International Journal of
Intelligent Systems and
Applications in Engineering
ISSN:2147-6799 www.atscience.org/IJISAE
Advanced Technology and Science
Original Research Paper

# Predicting the Severity of Motor Vehicle Accident Injuries in Adana-Turkey Using Machine Learning Methods and Detailed Meteorological Data

**Çiğdem ACI[*1], Cevher ÖZDEN[2]**

**Accepted : 16/02/2018 Published: 30/03/2018**

***Abstract:*** Traffic accidents are among the most important issues facing every nation in the world as they cause many deaths and injuries as well as economic losses every year. In this study, the traffic accidents that took place in Adana, have been classified according to injury severity (i.e. fatal or non-fatal) and the factors affecting the accident outcome are investigated. The study included the traffic accident reports kept by Regional Traffic Division and the weather data provided by the Regional Directorate of Meteorology during 2005-2015. Five major machine learning methods (i.e. k-Nearest Neighbor, Naive Bayes, Multilayer Perceptron, Decision Tree, Support Vector Machine) and one statistical method, Logistic Regression, were employed for prediction models and performances of the models as well as the effective parameters were compared. The main objective of the study is to determine how important weather and other phenomena are for the occurrence of traffic accidents. Decision Tree, k-Nearest Neighbor, and Multilayer Perceptron based models yielded higher accuracy in classification of accidents compared to other models. Furthermore, in Area Under Curve based analysis of factor importance, it was determined that Mean Cloudiness, Existence of Traffic Control and Ground Surface Temperature had higher positive effects, while Maximum Temperature and Weather (kept by traffic officers) parameters decreased the accuracy of models.

***Keywords:*** *injury severity, machine learning, prediction, predictor importance, traffic accident.*

## 1. Introduction

Traffic accidents kill around 1.3 million people every year in the world. In addition, 20 to 50 million people suffer non-fatal injuries, with many incurring a disability as a result of their injury [1]. Also, traffic injuries cause considerable economic losses to victims, their families, and to nations as a whole. These losses arise from the cost of treatment (including rehabilitation and incident investigation) as well as reduced/lost productivity (e.g. in wages) for those killed or disabled by their injuries, and for family members who need to take time off work (or school) to care for the injured. There are few global estimates of the costs of injury, but research carried out in 2010 suggests that road traffic crashes cost countries approximately 3% of their gross national product. This figure rises to 5% in some low- and middle-income countries [2].

In Turkey's case, the number of motor vehicles increases every year, nearly doubled between 2005 and 2015 [3]. However, the number of traffic accidents increased faster compared to vehicle number in the same period, which resulted in a higher ratio of killed and injured persons in the total population. Nearly 1% of the total population dies and 4% suffers from injuries due to traffic accidents, which have become an important risk factor of life in Turkey.

According to the records kept by officers after fatal traffic accidents in 2015, driver faults accounted for 89.30% of total faults, pedestrian faults 8.80%, road defects 0.91%, vehicle defects 0.55% and passenger's faults 0.43% [3]. The records were kept only for fatal injuries and lacked many possible additional factors that could contribute to the occurrence of the accidents, one of which is weather.

The relationship between traffic accidents and weather condition is a well-known fact. A number of studies have attempted to develop injury severity models using weather data. The previous studies in this area can be divided into two main categories from the methodological perspective: statistically based models and the machine learning based models: The statistical-based models explore the characteristics of crashes using Logistic Regression (LR). Reference [4] used Negative Binominal modeling technique to model the frequency of the accident occurrences and involvements over 1.606 accidents on a principal highway in Florida-USA. They used road and driver characteristics as explanatory variables. Their result showed that heavy traffic volume, speeding, narrow lane width, larger number of lanes, urban roadway sections, narrow shoulder width and reduced median width increase the likelihood of accident occurrence. They also reported that female drivers experience more accidents than male drivers in heavy traffic volume and younger drivers have a greater tendency of being involved in accidents. Reference [5] analyzed the pattern of traffic accidents based on several severity types. They included a total of 11.564 accidents reported in Seul-Korea and 22 factors such as vehicle and road characteristics. They employed Multilayer Perceptron (MLP), LR and Decision Tree Classifier (DTC) models to classify accidents into three main subgroups, (1) death or major injury, (2) minor injury and (3) property damage only. Consequently, they observed no significant difference in the classification accuracy of the models. Reference [6] analyzed driver injury severities for single-vehicle crashes occurring in rural and urban areas using data collected in New Mexico from 2010 to 2011. They used nested logit models and mixed logit models to identify contributing factors for driver injury

[1] *Dept. of Compt. Eng., Mersin University, 33343, Mersin, TURKEY*
[2] *Dept. of Compt. Eng., Çukurova University, 01330, Adana, TURKEY*
[*] *Corresponding Author: Email: caci@mersin.edu.tr*

severities. The data used in the study include weather information such as Clear, Fog, Rain etc. They identified five factors only significant for the rural model, including animal involved crashes, rainy condition, icy condition, crashes in no passing zone and pick-up involve crashes. On the other hand, they determined six factors significantly influencing driver injury severity in urban crashes, which includes crashes during peak hours, curved roadways, roadways with multi-lanes, tractor involved crashes, drug-impaired drivers, and drivers between 16 and 20-year-old.

