IJTE, Vol. 10, No. 3, Winter 2022, pp. 1089-1102
Research Paper

New Optimization Approach for Handling Imbalanced Data in Road Crash Severity

Abbas Rouhi Mashhadsari[1], Gholamali Behzadi[2,*]

Received: 2021/10/19 Accepted: 2021/12/13

Abstract

Accidents are a major problem that claim the lives of many people in the world each year. Fatalities and severe injuries could leave adverse and irreversible impacts on public health and economic prospects. A review of the variables affecting the severity of crash injuries can help reduce fatal accidents. However, a detailed prediction of fatal crashes as a smaller-data class than other classes is seen as a challenge. This study uses three robust machine learning such as Bayesian classifier, random forest, and support vector machine techniques. First, three imbalanced data prediction models were developed, suggesting they could not differentiate fatal data from injury data. To address this problem, three random, k-means clustering, meta-heuristic algorithms clustering techniques were used to balance the data. It should be noted that the genetic algorithm performed better than the particles swarm. Models developed by intelligent optimization methods, k-means clustering, and random methods were found to be more accurate, respectively. These criteria helped evaluate the models developed, which yielded the best model. The support vector machine method for genetic clustering-balanced data could predict fatal, and injury crashes with a 0.96% accuracy, becoming the best model. Finally, sensitivity analysis was performed on the best model, indicating that the highway, horizontal curves, and head-on variables contributed to fatal accidents.

Keywords: Genetic algorithm, particles swarm, optimization, crashes, machine learning

[*] Corresponding author. E-mail: Ga.behzadi@yahoo.com
[1]Ph.D. Candidate, School of Civil Engineering, Shomal University, Mazandaran, Amol, Iran
[2]Assistant Professor, School of Civil Engineering, Shomal University, Mazandaran, Amol, Iran

1. Introduction

Like other industries, the transportation industry has various advantages and disadvantages. Over time, vehicles become more popular and this, in turn, increases the number of vehicles on the roads. To meet this end, the safety of users is one of the most important issues which assumes special consideration. By safety, it is meant minimizing the number and severity of accidents by examining the major factors affecting them. One of the most important solutions to the safety problem is to predict the severity of accidents, which has received growing attention over the past years. Prediction of accidents will positively help lower direct costs, reduce fatalities and financial losses while lowering indirect costs from energy waste and loss of manpower [Le Yu et al., 2021]. Accident figures released by the World Health Organization (WHO) suggest that approximately 1.35 million people die each year as a result of road traffic crashes, and about fifty million are injured across the world. Iran is no exception and sees its rate of fatalities and damages from accidents rising as it ranks fifth for high-risk road driving. Therefore, in recent years, some research has been done on accident prediction [WHO, 2018, Mirbaha et al. 2013, Karami et al. 2020, Washington et al. 2014]. Statistical methods-based traditional techniques such as logistic regression, quantitative regression, and discriminant analysis have largely been used to examine and analyze accident prediction. For prediction, statistical models often require much historical data; however, the data provide unintended results when they have a lot of input variables [Niveditha et al. 2015, Washington et al. 2014, Ba et al. 2017, Tabachnick et al. 2007, Wang et al. 2019].Machine learning methods have widely been used in recent years for the following reasons:

1) Higher capacity to solving nonlinear problems;
2) Higher capability of prediction dependent variables in new data and
3) Higher capability of prediction by extracting rules from data [Basso et al. 2018, Theofilatos et al. 2019, Li et al. 2018, Li et al. 2016, Ahmed et al. 2011].

Machine learning has also been used to predict accidents, which has proved to have performed well. For example, Abdel Wahab and Abdel-Aty (2001) [Abdelwahab et al. 2001] used an artificial neural network to predict the severity of crashes. They examined such various factors as the driver, vehicle, road, and environmental features for modeling. They concluded that the neural network performed better than other instruments. Li et al. [Li et al. 2008] used support vector machine techniques to predict accidents. They also used the dual Bayesian models to evaluate the performance of the support vector machine, which was found to perform better and greatly predict the severity of accidents.

