'sparse': False,

  'multilabel': False,

  'multiclass': True,

  'target_type': 'classification',

  'signed': False})),
```

Figure 6. Auto-ML model hyperparameters



1	2	3	4	
3015	820	179	77	1
3541	317644	46158	9465	2
377	13807	24802	2912	3
22	784	486	3545	4

Figure 7. Confusion matrix of the model

Tables

Table 1. Data used in accident analysis (These data were compiled by Moosavi et al., [27])

Data	Definition
Severity	A number between 1 and 4 indicates the severity of the accident. Here, 1 indicates the most negligible impact on traffic, and 4 shows the most significant impact.
State	The state in the address field
Temperature(F)	Displays the temperature
Wind Chill(F)	Wind indicates cold
Humidity (%)	Displays humidity
Pressure(in)	Indicates air pressure
Visibility(mi)	Shows the visibility
Wind Speed(mph)	Shows wind speed
Precipitation(in)	Shows the amount of precipitation, if any
Weather Condition	Weather (rain, snow, storm, fog, etc.)
Year	Indicates the year in which the accident occurred.
Month	Indicates the month in which the accident occurred.
Day	Indicates the day the accident occurred.
POI Number	POI parameters are taken as position numbers to calculate the decimal number from the binary system. Thus, the 13 input parameters are clustered into 1 to improve the system's performance.
Sunrise Number	Day/night parameters are taken as position numbers to calculate the decimal number from the binary system. Thus, the four input parameters are clustered into 1 to improve the system's performance.
Weekday	Indicates the weekday on which the accident occurred.
Hour	Indicates the time when the accident occurred.

Table 2. The statistical information of the dataset

	count	mean	std	min	%25	%50	%75	max
State	1710535	20.79	15.19	1.00	4.00	19.00	36.00	49.00
Temperature(F)	1710535	56.98	19.45	-29.00	43.00	58.00	72.00	174.00
Wind Chill(F)	1710535	55.04	22.17	-59.00	39.00	58.00	72.00	174.00
Humidity (%)	1710535	65.94	22.79	1.00	50.00	69.00	85.00	100.00
Pressure(in)	1710535	29.43	1.09	19.37	29.22	29.76	30.00	58.04
Visibility(mi)	1710535	9.01	2.82	0.00	10.00	10.00	10.00	101.00
Wind Speed(mph)	1710535	7.36	5.54	0.00	3.00	7.00	10.00	984.00
Precipitation(in)	1710535	0.01	0.07	0.00	0.00	0.00	0.00	24.00
Weather Condition	1710535	3.64	3.43	1.00	2.00	3.00	3.00	20.00
year	1710535	2019.39	0.89	2016	2019	2020	2020	2020
month	1710535	7.61	3.75	1.00	4.00	9.00	11.00	12.00
day	1710535	16.14	8.65	1.00	9.00	16.00	23.00	31.00
POI Number	1710535	144.74	523.75	0.00	0.00	0.00	2.00	7180.00
Sunrise Number	1710535	9.70	6.82	0.00	0.00	15.00	15.00	15.00
Weekday	1710535	2.59	1.82	0.00	1.00	3.00	4.00	6.00
hour	1710535	12.15	6.01	0.00	7.00	13.00	17.00	23.00
Severity	1710535	2.23	0.53	1.00	2.00	2.00	2.00	4.00

Table 3. Data on severity classes

	Accuracy	Precision	Recall	F1 score
Severity 1	0.99	0.74	0.43	0.55
Severity 2	0.83	0.84	0.95	0.89
Severity 3	0.31	0.59	0.35	0.44
Severity 4	0.21	0.73	0.22	0.34

T. Baykal

Baykal graduated with an MSc in Civil Engineering from Isparta Applied Sciences University in 2019. In 2019, he started his Ph.D. education at the Department of Civil Engineering at Süleyman Demirel University as a scholarship holder of 100/2000 Doctorate program of the Higher Education Institution, and he's still continuing his studies now. His research interests include smart transportation systems, artificial intelligence models, and geographic information systems.

F. Ergezer

Ergezer has been working as a Research Assistant at Süleyman Demirel University since 2017. He began his MSc education in the field of Transportation Engineering in Civil in 2017 and graduated in 2019. In 2019, he started his Ph.D. education in the Department of Civil Engineering at Süleyman Demirel University and is still continuing. His areas of study are related to transportation, specifically asphalt pavements and intelligent transportation systems.

E. Eriskin

Eriskin completed his PhD in 2019 on the area of Transportation Engineering. Currently he is working as Assistant Professor in the Department of Property Protection and Security at the Suleyman Demirel University. His research interest is about traffic engineering, transportation design and safety.

S. Terzi

Terzi got his PhD degree at 2004 in the area of Transportation Engineering in Suleyman Demirel University. He is working since 2012 as full time Professor in Civil Engineering department. His research interest focus on Intelligent Transportation Systems, bituminous pavements, and artificial intelligence applications in transportation engineering.