Machine learning based models have also been widely used in predicting the severity of road traffic crashes: Reference [7] analyzed 971 traffic accidents occurred in Abu Dhabi in 2014, consisting of 121 fatal and 135 severe injuries. They employed DTC, Rule Induction, Naive Bayes Classifier (NBC) and MLP methods. The results indicated that key factors associated with fatal severity were age, gender, nationality, year of the accident, casualty status and collision type and 18-30 years old group as the most vulnerable group. Reference [8] investigated the effects of certain traffic and weather parameters on the likelihood of a secondary accident following the occurrence of a traffic accident. They employed MLP and logit models. They identified that traffic speed, duration of the primary accident, hourly volume, rainfall intensity and a number of vehicles involved in the primary accident are the top five factors associated with the secondary accident likelihood. In addition, changes in traffic speed and volume, number of vehicles involved, blocked lanes, and percentage of trucks and upstream geometry also influence the probability of having a secondary incident. Reference [9] analyzed a total of 1.536 accidents on rural highways in Spain using Bayesian Network models. They aimed to determine the effects of several factors including driver characteristics, highway features, vehicle characteristics, accidents and weather parameters on accident severity. Consequently, they identified that accident type, driver age, lightning and number of injuries are most associated with accident severity. Reference [10] investigated the application of MLP, DTC and a hybrid combination of DTC and MLP to build models that could predict injury severity. Their dataset contained traffic accident records from 1995 to 2000, a total number of 417.670 cases. The total set included labels of year, month, region, primary sampling unit, the number describing the police jurisdiction, case number, person number, vehicle number, vehicle make and model; inputs of drivers' age, gender, alcohol usage, restraint system, eject, vehicle body type, vehicle age, vehicle role, initial point of impact, manner of collision, rollover, roadway surface condition, light condition, travel speed, speed limit and the output injury severity. The injury severity had five classes: no injury, possible injury, non-incapacitating injury, incapacitating injury, and fatal injury. Their results revealed that, for the non-incapacitating injury, the incapacitating injury, and the fatal injury classes, the hybrid approach performed better than MLP, DTC and Support Vector Machines (SVM). For the no injury and the possible injury classes, the hybrid approach performed better than MLP. The no injury and the possible injury classes could be best modeled directly by DTC. Reference [11] deals with some classification models to predict the severity of the injury that occurred during traffic accidents. For this purpose, they used the dataset that contains 34.575 accident cases belonging to the year 2008 produced by the transport department of the government of Hong Kong. They employed Naive Bayes, J48, AdaBoostM1, PART and Random Forest classifiers for predicting classification accuracy. They used Genetic Algorithm for feature selection to reduce the dimensionality of the dataset. They investigated three different cases such as Accident, Casualty, and Vehicle for finding the cause of the accident and the severity of the accident. Their final result showed that the Random Forest outperformed other four algorithms. Reference [12] investigated common features between accidents. They researched road accident data of major national highways that pass through Krishna district for the year 2013 by applying machine learning techniques. They formed clusters using K-medoids, and applied expectation maximization algorithms to discover hidden patterns using a priori algorithm. Their aim was to generate association rules that could analyze how to discover hidden patterns that are the root causes for accidents among different combinations of attributes of a larger dataset. They used density histograms for visualizing region-wise such as fatal versus weather, fatal versus time, time versus day, fatal versus month, fatal versus traffic, and fatal versus age. Their results showed that the selected machine learning techniques are able to extract hidden patterns from the data.

The motivation of this study is to examine the role of detailed meteorological weather reports in determining the results of (fatal or non-fatal) motor vehicle accidents. It is known that weather affects every single dimension of our daily life, even our moods. However, weather condition information in traffic accident datasets is kept very simple in previous studies. The injury severity of accidents can be estimated accurately if detailed meteorological weather reports could be combined with accident records.

In this work, machine learning based prediction models are developed to estimate the results of the accidents occurred in Adana (a southern city of Turkey); in addition, LR method is also used to give a statistical comparison basis for machine learning methods.

## 2. MATERIAL AND METHODS

### 2.1. Accident Data and Meteorological Data

This study is conducted based on ten-year crash data consisting of fatal and non-fatal traffic accidents and meteorological records collected in Adana from 2005 to 2015, provided by the General Directorate of Security-Traffic Services Department and Turkish State Meteorological Service. The dataset is composed of two major sub-datasets: The first one includes Day of Week, Crash Time Period, Location, Division of Road, Roadway Surface, Weather Information, Traffic Control, Pavement Marking, Shoulder, Slope, and Crossing. This dataset consists of 25.015 accident record, of which only 246 are fatal and the rest non-fatal. Due to the unbalanced distribution of the original accident records, it would be impossible to develop accurate prediction models because any method can just classify all cases as nonfatal and still achieve over 90% accuracy. Therefore, we kept all the fatal accident records in the dataset and arbitrarily reduced the size of non-fatal accidents to three-fold (1:3) and one-fold (1:1) of the fatal accidents. We obtained two different datasets after this process: the first one consisted of 246 fatal (25%) and 738 non-fatal (75%) accidents, while the second one included 246 fatal (50%) and 246 non-fatal (50%) accidents. In this way, all fatal accidents were included in both dataset, and nonfatal accidents were randomly selected. Then, 10-fold cross-validation was used for both datasets before the application of each method to eliminate chance factor. The detailed meteorological data is obtained from the Turkish State Meteorological Service. The parameters used are Mean Wind Speed (m/sec), Mean Pressure (hPa), Maximum Temperature (°C), Minimum Temperature (°C), Mean Cloudiness, Mean Relative Humidity (%), Total Global Solar Radiation (cal/cm²), Total Precipitation (mm) and Ground Surface Temperature (°C). All meteorological data are daily measured.

Table 1 shows descriptive statistics of the meteorological dataset. Considering the period of the study 2005-2015, only these parameters fulfilled the requirement of continuity. Some part of the meteorological observations had so many missing values that it would be impossible to complete the series with statistical methods.