Available data play a key role in predicting accidents. According to road accident data, injury samples usually outnumber fatal ones with lower fatal data as a result. Since machine learning methods try to reduce the overall model rates, they attach more importance to more data and prioritize them, leading to an over-prediction of a class. In the meantime, there are two approaches to reducing the number of class data, one with a great number of data, called under-sampling, and the other with an increased class data, whose data is fewer than other classes, called over-sampling [Abou Elassad et al. 2020, Elamrani Abou Elassad et al. 2020]. Under-sampling, if not made logically and correctly, could remove useful information by removing patterns, while an improper over-sampling may increase the likelihood of the model over-fitting [Abd Elrahman et al. 2013]. Fiorentini and Losa in 2020 used random under sampling the majority class (RUMC) method for unbalanced data and

they used random tree, k-nearest neighbor, logistic regression, and random forest machine learning methods to predict the Road Crash Severity. They achieved acceptable results. This method considers special areas for the minority class as the machine learning method attaches more importance to the class. The synthetic minority over-sampling technique can be used for binary problems and the continuous variables space [Nguyen et al. 2009]. Yen and Lee [Yen et al. 2009] used a clustering-based under-sampling method. They clustered the data and then selected a specific number of data from a majority class in each cluster using a predefined coefficient. Their findings showed that clustering-based sampling increases the accuracy of prediction, which proves to be more stable than other methods. A number of under-sampling methods have been mentioned in the literature review, but metaheuristic optimization methods have received less attention. Therefore, in this paper, an attempt has been made to develop an under-sampling method based on meta-heuristic optimization algorithms. This article uses three machine learning methods of naive Bayes classifier, random forest, and support vector machine to predict the severity of crashes. At first, without making changes to the data, three general prediction models were developed using naive Bayes classifiers, random forest, and support vector machine methods. Then the accuracy and error of the developed models were examined. Later, a random method was used to balance the data to develop three more separate models. Then, the conventional clustering method was used to balance the data, and three other models were developed to predict the severity of accidents. Finally, intelligent optimization methods were used to develop two new genetic clustering and particle swarm methods, which served as the basis for sampling and balancing the data. Along with the selected samples, two more separate models were developed using machine learning methods whose accuracy was compared with that of the previous models; thus, the performance of the developed sampling method was measured. Finally, the best model with greater accuracy and fewer error rates was proposed, and sensitivity analysis was performed on input variables. Also, the factors affecting the fatal crashes were examined.

2. Research Methodology

This section first concerns the data used, then explains the machine learning methods and how the model is developed. Finally, the types of under-sampling methods and the way the accuracy and error of the models are measured will also be discussed. All methods and algorithms used in this research were implemented in MATLAB software.

2.1. Data

The data used in this study have been taken from accidents in intercity areas across the Iran country from 2018 to 2020 (within three years). Different input variables including road type, road characteristics, type of collision, number of vehicles involved, type of vehicle, speed limits, night or day, season, holidays or non-holidays, and residential area were considered. This study included 16122 data as well as input variables to predict the type of accident (injury or fatal). Descriptive statistics of the variables used in this study are given in Table (1).

Machine learning methods try to minimize the difference between the output variable and the prediction values. To address this optimization problem, if the variables are not at the same level, the variable with larger numbers becomes more important because the objective function of machine learning methods usually uses the Euclidean distance [Safak et al. 2020]. For this, for each secondary variable in Table (1), a dummy variable was defined. In general, 34 input variables and one output variable were taken.

Table 1. Descriptive statistics of the data used

Main variables	Secondary variables	Frequency	Frequency percentage	Mode
Type of road	Freeway	6010	37.3	Main
	Highway	969	6	
	Main road	7146	44.3	
	rural road	339	1.2	
	Secondary	16.58	10.3	
Road characteristics	Arc	6112	37.9	Arc
	Intersection	4693	29.1	
	Arc -intersection	2822	17.5	
	Straight	2495	15.5	
Type of collision	Head-on	2351	14.6	7817
	Front to back	2499	15.5	
	Rollover	7817	48.5	
	Lateral barrier collision	965	6	
	Side collision	2490	15.4	
Number of vehicles	One	8782	54.5	One
	Two or more	7340	45.5	
Type of vehicles	Large	926	5.7	Small
	Small	13597	84.3	
	Small and large	1599	9.9	
Speed limits	Less than 60	1567	9.7	60-90
	60-90	8026	49.8	
	Over 90	6529	40.5	
Time	Days	9953	61.7	Day
	Nights	6169	38.3	
Season	Fall	3824	23.7	Summer
	Spring	39.86	24.7	
	Summer	5041	31.3	
	Winter	3271	20.3	
Holidays	No	15030	93.2	No
	Yes	1092	6.8	
Weekdays	No	11094	68.8	No
	Yes	5028	31.2	
Residential area	No	14438	89.5	No
	Yes	1684	10.5	
Type of crash	Fatal	2484	15.4	Injury
	Injury	13638	84.6	

2.2. Support Vector Machine

One of the highly powerful tools in machine learning for classification problems is the support vector machine, which was developed by Vapnik et al. in the 19990s [Vapnik et al. 1995]. The support vector machine tries to use the optimal margin method to determine the boundary between the two classes. By the margin, it is meant the sum of the distances of the nearest point from both classes to the hyper plane. This boundary must be selected to meet the following two conditions:

1) All samples whose output variable is +1 are on one side of the boundary and the ones labeled -1 on the other side.

2) The decision boundary should maximize the margin.

A decision boundary can be written mathematically, as shown in Equation (1).

$$W.X+b=0 \quad (1)$$

Where X is a point on the decision boundary and W is a n-dimensional vector perpendicular to the decision boundary. In fact, W and b should minimize the error and maximize the margin [Cortes et al. 1995].

In this study, the vector machine method was used to predict the severity of accidents. Data were randomly divided into two training (80%) and testing (20%) classes, yielding four models of support vector machines.

2.3. Naive Bayes Classifier

One of the machine learning tools for classification problems is the Naive Bayes Classifier [Keogh, 2006]. This method operates on Bayes' theorem and provides a way to calculate the probability of a secondary probability based on its prior probability. Bayes' theorem is shown in Equation (2).

$$P\left(A|B\right)=\frac{P\left(B|A\right)P\left(A\right)}{P\left(B\right)} \quad (2)$$

According to the Bayes law, the probability of a occurring can be considered given the B event. Here, B is the evidence, and A is the theorem.

To predict the severity of crashes, input variables such as road type, road characteristics, type of collision, etc., should be used. As stated, these input variables are assumed to be completely independent of each other and equally contribute to predicting the severity of accidents. According to the above, the Bayes law can be rewritten as Equation (3).

$$P\left(y|X\right)=\frac{P\left(X|y\right)P\left(y\right)}{P\left(X\right)} \quad (3)$$

Where y is different classes of accidents (injury and fatal), the variable X represents the input variables or parameters, defined as Equation (4).

$$X=\left(x_1,x_2,x_3,...,x_n\right) \quad (4)$$

Where x_i is the input variables, which include the type of road, road characteristics, type of collision, etc. By inserting the xs and extending Equation (3), Equation (5) is obtained.

$$P\left(y|x_1,...,x_n\right)=\frac{P\left(x_1|y\right)P\left(x_2|y\right)...P\left(x_4|y\right)P\left(y\right)}{P\left(x_1\right)P\left(x_2\right)...P\left(x_4\right)} \quad (5)$$

Now the values on the left can be determined by inserting the values on the right side of Equation (5) from the data. Since the denominator of all cases is constant, it can be removed and reported as follows.

$$P\left(y|x_1,...,x_n\right)\propto P\left(y\right)\prod_{i=1}^{4}P\left(x_i|y\right) \quad (6)$$

As in the support vector machine method, the model was developed with 80% of the data (training data), while it was evaluated with the remaining 20%. As noted below, four separate models were developed using this method, and the severity of accidents was predicted.

2.4. Random Forest

The Random Forest (RF) is a popular algorithm developed by Breiman to solve classification and regression problems. This algorithm consists of lots of decision trees. The decision tree is a non-parametric supervised learning method that can make decisions on the data set and use the tree structure to solve classification and regression problems. RF is a combination of prediction trees so that each

tree is sampled independently based on a random vector with the same distribution. This algorithm uses the existing decision trees to predict the output value based on the average output of the used decision trees. In the RF model, all variables are not equally correlated with the output variable.

2.5. Developing Prediction Models using Initial Data

This section concerns the development of two accident severity models using a support vector machine and a naive Bayes classifier with initial data. Initial data includes 16122 data lines of different specifications referred to in Section 2.1. The first and second prediction models were developed using the support vector machine and naive Bayes classifier methods, respectively. To develop models and prevent model over fitting, data must first be divided into two categories: training (80%) and testing (20%) data. Each machine learning model has a number of parameters to be based on the data with the most optimal value so that the model can have the best performance. For this, the K-fold cross-validation was used. Then, the training data were divided into four classes, three of which trained the model, while one remaining class evaluated the model performance. This process is repeated four times in each iteration to evaluate the performance of the model with all the four classes, the average of which helped calculate the model error. By considering different values for the model parameters using the trial-and-error method, the optimal value of these parameters can be determined. After the optimal parameters were determined, the first and second models were developed by the support vector machine and naive Bayes classifier along with initial data, respectively.

2.6. Development of Prediction Models by Balancing the Data using a Random Method

As suggested in the literature review, one way to improve the accuracy of accident prediction models was to balance the data. The total data from this study amounted to 16122 data, of which 2484 data pertained to fatal accidents and 13638 to injury accidents. To balance the data, injury accident data were separated, and 2484 data were selected quite randomly. The selected data were integrated with the fatal data, and a new dataset was created. Using the method described in Section 2-4, the third and fourth models for crash prediction were developed using a support vector machine and naive Bayes classifier along with developed data.