**Table 1.** Descriptive Statistics for Meteorological Data

| Attribute Name | Min. | Max. | Mean | St.Dev. |
|---|---|---|---|---|
| Mean Wind Speed (m/sec) | 0.0 | 4.2 | 1.2 | 0.5 |
| Mean Pressure (hPa) | 999.9 | 1027.4 | 1010.6 | 5.2 |
| Maximum Temperature(°C) | 8.0 | 39.9 | 26.7 | 7.6 |
| Minimum Temperature (°C) | -3.0 | 27.6 | 15.5 | 7.1 |
| Mean Cloudiness* | 0.0 | 10.0 | 3.8 | 1.5 |
| Mean Relative Humidity (%) | 27.8 | 95.3 | 70.2 | 12.9 |
| Total Global Solar Radiation (cal/cm²) | 29.4 | 673.8 | 390.4 | 137.8 |
| Total Precipitation (mm) | 0.0 | 53.0 | 1.1 | 4.9 |
| Ground Surface Temperature (°C) | -6.1 | 26.4 | 13.2 | 7.8 |

*** Mean cloudiness is measured based on the division of the sky into 10 equal parts. Thus, 10.0 means fully cloud covered the sky, while 0.0 means clear cloudless sky.**

### 2.2. Naive Bayes Classifier (NBC)

Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$ and $P(x|c)$. NBC assumes that the effect of the value of a predictor ($x$) on a given class ($c$) is independent of the values of other predictors. This assumption is called class conditional independence [13-14].

$$P(y|\vec{x}) = \frac{P(\vec{x}|c)P(y)}{P(\vec{x})} \quad (1)$$

- $P(y|\vec{x})$ is the posterior probability of class (target) given predictor (attribute).
- $P(y)$ is the prior probability of class.
- $P(\vec{x}|c)$ is the likelihood which is the probability of predictor given class.
- $P(\vec{x})$ is the prior probability of predictor.

Estimating $P(\vec{x}|y)$, however, is not easy. The additional assumption is the *Naive Bayes* assumption:

$$P(\vec{x}|y) = \prod_{\alpha=1}^{d} P(x_\alpha|y) \quad (2)$$

$$h(\vec{x}) = argmax_y P(y|\vec{x}) \quad (3)$$

$$= argmax_y \frac{P(\vec{x}|y)\,P(y)}{P(\vec{x})} \quad (4)$$

$$= argmax_y P(\vec{x}|y)P(y) \quad (5)$$

$$= argmax_y \prod_{\alpha=1}^{d} P(X_\alpha|y)P(y) \quad (6)$$

$$= argmax_y \sum_{\alpha=1}^{d} log\big(P(x_\alpha|y)\big) + log(P(y)) \quad (7)$$

Estimating $\log(P(x\alpha|y))$ is easy as we only need to consider one dimension. And estimating $P(y)$ is not affected by the assumption.

### 2.3. k-Nearest Neighbor (kNN) Method

The kNN methodology relies on a simple distance learning approach whereby an unknown member is classified according to the majority of its k-nearest neighbors in the training set. The nearness is measured by an appropriate distance metric [15]. It is used for classifying objects based on the closest training examples in the feature space. kNN algorithm is among the simplest of all machine learning algorithms. In the classification process, the unlabeled query point is simply assigned to the label of its k-nearest neighbors. Typically, the object is classified based on the labels of its k-nearest neighbors by majority vote [13], [16]. If k equals 1, the object is simply classified as the class of the object nearest to it. When there are only two classes, k must be an odd integer. However, there can still be a tie when k is an odd integer during multiclass classification. In the study, Euclidean distance is used as the distance function for kNN:

$$d(x,y) = \|x - y\| = \sqrt{(x-y).(x-y)} \quad (8)$$

$$= \left(\sum_{i\_1}^{m}((x_i - y_i)^2)\right)^{1/2} \quad (9)$$

where x and y are in X = Rm.

### 2.4. Decision Tree Classifier (DTC)

DTCs are decision trees used for classification. As any other classifier, the DTCs use values of attributes/features of the data to make a class label (discrete) prediction. Structurally, DTCs are organized like a decision tree in which simple conditions on (usually single) attributes label the edge between an intermediate node and its children [13], [14], [16], [17]. In the study, CART implementation of MATLAB according to Breiman et al. 1984 was used. In this model, Gini Index (GI) was used as splitting measure. GI is an impurity-based criterion that measures the divergences between the probability distributions of the target attributes. The Gini index is defined as:

$$Gini(y,S) = 1 - \sum_{c_j \epsilon dom(y)} \left(\frac{|\sigma_{y=cj}S|}{|S|}\right)^2 \quad (10)$$

And, the evaluation criterion for selecting the attribute $a_i$ is defined as:

$$GiniGain(a_i,S) = Gini(y,S) - \sum_{v_i \in dom(a_i)} \frac{|\sigma_{a_i=v_{i,j}}S|}{|S|}.Gini\left(y, \sigma_{a_i=v_{i,j}}S\right) \quad (11)$$

Error–based pruning is employed in the model. The error rate is estimated using the upper bound of the statistical confidence interval for proportions.

$$\epsilon UB(T,S) = \epsilon(T,S) + Z_\alpha.\sqrt{\frac{\epsilon(T,S).(1-\epsilon(T,S))}{|S|}} \quad (12)$$

Where, $\epsilon(T,S)$ denotes the misclassification rate of the tree T on the training set S. Z is the inverse of the standard normal cumulative distribution and α is the desired significance level. The growing phase continues until a stopping criterion is triggered [19].