2.7. Development of Prediction Models by Balancing the Data using the Clustering Method

2.7.1. K-Means Clustering Method

Data mining is aimed to analyze data to arrive at meaningful patterns and rules. The issues data mining delves into are divided into two categories. The first category includes data that have an output variable, i.e., the goal is to predict the output variable using the input variable. In the second category, however, there is no output variable, and the goal is to classify the data based on their similarity. The k-means method also tries to divide the data into various clusters, with the data of each class being highly similar to each other and the data of the other classes being different from each other [Likas et al. 2003, Sinaga et al. 2020].

2.7.2. Metaheuristic Particle Swarm Algorithm

The particle swarm optimization method was proposed by James Kennedy and Russell in 1995 [Wang et al. 2018]. They initially aimed to use existing social models and relations to create a type of computational intelligence. Their research led to the creation of a robust algorithm. This method has been adopted from the collective action of swarms of animals such as birds and fish. In this algorithm, there are a number of creatures which are called particles scattered in the search space. The objective

function of each particle in a location in the space where it is located is calculated. Then, the information of the current location and of the best location the particle was previously in, as well as the information from one or more particles of the best particles, are combined to select a direction for movement. After the collective movement, an iteration of the algorithm ends, and these steps are repeated to gain the optimal answer [Yang, 2007]. Figure (1) shows the flowchart of the study steps.

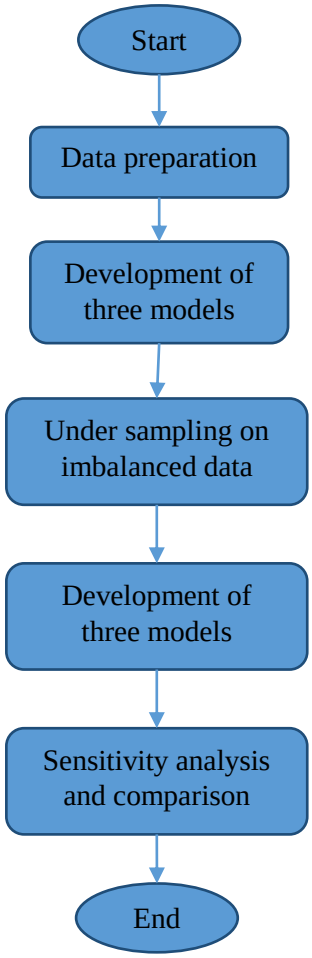


Figure 1. Flowchart of study steps

2.7.3. Metaheuristic Genetic Algorithm (GA)

The genetic algorithm is one of the most conventional and oldest metaheuristic methods proposed by John Holland in 1976 [Holland, 1992]. This algorithm is one of the evolutionary algorithms inspired by nature. It is thus an algorithm that frequently changes the generated population to find the optimal answer, with the changes known as evolution. In each iteration of this evolution, two members of the population are selected as parents who produce children for the next generation. This algorithm can be used to solve high-complex optimization problems.

The genetic algorithm steps are as follows

Step 1: Generating an initial population

Step 2: Selecting the parents; performing crossover, and producing a population of children

Step 3: Selecting the parents; performing the mutation, and generating the mutant population

Step 4: Selecting members of the new main population from among the main population, children, and mutants

Step 5: If the end conditions are not met, step 2, otherwise step 6 is considered.

Step 6: end

In step 4, after crossover and mutation operations are performed in each iteration, there will be three new classes of the population, which are the initial population, the crossover-generated population, and the mutation-generated population. To begin the next iteration, the algorithm must select from the total population as much as the size of the initial population. For this, the three populations generated are evaluated according to the objective function, with the members having a better answer selected according to the number of the initial population and the rest of the total population removed [Mirjalili, 2019, Whitley, 1994].

Where x_i is a member vector d and Mj a vector of the j^{th}center of the cluster (centroid) coordinates which is member d, and k is the number of clusters intended. Here, the objective function and decision variables are defined according to Equations (7) and (8).

$$f = \text{Minimize} \sum_{i=1}^{n} \min d(x_i, m_j), \quad 1 \le j \le k \tag{7}$$

$$DV = k \times d \tag{8}$$

Where f is the objective function, d (x_j, m_j) is the distance between the i^{th} data to the j^{th} cluster, and DV is the number of problem decision variables.