### 2.5. Support Vector Machine (SVM)

SVM is a kernel-based learning algorithm in which only a fraction of the training examples is used in the solution (these are called the support vectors), and where the objective of learning is to maximize a margin around the decision surface. The basic idea of applying SVM to classification can be stated briefly as first map the input vectors into one feature space (possibly with a higher dimension), either linearly or nonlinearly, which is relevant with

the selection of the kernel function; then within the feature space, seek an optimized linear division, i.e. construct a hyperplane which separates two classes [18]. Considering classification for two classes with training vectors $x_i \epsilon R^+, i = 1 \ldots n$ and $y \epsilon \{1, -1\}$, SVC solves the following problems [20]:

$$\min_{w,b,\delta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \delta_i \; subject\ to\ y_i\left(w^T \varphi(x_i)\right) + \; b \; \geq 1 - \delta_i, \; \delta_i \geq 0, i = 1, \ldots n \quad (13)$$

The dual of this formula is,

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \; subject\ to\ y^T \alpha = 0, \;\; 0 \leq \alpha_i \leq C, \; i \; = 1, \ldots, n \quad (14)$$

Where $e$ is the vector of all ones, C > 0 is the upper bound, Q is an n by n positive semidefinite matrix, $Q_{ij \equiv y_i y_j K(x_i x_j)}$ where $K\left(x_i, x_j\right) = \varphi(x_i)^T \varphi(x_j)$. Here training vectors are implicitly mapped into a higher dimensional psace by the function $\varphi$. The decision function is:

$$sgn(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + \; p) \quad (15)$$

### 2.6. Multilayer Perceptron (MLP)

MLP is a supervised neural network based on the original simple perceptron model with back propagation for training the network. It commonly consists of an input layer of source nodes, an output layer and one or more hidden layers of computation nodes (neurons) that increasing the learning power of the MLP model. The number of hidden neurons determines the learning capacity of MLP network. It is most recommended to select the network which performs best with the least possible number of hidden neurons [19]. Considering an MLP consisting of a single input, hidden and output layers, n-dimensional feature or input vector is denoted by *X = (x1, …, xn)* and weight vector by *W= (w1, …,wn)*, then the weighted outputs of each neuron in the hidden layer will be:

$$y = \sum_{i=1}^{n} w_i x_i \quad (16)$$

After that, the calculated value is passed through an activation function to yield an output value. Taking the activation function at layer $j$ as $f^{(j)}(x)$, the output can be determined as follows:

$$net = f^{(2)}\left(\sum_j f^{(1)}(\sum_i w_i x_i).w_j\right) \quad (17)$$

Hyperbolic tangent is commonly used as the standard sigmoid activation function in MLPs, of which values range between 0 and 1.

$$f(x) = \tanh(x) = 2sigmoid(2x) - 1, f(-x) = -f(x) \quad (18)$$

The hyperbolic tangent is the solution to the differential equation $f'(-x) = 1 - f(x)^2$ with $f(0) = 0$ and the non-linear boundary value problem: $\frac{1}{2} f'' = f^3 - f; f(0) = f'(\infty) = 0$.

$$y = \frac{1}{1+e^{-net}} \quad (19)$$

Where $e$ is the euler number and $y$ is the output of the MLP. Thereafter, the error for the computation is calculated as follows:

$$err^{(t)} = T^{(t)} - y^{(t)} \quad (20)$$

Where *err* denotes the difference between the real target (*T*) and the obtained output of the MLP ($y_{(out)}$). Then, the system can be optimized by minimizing the following equation:

$$E = \frac{1}{2} \sum_i (err_i)^2 \quad (21)$$

In order to update the weights, the amount of change is calculated for each weight by partial differentiation and the chain rule as follows:

$$\Delta w_{ij} = -\varphi \frac{\partial E}{\partial w_{ij}} = -\varphi \left[\frac{\partial E}{\partial y_j}\right] \left[\frac{\partial y_j}{\partial net_i}\right] \left[\frac{\partial net_i}{\partial w_{ij}}\right] \quad (22)$$

Where $\varphi$ denotes the learning rate. In the final step, each weight is updated as follows:

$$w_{ij}^{new} = w_{ij}^{old} + \; \Delta w_{ij} \quad (23)$$

The procedure given above comprises only one "*epoch"* and the same calculations are repeated until reaching the stopping criteria. MLP is capable of modeling complex functions, good at ignoring irrelevant inputs and noise, it can adapt its weights and it is easy to use. MLPs have been used as the main method in many studies conducted in the field of traffic accidents or to make comparisons.

### 2.7. Logistic Regression (LR)

LR is a predictive analysis and used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. It is a frequently used and well known statistical analysis [13], [21]. Suppose we have a binary output variable Y, and we want to model the conditional probability Pr(Y=1|X=x) as a function of x, any unknown parameters in the functions are to be estimated by maximum likelihood. Let log p(x) be a linear function of x so that changing an input variable multiplies the probability by a fixed amount. The logarithms are unbounded in only one direction, and linear functions are not. Through logistic (or logit) transformation of log p(x), we obtain:

$$log \frac{p(x)}{1-p(x)} = \beta_0 + x.\beta \quad (24)$$

Solving this equation for p gives:

$$p(x; b, w) = \frac{e^{\beta_0 + x.\beta}}{1+e^{\beta_0 + x.\beta}} = \frac{1}{1+e^{-(\beta_0 + x.\beta)}} \quad (25)$$

When $p \geq 0.5, Y = 1$ and when $p < 0.5, Y = 0$. This means guessing 1 whenever $\beta_0 + x.\beta$ is non-negative, and 0 otherwise. So, LR gives a linear classifier. The decision boundary separating the two predicted classes is the solution of $\beta_0 + x.\beta = 0$, which is a point if x is one dimensional, a line if it is two dimensional [22].