2.7.4. Clustering-Based under-Sampling Method

Yen and Lee [Yen et al. 2009] provided a clustering-based under-sampling method in 2009. They concluded that clustering could balance the data set. To this end, if the total number of samples in the total data (imbalanced data) equals N, which includes the samples of the majority class (MA) and those of the minority class (MI), the size of MA is represented by $Size_{MA}$ and the size of MI by $Size_{MI}$. It is clear that in the imbalanced dataset, $Size_{MA}$is larger than $Size_{MI}$. In this method, the data are first divided into 4 clusters based on the clustering methods mentioned. Each cluster will involve a number of majority class and a number of the minority class. For example, the number of majority class samples and of minority class samples in the i^{th} cluster (i is a number between 1 and k) are represented by $Size^i_{MA}$ and $Size^i_{MI}$, respectively. M is also the balance ratio between the minority class and the majority class, which has the value of 1 in this study. The number of samples selected from the majority class in the i^{th} cluster can be calculated from the following Equation.

$$SSize^i_{MA} = \left(M \times Size_{MI}\right) \times \frac{Size^i_{MA} / Size^i_{MI}}{\sum_{i=1}^{k} Size^i_{MA} / Size^i_{MI}} \quad (9)$$

Using Equation (9), the number of data selected from the majority class in each cluster can be measured. In this study, Equation (9) was used to select a specific number of injury classes from each of the 4 clusters, and the data were then balanced.

2.8. Evaluating the Accuracy of the Developed Models

In problems where the output variable is classified, the accuracy of the models is evaluated by various indicators. The following should be explained:

True Positive (TP): The outcome of a fatal accident in an accident that is truly fatal in nature.

False Positive (FP): The outcome of a fatal (unnatural) accident in an accident that is truly injury in nature.

False Negative (FN): The outcome of an injury accident in an accident that is truly fatal in nature.

True Negative (TN): The outcome of an injury accident in an accident that is truly injury in nature.

The accuracy and error of the models developed in this study can be measured from equations (10) to (14).

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (10)$$

$$Sensitivity = \frac{tp}{tp + fn} \quad (11)$$

$$Specificity = \frac{tn}{tn + fp} \quad (12)$$

$$FPR = \frac{fp}{tn + fp} \quad (13)$$

$$\Pr ecision: \frac{tp}{tp + fp} \quad (14)$$

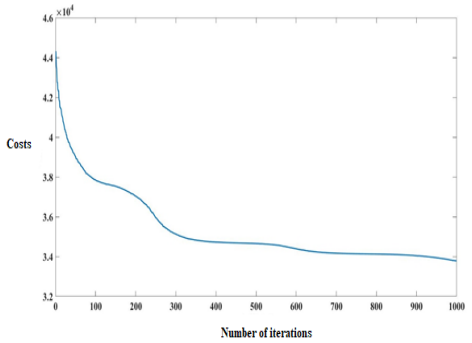
3. Results

This section first provides the results from meta-heuristic genetic optimization and particle swarm algorithms for the development of clustering methods. Then, the results from the prediction models using different machine learning and data balancing methods are presented. After the best prediction model is introduced, sensitivity analysis will be performed on it, and the sensitivity of input variables on the severity of accidents be examined.

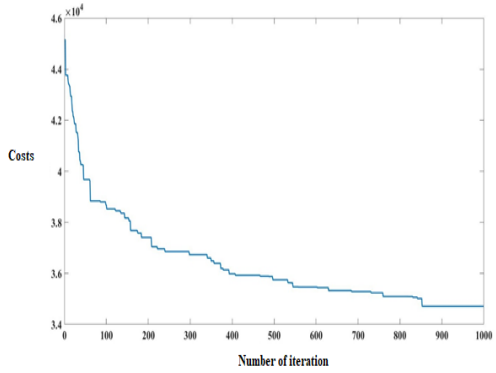
3.1. Results from Genetic Optimization and Particle Swarm Methods for Clustering

The two clustering methods having been developed by two intelligent genetic optimization and particle swarm methods; the method that makes the clustering better is

selected. As previously shown in Equation (7), the goal is to minimize the distance between the members of the cluster. If the cost function of each method (genetic or particle swarm methods) is lower, it suggests its better performance in clustering. Figure (2) demonstrates the cost from each method as plotted by the number of iterations of the algorithm.



A) Genetic method



B) Particle swarm method

Figure 2. Results of meta-heuristic optimization methods to develop clustering method

Figure (2) shows that the objective function (cost) decreases with increasing the number of repetitions, indicating it is nearing the optimal answer. After 1000 iterations, the cost function of the genetic algorithm is seen to have decreased than that of the particle swarm algorithm, showing the better performance of the former algorithm in clustering. Thus, this study used the genetic algorithm for clustering, while under-sampling and clustering methods were based on a) Genetics b) Particle swarm methods to balance the data.

3.2. Results from Accident Prediction Models using Machine Learning

In general, in this study, eight severity of accident prediction models were developed using three support vector machines and naive Bayes classification methods as well as four classes of balanced and imbalanced data sets, with the results in terms of training and testing data are given in Tables (2) and (3). Initially, the first, second, and third prediction models of the severity of accidents were developed based on initial (imbalanced) data, with all data predicted as relating to injury accidents in both models of training and testing data. Since fatal accidents accounted for 17% of the data, the model decided to predict all the data pertaining to injury accidents to reduce the error rate, while this study aimed to accurately predict fatal and injury accident data. In other words, the accuracy of the first, ninth and second models is generally good but has a very low validity due to failure to predict fatal data. To address the problem and to have the model accurately predict fatal crashes, various dimension reduction methods were proposed. The third, tenth, and fourth models were developed based on the random method, and the data were balanced. Although other parameters had lower accuracy compared to the first, ninth and second models, the models managed to differentiate injury data from fatal data, though with low accuracy. The fifth, eleventh and sixth models had better accuracy, with the genetic algorithms demonstrating the best performance for clustering and development of the seventh, twelfth and eighth models. In general, support vector machine models also performed better than the naive Bayes and random forest classifier and enjoyed better accuracy. Thus, the seventh model was found to be the superior model of this study.

Table 2. Evaluation of the accuracy of training data-developed models

Method	Data	Model Name	Specificity	sensitivity	Accuracy
SVM	Imbalanced	First	1	0	0.83
	Balanced (random)	Third	0.67	0.66	0.66
	Balanced (k-means)	Fifth	1	0.82	0.90
	Balanced (genetic)	Seventh	1	0.93	0.96
RF	Imbalanced	Ninth	1	0.02	0.83
	Balanced (random)	Tenth	0.65	0.65	0.65
	Balanced (k-means)	Eleventh	0.97	0.81	0.89
	Balanced (genetic)	Twelfth	1	0.92	0.95
Naïve Bayes	Imbalanced	Second	1	0	0.83
	Balanced (random)	Fourth	0.67	0.58	0.58
	Balanced (k-means)	Sixth	0.99	0.75	0.75
	Balanced (genetic)	Eighth	1	0.92	0.91

Table 3. Evaluation of the accuracy of testing data-developed models

Method	Data	Model Name	Specificity	sensitivity	Accuracy
SVM	Imbalanced	First	1	0	0.83
	Balanced (random)	Third	0.74	0.71	0.71
	Balanced (k-means)	Fifth	0.99	0.86	0.92
	Balanced (genetic)	Seventh	1	0.97	0.98
RF	Imbalanced	Ninth	1	0.01	0.82
	Balanced (random)	Tenth	0.61	0.61	0.61
	Balanced (k-means)	Eleventh	0.93	0.75	0.83
	Balanced (genetic)	Twelfth	1	0.86	0.92
Naïve Bayes	Imbalanced	Second	1	0	0.83
	Balanced (random)	Fourth	0.68	0.59	0.63
	Balanced (k-means)	Sixth	1	0.75	0.86
	Balanced (genetic)	Eighth	1	0.91	0.95

After the eight prediction models on the severity of accidents were evaluated, it was concluded that the seventh model, i.e., the support vector machine-based prediction model, used the data balanced by the meta-heuristic genetic optimization algorithm to yield the best performance.

4. Sensitivity Analysis

Because machine learning techniques serve as a black box and do not provide a relationship between input and output variables, sensitivity analysis on input variables can make up for this weakness. For this purpose, the best model for predicting the severity of accidents, which uses supported vector machine method data balanced by the meta-heuristic genetic algorithm, was selected to examine the effect of input variables on the severity of accidents. For example, if the goal is to examine the type of roads on the severity of accidents, the rest of the available data variables are assumed to be constant, with the variable of the type of road examined at different speeds. Sensitivity analysis results are given in Tables (4). As seen, the percentage of fatal accidents under different conditions has risen with increasing speed levels. Variables of the type of highway and freeway showed the highest number of fatal accidents with rising speed. Highway fatalities can be reduced by about 10% if measures are taken to get the speed limit to an average level. Also, concerning road characteristics, arcs are said to increase fatal accidents at high speeds, and if speed limits are reduced to a medium level before the arcs, fatal accidents will decrease by about 10%.

Reducing two-way roads could reduce head-on collisions as 14% of fatal accidents can be avoided.