## 3. Results and Discussion

### 3.1. Performance Metrics

The performance of a classifier model is defined from a matrix, known as confusion matrix, which shows the correctly and incorrectly classified instances for each class. TP, TN, FP, FN metrics can be described as follows [23]:

- True Positive (TP): instances that are positive and classified as positive
- True Negative (TN): instances that are negative and classified as negative
- False Positive (FP): instances that are negative but classified as positive
- False Negative (FN): instances that are positive but classified as negative instances that are negative and classified as negative

The measures that are used to evaluate the performance of a classifier are computed from the generated confusion matrix. Sensitivity and specificity are the most widely used statistics in a diagnostic test. Sensitivity (True Positive Rate (TPR)) is the proportion of instances correctly labeled positive in all positive instances tested (1); Specificity (True Negative Rate (TNR)) is the proportion of instances correctly labeled negative in all the negative instances tested (2); the Positive Predictive Value (PPV) is defined as (3) where a "true positive" is the event that the test makes a positive prediction; the Negative Predictive Value (NPV) is defined as (4) where a "true negative" is the event that the test makes a negative prediction. Accuracy (ACC) is the likelihood of a correctly predicted total number of modules (5) [24];

$$TPR = TP/(TP + FN) \quad (26)$$
$$TNR = TN/(TN + FP) \quad (27)$$
$$PPV = TP/(TP + FP) \quad (28)$$
$$NPV = TN/(TN + FN) \quad (29)$$
$$ACC\ (\%) = [(TP + TN)/(TP + TN + FP + FN)] * 100 \quad (30)$$

In other words, sensitivity measures the ability of a test to detect the condition when the condition is present. Specificity measures the ability of a test to correctly exclude the condition (not detect the condition) when the condition is absent. Nonfatal predictive value is the proportion of nonfatal cases that correspond to the presence of the condition. Fatal predictive value is the proportion of fatal cases that correspond to the absence of the condition [23].

Receiver Operating Characteristic (ROC) curve is a plot of the TPR against the False Positive Rate (FPR) (6) at various threshold settings which shows the trade-offs between true positive (benefits) and false positive (costs). The area under the ROC curve (AUC) quantifies the overall discriminative ability of a test. An entirely random test has an AUC of 0.5, whereas a perfect test has an AUC of 1.00 [27] [28] [29].

$$FPR = FP/(TN + FP) \quad (31)$$

### 3.2. Performance of the Prediction Models

In the study, NBC, kNN, DTC, SVM, MLP and LR methods were chosen as classifiers in order to provide a deep understanding into the nature of classification with a wide range of machine learning methods. The machine learning methods were all applied using MATLAB software, and LR method was performed with IBM SPSS Statistics software. For each method, twenty predictor variables obtained from two abovementioned datasets were provided as input variables. And the severity of accident (fatal and nonfatal) was set as output. Specificity, Sensitivity, Accuracy, ROC and AUC results were measured to compare the performances of classifiers.

As mentioned in Section 2.1, the original accident dataset has an unbalanced structure; therefore, two different datasets were created (the first one has (25-75%) fatal/nonfatal ratio; the second one has (50-50%) fatal/nonfatal ratio. Then, 10-fold cross-validation was carried out for these two different combinations of inputs. For this purpose, cvpartition method of MATLAB software was used with a for-loop. 10-fold cross-validation is performed by separating 90% training and 10% test data in each fold randomly. Then, all analysis methods are applied in turn. kNN classification is achieved by 'fitcknn' method of MATLAB, where distance is set to Euclidean and number of neighbors is set to 1, as there are two output classes, fatal and nonfatal. NBC is made by 'fitcnb' method of MATLAB, where Gaussian distribution is specified to model data. DTC is performed by 'fitctree' method of MATLAB, which automatically selects the optimal subset of algorithms for each split using the known number of classes and levels of a categorical predictor. The parameters are chosen as follows: 'prune=on','minparentsize=10','AlgorithmForCategorical = PCA','qetoler=1E-6' and 'mergeleaves=on'. SVM classification model was designed by running 'fitcsvm' function in MATLAB. Several combinations have been tried and Radial Basis Function is chosen as the kernel for performance comparison, while Sequential Minimal Optimization is chosen as a solver for gradient difference between upper and lower violators.

MLP model, a supervised neural network model, is designed with 'patternnet' function of MATLAB. It takes three parameters, which are set as 'hiddenSizes=10','trainFcn=trainscg' and 'performFcn=crossentropy'.

The mean results of the first dataset (25/75 fatal/nonfatal) are given in Table 2.

**Table 2.** Prediction results of methods on the initial dataset (25/75 dataset of fatal/nonfatal)

| Performance Metrics | kNN | NBC | DTC | SVM | MLP | LR |
|---|---|---|---|---|---|---|
| TNFR | 0.884 | 0.881 | 0.914 | 0.849 | 0.856 | 0.961 |
| TFR | 0.857 | 0.613 | 0.846 | 0.762 | 0.766 | 0.441 |
| NFPV | 0.963 | 0.862 | 0.955 | 0.946 | 0.946 | 0.838 |
| FPV | 0.622 | 0.651 | 0.727 | 0.492 | 0.521 | 0.791 |
| ACC | 0.878 | 0.809 | 0.898 | 0.832 | 0.839 | 0.831 |
| AUC | 0.792 | 0.756 | 0.841 | 0.824 | 0.719 | 0.798 |

R-square is 0.467 for LR. Based on the results given in Table 2, the following remarks can be made:

- In terms of TNFR values, LR had the highest results (0.961) and it was followed by DTC (0.914), kNN (0.884), NBC (0.881), MLP (0.856), and SVM (0.849).
- In terms of TFR values, kNN yielded the highest result (0.857) and it was followed by DTC (0.845), MLP (0.766), SVM (0.762), NBC (0.613) and LR (0.441).
- In terms of NFPV, kNN yielded the highest result (0.963), and it was followed by DTC (0.955), SVM (0.946), MLP (0.946), NBC (0.862) and LR (0.838). All the models were quite good at predicting nonfatal instances. Actually, this was an expected and normal outcome considering that the percentage of nonfatal instances is quite high.
- In terms of FPV, LR had the highest result (0791) and it was followed by DTC (0.727), NBC (0.651), kNN (0.622), MLP (0.521) and SVM (0.492).
- In terms of ACC, DTC yielded the highest total accuracy (0.898) and it was followed by kNN (0.878), MLP (0.839), SVM (0.832), LR (0.831) and NBC (0.809).
- In terms of AUC values, DTC yielded the highest result (0.841) and it was followed by MLP (0.824), LR (0.798), kNN (0.792), NBC (0.756) and SVM (0.719).
- In conclusion, only the DTC and LR methods provided a fair classification for fatal instances. As known, the highly non-linear relationship between variables will result in failure for models and thus make such models invalid. However, DTC does not require any assumptions of linearity in the data. This could have triggered the success of DTC in the analysis. On the other hand, LR's success mainly derived from its high accuracy of the non-fatal instance. This could be attributed to the fact that LR is great at simple classification problems. The initial data set contains a large amount of non-fatal instances, which could have contributed to the success of LR. As a result, DTC and LR can be seen as good classifiers due to their high AUC values despite their relatively low

FPVs.

For the next analysis, the methods with the same parameter settings were applied to the second dataset consisting of an equal number of nonfatal and fatal (50/50) instances. The obtained results are given in Table 3.

**Table 3.** Prediction results of methods on the second dataset (50/50 dataset of fatal/nonfatal)

| Performance Metrics | kNN | NBC | DTC | SVM | MLP | LR |
|---|---|---|---|---|---|---|
| TNFR | 0.864 | 0.828 | 0.902 | 0.858 | 0.830 | 0.781 |
| TFR | 0.956 | 0.931 | 0.912 | 0.926 | 0.943 | 0.714 |
| NFPV | 0.960 | 0.935 | 0.910 | 0.931 | 0.947 | 0.732 |
| FPV | 0.845 | 0.793 | 0.894 | 0.838 | 0.797 | 0.766 |
| ACC | 0.903 | 0.864 | 0.902 | 0.884 | 0.872 | 0.748 |
| AUC | 0.902 | 0.867 | 0.904 | 0.894 | 0.925 | 0.824 |

R-square is 0.401 for LR. Based on the analysis results given in Table 3, the following remarks can be made:

- In terms of TNFR values, DTC yielded the highest result (0.902), which was followed by kNN (0.864), SVM (0.858), MLP (0.830), NBC (0.828) and LR (0.781).
- In terms of TFR values, kNN had the highest score (0.956), and it was followed by MLP (0.943), NBC (0.931), SVM (0.926), DTC (0.912) and LR (0.714). All methods achieved quite high and similar scores in this parameter except for LR.
- In terms of NFPV, kNN had the highest score (0.960) and it was followed by MLP (0.947), NBC (0.935), SVM (0.931), DTC (0.910) and LR (0.732). Machine learning methods are quite successful in classifying nonfatal instances with over 90% accuracy, while LR could not achieve a significant classification rate.
- In terms of FPV, DTC yielded the highest result (0.894), which was followed by kNN (0.845), SVM (0.838), MLP (0.797), NBC (0.793) and LR (0.766). The most important problem encountered in the study is to reach an acceptably high precision in the classification of fatal instances. From this respect, DTC, kNN, and SVM attained good classification rates between 80-90%, while MLP, NBC, and LR scored only a little below 80%. Especially, DTC came first in both analyses, and kNN is another successful method for this parameter.
- In terms of AUC values, MLP had the highest score (0.925), and it was followed by kNN (0.902), DTC (0.902), SVM (0.884), NBC
- (0.864) and LR (0.748). AUC is an important parameter to illustrate classification accuracy, and in this regard, MLP, DTC, and kNN are distinguished from the other methods as they showed success in the classification of both fatal and nonfatal instances.
- In terms of ACC (overall accuracy), kNN yielded the highest result (0.903) and it was closely followed by DTC (0.902), SVM (0.884), MLP (0.872), NBC (0.864) and LR (0.824). Compared to the results of the previous analyses, all methods improved their results, DTC and kNN continued their superiority over other methods; however, the classification accuracy of LR degraded significantly.
- In conclusion, the results of the second analysis indicated that the accuracy of complex classification methods significantly increased, and all methods except for LR achieved similar rates of overall accuracy, with slightly better results of kNN and DTC. kNN's performance depends on the close neighborhood of similar target and can yield good predictive accuracy in low dimensions. Likewise, DTC proved to be more accurate in low dimensions, but both kNN and DTC have poor run-time performance when the data set becomes large. This is because each new node requires the computation of distance to every other node in the model for kNN, and similarly, the more decisions there are in a tree, the less accurate any expected outcomes are likely to be. On the other hand, MLP and SVM are known to produce even more accurate results with high dimensions and large datasets.

### 3.3. Predictor Importance Analysis

The AUC-based method was carried out with the second dataset consisting of an equal number of fatal and nonfatal accidents using 10-fold cross-validation. The obtained results are given for each classification method in Table 4. As a whole, no single parameter made big difference alone in the classification methods, and they all made similar small contributions to the model outputs either it was negative or positive.

The input parameters are listed in descending order by AUC level. The following inferences can be made from the results in Table 4:

- In NBC model, Mean Cloudiness is slightly more effective over the classification result and it is somewhat differentiated from other variables. It was followed by Mean Pressure and Ground Surface Temperature.
- In kNN, Global Solar Radiation is slightly more effective over the classification and similarly, its removal resulted in the highest loss of AUC value.
- In DTC, Slope, Traffic Control, Ground Surface Temperature and Division of Road are more effective input variables, respectively. On the other hand, removal of Pavement Marking, Total Precipitation, Crash Time Period, Crash Location and Maximum Temperature variables improved the AUC result for good, and there is no such big improvement in other models.

- In SVM, Mean Cloudiness is determined to be more effective and it is followed by Traffic Control, Ground Surface Temperature and Mean Wind Speed, respectively. It is notable that most of the input variables have positive effects on the AUC value, while only a few have rather small negative effects.
- In MLP, Division of Road is more effective, which is followed by Pavement Marking and Mean Cloudiness with close rates.
- Considering all models, Mean Cloudiness, Traffic Control, and Ground Surface Temperature variables are observed to be slightly more effective on the results, while Maximum Temperature and Weather had a negative effect on AUC in all models. Weather is a parameter recorded by the traffic officer at the accident location, and it only gives the superficial description of the weather. Therefore, it is quite normal for this parameter to have an inconsistent effect on the result.