Table 4. Variations of the percentage of fatal accidents

Main variables	Secondary variables	Low speed	Medium speed	High speed
Type of roads	Freeway	-2.7	-1.3	11.9
	Highway	-0.7	3	13.2
	Main road	-1.8	0.9	2.8
	rural road	-3.8	-2	3.9
	Secondary	-3.1	0.4	4.1
Road characteristics	Arc	-0.7	4.7	14.5
	Intersection	-2.8	-0.5	13
	Arc - intersection	-2.1	1.4	6.6
	Straight	-2.7	0.6	5.2
Type of collision	Head-on	-0.8	4.6	13.6
	Front to back	-3.3	-1.3	11.7
	Rollover	-3.3	-0.5	4.4
	Lateral barrier collision	-2.1	-1.6	5.8
	Side collision	-3.7	-2.2.	-1.6

5. Conclusion

This study used three machine learning methods of naive Bayes classifier, random forest, and support vector machine to predict accidents. Also, various methods of data balancing, including the random method, clustering-based methods such as the k-means method, and intelligent genetic optimization, and particle swarm methods, were evaluated. In general, the following conclusions can be learned from this study.

• Generally speaking, the accuracy of the models developed for handling balanced data was greater than that of models developed with raw data because machine learning models ignored injury crashes in raw data.

• With regards to the data balancing methods, the meta-heuristic optimization methods had a better performance than the random and k-means methods because the accuracy of SVM method for data balanced with random, k-means and genetic algorithm methods are equal to 0.71, 0.92, and 0.98 respectively.

• The support vector machine method had greater accuracy than the Bayes classification and random forest methods, and the model developed by the support vector machine was found to be the best model presented in this study as it used data balanced by the genetic algorithm method. Here, the accuracy of training and testing data was 0.98 and 0.96, respectively.

• According to the results of sensitivity analysis, the probability of fatal accidents was much higher when two vehicles were involved than one single vehicle.

• According to sensitivity analysis, the probability of fatal accidents at high speeds in the arc was about three times that of the straight roads, which could be decreased by taking measures to reduce the speed limits.

• Highway, arc, and head-on variables were the most critical variables causing fatal crashes at high speeds according to sensitivity analysis.

6. Future Research

Although in this study an attempt has been made to build a comprehensive road crash severity prediction model with a machine learning approach, but it is suggested that parameters such as different weather conditions, topographies, human behaviors and among others in each geographical area in future studies.

Also, the effect of roadside condition variable along with the variables expressed in this study on crash severity can be an interesting topic for future studies.

In addition, it can be interesting to study the over-sampling methods and compare them with the method presented in this study.

7. References

- Abd Elrahman, S. M. and Abraham, A. (2013) ″A review of class imbalance problem″, Journal of Network and Innovative Computing, vol. 1, no. 2013, pp. 332–340.

- Abdelwahab, H. T. and Abdel-Aty, M. A. (2001) "Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections", Transportation Resarch Record, vol. 1746, no. 1, 2001, pp. 6–13.

- Abou Elassad, Z. E., Mousannif, H. and H. Al Moatassime, (2020) "A real-time crash prediction fusion framework: An imbalance-aware strategy for collision avoidance systems", Transportation Research Part C: Emerging Technologies, vol. 118, 2020, p. 102708.

- Ahmed, M.M. and Abdel-Aty, M. A. (2011) "The viability of using automatic vehicle identification data for real-time crash prediction" IEEE Transactions On Intelligent Transportation Systems, vol. 13, no. 2, 2011, pp. 459–468.

- Ba, Y., Zhang, W., Wang, Q., Zhou, R. and C. Ren, (2017) "Crash prediction with behavioral and physiological features for advanced vehicle collision avoidance system", Transportation Research Part C: Emerging Technologies, vol. 74, 2017, pp. 22–33.

- Basso, F., Basso, L. J., Bravo, F. and Pezoa, R. (2018) "Real-time crash prediction in an urban expressway using disaggregated data", Transportation Research Part C: Emerging Technologies, vol. 86, 2018, pp. 202–219.

- Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002) ″SMOTE: synthetic minority over-sampling technique", Journal of Artificial Intelligence Resaerch, vol. 16, 2002, pp. 321–357.

- Cortes, C. and Vapnik, V. (1995) "Support-vector networks", Machine Learning, vol. 20, no. 3, 1995, pp. 273–297.

- Elamrani, Z., Abou Elassad, Mousannif, H. and Al Moatassime, H. (2020) "Class-imbalanced crash prediction based on real-time traffic and weather data: a driving simulator study", Traffic Injury Prevention, vol. 21, no. 3, 2020, pp. 201–208.

- Holland, J. H. (1992) "Genetic algorithms", Scientific American, vol. 267, no. 1, 1992, pp. 66–73.