## 4. Conclusions

In this paper, NBC, kNN, DTC, SVM and MLP methods and LR statistic method are used to analyze motor vehicle accident data according to the accident result (i.e. fatal and non-fatal), also the significant factors that are associated with detailed meteorological reports in traffic accidents are identified. Property damage–only accidents are not included in this study

**Table 4.** Predictor importance analysis results for NBC, kNN, DTC, SVM and MLP models

| *Input Sym.* | *NBC* Change in AUC | % | *Input Sym.* | *kNN* Change in AUC | % | *Input Sym.* | *DTC* Change in AUC | % | *Input Sym.* | *SVM* Change in AUC | % | *Input Sym.* | *MLP* Change in AUC | % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MCL | 0.036 | 4.136 | GSR | 0.02 | 2.226 | SLO | 0.02 | 2.189 | MCL | 0.034 | 3.807 | DIV | 0.026 | 2.798 |
| MXMP | 0.013 | 1.465 | CRO | 0.008 | 0.896 | TRA | 0.018 | 1.986 | TRA | 0.019 | 2.111 | MAR | 0.016 | 1.732 |
| GST | 0.011 | 1.292 | PER | 0.006 | 0.683 | GST | 0.014 | 1.543 | GST | 0.012 | 1.376 | MCL | 0.016 | 1.716 |
| MNTP | 0.009 | 1.022 | GST | 0.006 | 0.674 | DIV | 0.014 | 1.524 | MWS | 0.01 | 1.121 | SLO | 0.011 | 1.22 |
| MWS | 0.008 | 0.964 | MNTP | 0.004 | 0.48 | CRO | 0.01 | 1.109 | MXTP | 0.008 | 0.933 | TRA | 0.01 | 1.069 |
| MRH | 0.006 | 0.742 | LOC | 0.002 | 0.212 | MCL | 0.006 | 0.711 | SUR | 0.008 | 0.924 | GSR | 0.01 | 1.04 |
| GSR | 0.004 | 0.463 | DAY | 0 | 0 | SHO | 0.006 | 0.619 | GSR | 0.008 | 0.858 | CRO | 0.009 | 0.937 |
| DAY | 0.003 | 0.309 | MAR | 0 | -0.009 | MRH | 0.004 | 0.388 | SHO | 0.006 | 0.707 | MWS | 0.008 | 0.903 |
| MAR | 0.003 | 0.299 | MWS | 0 | -0.037 | GSR | 0.003 | 0.388 | MXMP | 0.006 | 0.679 | MXMP | 0.004 | 0.43 |
| TRA | 0.002 | 0.289 | MRH | 0 | -0.037 | MXMP | -0.002 | -0.203 | CRO | 0.004 | 0.499 | DAY | 0.004 | 0.391 |
| MXTP | 0.002 | 0.27 | TRA | -0.002 | -0.222 | DAY | -0.002 | -0.231 | MNTP | 0.004 | 0.462 | MRH | 0.003 | 0.369 |
| DIV | 0.002 | 0.241 | SLO | -0.002 | -0.268 | MWS | -0.002 | -0.268 | PER | 0.002 | 0.254 | PER | 0.002 | 0.202 |
| LOC | 0.002 | 0.231 | SUR | -0.002 | -0.277 | WET | -0.005 | -0.508 | DAY | 0.001 | 0.066 | SHO | 0.001 | 0.073 |
| SLO | 0 | -0.019 | MXMP | -0.003 | -0.388 | MNTP | -0.006 | -0.665 | TP | 0 | 0.019 | TP | -0.003 | -0.374 |
| TP | -0.001 | -0.135 | TP | -0.004 | -0.425 | SUR | -0.009 | -0.97 | LOC | -0.002 | -0.207 | GST | -0.004 | -0.416 |
| PER | -0.003 | -0.337 | MXTP | -0.004 | -0.434 | MXTP | -0.01 | -1.145 | DIV | -0.002 | -0.283 | MNTP | -0.005 | -0.586 |
| CRO | -0.003 | -0.386 | WET | -0.004 | -0.452 | LOC | -0.011 | -1.164 | WET | -0.004 | -0.452 | SUR | -0.006 | -0.634 |
| WET | -0.005 | -0.607 | SHO | -0.004 | -0.48 | PER | -0.013 | -1.386 | MRH | -0.004 | -0.471 | LOC | -0.006 | -0.657 |
| SUR | -0.005 | -0.627 | MCL | -0.008 | -0.859 | TP | -0.018 | -2.042 | MAR | -0.004 | -0.49 | MXTP | -0.01 | -1.114 |
| SHO | -0.006 | -0.685 | DIV | -0.008 | -0.877 | MAR | -0.023 | -2.55 | SLO | -0.004 | -0.509 | WET | -0.011 | -1.211 |

MCL: Mean Cloudiness; MXMP: Mean Pressure; GST: Ground Surface Temperature; MNTP: Minimum Temperature; MWS: Mean Wind Speed; MRH: Mean Relative Humidity; GSR: Global Solar Radiation; DAY: Day of Week; MAR: Pavement Marking; TRA: Traffic Control; MXTP: Maximum Temperature; DIV: Division of Road; LOC: Location; SLO: Slope; TP: Total Precipitation; PER: Crash Time Period; CRO: Crossing; WET: Weather; SUR: Roadway Surface; SHO: Shoulder.

A total of 20 parameters were used as input to classify accidents into two classes, fatal or non-fatal. The most difficult part of the study is to classify fatal instances accurately due to their low percentage value in total. The first dataset consisted of 246 fatal and 738 non-fatal cases, while the second included 246 fatal and 246 non-fatal cases.