- Karami, Ali., Hadji Hosseinlou, Mansour., Abbasi, Mohammad Hossein. and Figuerira, Monteiro. (2020) "Priority Order for Improvement of Intersections using Pedestrian Crash Prediction Model", International Journal of Transportation Engineering, vol.7, 2020, pp. 297-313.

- Keogh, E. (2017) "Naive bayes classifier", Accessed Nov, vol. 5, 2006, p. 2017.

- Le Yu, Bowen Du, Xiao Hu, Leilei Sun, Liangzhe Han, Weifeng Lv, (2021) "deep spatio-temporal graph convolutional network for traffic accident prediction", Neurocomputing, vol.423, 2021,pp. 135-147.

- Li, L., He, S., Zhang, J. and Ran, B. (2016) "Short-term highway traffic flow prediction based on a hybrid strategy considering temporal–spatial information", Journal Of Advanced Transportation, vol. 50, no. 8, 2016, pp. 2029–2040.

- Li, Y., Ma, D., Zhu, M., Zeng, Z. and Wang, Y. (2018) "Identification of significant factors in fatal-injury highway crashes using genetic algorithm and neural network", Accident

Analysis & Prevention, vol. 111, 2018, pp. 354–363.

- Li, X., Lord, D., Zhang, Y. and Xie, Y. (2008) "Predicting motor vehicle crashes using support vector machine models", Accident Analysis & Prevention, vol. 40, no. 4, 2008, pp. 1611–1618.

- Likas, A., Vlassis, N. and Verbeek, J. J. (2003) "The global k-means clustering algorithm", Pattern Recognition, vol. 36, no. 2, 2003, pp. 451–461.

- Mirbaha, Babak., Saffarzadeh, Mahmoud. and Noruzoliaee, Mohammad Hossein. (2013) "A Model for Predicting Schoolchildren Accidents in the Vicinity of Rural Roads based on Geometric Design and Traffic Conditions", International Journal of Transportation Engineering, vol.1, 2013, pp. 25-33.

- Mirjalili, S. (2019) "Genetic algorithm", Evolutionary Algorithms and Neural Networks, vol. 780, Springer, 2019, pp. 43–55.

- Nguyen, G. H., Bouzerdoum, A. and Phung, S. L. (2009) "Learning pattern classification tasks with imbalanced data sets", Pattern Recognit, 2009, pp. 193–208.

- Niveditha, V., Ramesh, A., Kumar, M. (2015) "Development of Models for Crash Prediction and Collision Estimation- A Case Study for Hyderabad City", International Journal of Transportation Engineering, vol.3, 2015, pp. 143-150.

- Safak, V. (2020) "Min-Mid-Max Scaling, Limits of Agreement, and Agreement Score", arXiv, 2020.

- Sinaga, K. P. and Yang, M.-S. (2020) "Unsupervised K-means clustering algorithm", IEEE Access, vol. 8, 2020, pp. 80716–80727.

- Tabachnick, B. G., Fidell, L. S. and Ullman, J. B. (2007) “Using multivariate statistics”, vol. 5. Pearson Boston, MA, 2007.

- Theofilatos, A., Chen, C. and Antoniou, C. (2019) "Comparing machine learning and deep learning methods for real-time crash prediction", Transportation Resarch Record, vol. 2673, no. 8, 2019, pp. 169–178.

- Vapnik, V., Guyon, I. and Hastie, T. (1995) "Support vector machines", Machine Learning, vol. 20, no. 3, 1995, pp. 273–297.

- W. H. Organization, Global status report on road safety 2018. World Health Organization, 2018.

- Wang, C., Xu, C. and Dai, Y. (2019) "A crash prediction method based on bivariate extreme value theory and video-based vehicle trajectory data", Accident Analysis & Prevention, vol. 123, 2019, pp. 365–373.

- Wang, D., Tan, D. and Liu, L. (2018) "Particle swarm optimization algorithm: an overview", Soft Computing, vol. 22, no. 2, 2018, pp. 387–408.

- Washington, S., Haque, M. M., Oh, J. and Lee, D. (2014) "Applying quantile regression for modeling equivalent property damage only crashes to identify accident blackspots", Accident Analysis & Prevention, vol. 66, 2014, pp. 136–146.

- Whitley, D. (1994) "A genetic algorithm tutorial", Statistics and Computing, vol. 4, no. 2,1994, pp. 65–85.

- Yen, S.-J. and Lee, Y.-S. (2009) "Cluster-based under-sampling approaches for imbalanced data distributions", Expert Systems With Applications, vol. 36, no. 3, 2009, pp. 5718–5727.

- Yang, I.T. (2007) "Performing complex project crashing analysis with aid of particle swarm optimization algorithm", International Journal of Projoject Management, vol. 25, no. 6, 2007, pp. 637–646.