As a result, DTC and kNN algorithms yielded slightly more accurate results in classifying fatal instances in both datasets. On the other hand, MLP yielded the highest accuracy in both nonfatal and fatal instances combined as well as the highest AUC rate. Although LR performed well in the first dataset, its accuracy significantly decreased with the second dataset. The success of kNN and DTC could be attributed to the low dimensionality of the datasets.

To analyze predictor importance of the prediction models, AUC-based input ranking method is used. Based on this method, Mean Cloudiness, Traffic Control and Ground Surface Temperature variables were found to have a higher weight on classification results; in addition, Maximum Temperature and weather parameters negatively affected the classification performance of all models.

The dataset lacks information on driver and vehicle characteristics, which was the main disadvantage of the study. The current traffic accident report should include information about driver characteristics like age, gender, education, etc. as well as vehicle characteristics like model, age, and type. With this additional information, the more detailed analysis could be carried out in the future.

## References


[1] WHO, "Global status report on road safety," 2016. [Online]. Available: http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/. [Accessed: 19-Jun-2017].

[2] WHO, "Road traffic injuries," 2017. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs358/en/. [Accessed: 19-Jun-2017].

[3] TSI, "Turkish Statistical Institute," Turkish Statistical Institute, 2017. [Online]. Available: http://www.turkstat.gov.tr/Start.do. [Accessed: 19-Jun-2017].

[4] M. A. Abdel-Aty and A. E. Radwan, "Modeling traffic accident occurrence and involvement" Accid. Anal. Prev., vol. 32, no. 5, pp. 633–42, Sep. 2000.

[5] S. Y. Sohn and H. Shin, "Pattern recognition for road traffic accident severity in Korea," Ergonomics, vol. 44, no. 1, pp. 107–117, Jan. 2001.

[6] Q. Wu, G. Zhang, X. Zhu, X. C. Liu, and R. Tarefder, "Analysis of driver injury severity in single-vehicle crashes on rural and urban roadways," Accid. Anal. Prev., vol. 94, pp. 35–45, 2016.

[7] M. Taamneh, S. Alkheder, and S. Taamneh, "Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates," J. Transp. Saf. Secur., pp. 1–21, Apr. 2016.

[8] E. I. Vlahogianni, M. G. Karlaftis, and F. P. Orfanou, "Modeling the Effects of Weather and Traffic on the Risk of Secondary Incidents," J. Intell. Transp. Syst., vol. 16, no. 3, pp. 109–117, Jul. 2012.

[9] J. Ona, R. O. Mujalli, and F. J. Calvo, "Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks," Accid. Anal. Prev., vol. 43, no. 1, pp. 402–411, Jan. 2011.

[10] M. Chong, A. Abraham, M. Paprzycki, "Traffic Accident Analysis Using Machine Learning" Informatica 29 (2005) 89–98

[11] S. Krishnaveni, M. Hemalatha, "A Perspective Analysis of Traffic Accident using Data Mining Techniques", International Journal of Computer Applications (0975 – 8887) Volume 23– No.7, June 2011

[12] S. Vasavi, "Extracting Hidden Patterns Within Road Accident Data Using Machine Learning Techniques" Information and Communication Technology, Advances in Intelligent Systems and Computing 625, https://doi.org/10.1007/978-981-10-5508-9_2

[13] C. M. Bishop, Pattern recognition and machine learning. Springer,

2006.
[14] T. Mitchell, Machine Learning. McGraw-Hill, 1997.
[15] S. Ajmani, K. Jadhav, and S. A. Kulkarni, "Three-Dimensional QSAR Using the k-Nearest Neighbor Method and Its Interpretation," J. Chem. Inf. Model., vol. 46, no. 1, pp. 24–31, Jan. 2006.
[16] O. Z. Maimon and L. Rokach, Soft computing for knowledge discovery and data mining. Springer, 2011.
[17] S. Shalev-Schwartz and S. Ben-David, Understanding machine learning : from theory to algorithms. Cambridge: Cambridge University Press, 2014.
[18] M. Chong, A. Abraham, and M. Paprzycki, "Traffic Accident Analysis Using Machine Learning Paradigms," Informatica, vol. 29, no. 1, pp. 89–98, 2005.
[19] Rokach, L. "Classification and Regression Tree Lecture Notes-Chapter 9", http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf (access date: November 19, 2017).
[20] Scikit Learn Documentation, http://scikitlearn.org/stable/modules/svm.html#svc (access date: February 2, 2018).
[21] X. Fan, L. Wang, and S. Li, "Predicting chaotic coal prices using a multi-layer perceptron network model," Resour. Policy, vol. 50, pp. 86–92, Dec. 2016.
[22] Shalizi, 2012. Advanced Data Analysis from an Elementary Point of View. Cambridge University Press.
[23] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. New York, NY: Springer New York, 2009.
[24] M. Taamneh, S. Taamneh, and S. Alkheder, "Clustering-based classification of road traffic accidents using hierarchical clustering and artificial neural networks," Int. J. Inj. Contr. Saf. Promot., pp. 1–8, Sep. 2016.
[25] D. G. Altman and J. M. Bland, "Statistics Notes: Diagnostic tests 2: predictive values," BMJ, vol. 309, no. 6947, 1994.
[26] BIML, "Test Statistics," 2017. [Online]. Available: http://groups.bme.gatech.edu/groups/biml/resources/useful_documents/Test_Statistics.pdf. [Accessed: 19-Jun-2017].
[27] Garson and G. David, "Interpreting neural-network connection weights," AI Expert, vol. 6, no. 4, pp. 46–51, 1991.
[28] Y.-W. Chang and C.-J. Lin, "Feature Ranking Using Linear SVM," in JMLR: Workshop and Conference Proceedings 3, 2008, pp. 53–64.
[29] K. Subbian and P. Melville, "Supervised Rank Aggregation for Predicting Influencers in Twitter," in 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing, 2011, pp. 661–665.