Crash Severity Predictive Models Using Machine Learning Algorithms: a Case Study of Riyadh, Saudi Arabia

Hussein Bachir
Master Student at the Civil Engineering Department
College of Engineering, King Saud University
Riyadh, Saudi Arabia
442106444@student.ksu.edu.sa

Mohammed Almannaa
Assistant Professor at the Civil Engineering Department
College of Engineering, King Saud University
Riyadh, Saudi Arabia
malmannaa@ksu.edu.sa

Abstract

Machine learning models have shown high prediction accuracy as a result of their freedom from the limitations of data distribution assumptions in classical statistical methods, non-compliance with which leads to incorrect and inaccurate results. One of the most important applications of machine learning is to predict the severity of traffic crashes according to a number of independent factors related to the crash. This paper aims to compare four machine learning methods: logistic regression, k-nearest neighbor, decision trees, and random forests. More than 40,000 crashes occurred in Riyadh, Saudi Arabia, during 2012-2016 were used. It was found that Decision Trees and Random Forests are the best algorithms in terms of accuracy, and the logistic regression is the weakest. The type of crash was the most important factor to the crash severity, followed by the time of crash. On the other hand, the number of parties involved in the crash and the lighting condition were the lowest.

Keywords

Machine Learning; Crash Severity; Predictive models; Riyadh; Road Safety.

1. Introduction

The interest in studying traffic crashes and their consequences has grown in recent years as the number of traffic crashes has increased due to increased urbanization and car use around the world. According to a World Health Organization report from 2018, the number of people killed in traffic crashes has risen to more than 1.35 million (de Groot 2018). These statistics prompted traffic specialists and decision-makers to look for analytical methods to improve road safety. One of which is identifying factors related to the risk of injury to the parties involved in the crash (Santos 2021). Predicting the severity of crash can be used by three main parties: transportation safety engineers who want to predict the costs of crash in the coming years, hospitals and emergency health care providers who need to anticipate the severity of the injury so that they can provide appropriate medical care in advance, and insurance companies whose customers' premiums are determined by a number of factors including the costs of the traffic crash (Iranitalab and Khattak 2017).

Traditional statistical methods have been widely used in traffic safety research, in which models provide good indicators of collision probability and facilitate results interpretation (Iranitalab and Khattak 2017). However, main attributes of crash data frequently lead to methodological limitations that are not fully considered, and these techniques are subject to strict assumptions about data distribution and predetermined relationships that may not be valid, resulting in erroneous estimates and incorrect conclusions (Iranitalab and Khattak 2017, Chakraborty 2021). Machine learning techniques have emerged as a new method for analysing traffic crash data to address this issue. Machine learning techniques are distinguished by their computational power and the absence of the need for assumptions about the fundamental relationships between variables. Public authorities are the primary source of traffic crash data used in the development of predictive models (Al-Mistarehi,2022). However, some of these data sets have an imbalance problem because the number of crashes resulting in death or serious injury is much lower than the number of crashes resulting in minor injuries or property damage, and thus it influences negatively the accuracy of the model (i.e. increases the bias).

1.1 Objectives

This paper makes use of four machine learning algorithms in predicting the severity of traffic crashes: logistic regression, K-nearest neighbor, decision trees, and random forest. In Section Two, the paper presents some highlights of previous work that used these algorithms to build predictive models. The Third Section discusses the research methodology, which includes a presentation of the study area, a description of the data set used, its strengthens and weakness, and four methods used for implementing the four algorithms. In Section Four, the results inducing the accuracy for each algorithm are presented separately, and finally the results are discussed in Section Five.

2. Literature Review

This review of the literature focuses on crash severity modelling and the application of prediction and classification methods in traffic crash research. There are two types of traffic crash severity prediction models: statistical and machine learning. Classification algorithms predict only discrete (or categorical) outputs, whereas regression algorithms predict both discrete and continuous outputs. Huang et al. used Byesian binomial hierarchical models with random effects in order to calculate the common unobserved factors shared by the crash parties involved (Huang et al. 2008). Their model outperformed the standard logistic model in terms of accuracy of the risk factor estimates as the standard logistic model was unable to explain the model's residual variance of 28.9%. Winston et al. investigated the relationship between drivers' collision exposure and crash risk and the presence of airbags or anti-lock brakes in their vehicles (Winston et al. 2006). A multivariate model was developed that simultaneously modelled a series of four binary outcomes: the presence of airbags, the presence of anti-lock brakes, the probability of occurring the crash, and the possibility of injury. Ouyang et al. used a simultaneous binary logical model to study injuries in car and truck collisions (Ouyang et al. 2008). They compared their results with simple binary logical models for individual injuries that assume there is no correlation between injuries that occur in the same collision. Their model showed significant gains in efficiency when there is a high correlation between injury severity. It is simple to apply to multiple severity classes in multiple vehicle collisions. Milton et al. created a mixed logit model that identifies areas with an unusually high number of traffic fatalities (Milton et al., 2008). The mixed log model provides flexibility in capturing road sector specific heterogeneity, which can result from a variety of factors related to the road, environment, driver behavior, and their interactions.

In the field of Machine learning and Artificial Intelligence, Iranitalab and Khattak (2017) developed a cost-based approach to compare the performance of four machine learning algorithms: Multinomial logit (MNL), Nearest Neighbor Classification (NNC), Support Vector Machines (SVM), and Random Forests (RF). These four models were used to predict the severity of traffic crashes. The result revealed that MNL outperformed others in predicting PDO (Property damage Only), while NNC outperformed others in the severity level of crashes. The results confirm previous works that the MNL model's strong assumptions about data cause its poor performance, whereas the other three functional learning methods are more flexible. Rahim et al. (2021) proposed a new deep learning approach with F1 loss function dedicated to predicting the severity of traffic crashes in work areas. They used recall and model accuracy as matrices, relying on transforming variables into images using t-SNE technology to reduce nonlinear dimensions and convex algorithm hull.

The results showed that the proposed deep learning model outperformed the SVM model in classifying fatal crashes and injuries. Based on 5973 records of traffic crashes that occurred in Abu Dhabi between 2008 and 2013, an artificial neural network was used to predict the severity of traffic crashes. These data were used to create two different classifiers. To train and validate the first classifier, the entire data set was used. For the second classifier, 90% of the data was used for training, and 10% was used for testing. The overall prediction performance of the training and testing data were 81.6% and 74.6%, respectively, and the neural network's prediction accuracy for fatal, severe, moderate, and low-severity crashes were 4.5%, 10.2%, 80.1%, and 94.5% respectively. For comparison purposes, the ordered probability model was used and compared with the artificial neural network, and they found the latter outperformed the former in terms of prediction accuracy. Chen et al. (2017) developed Support Vector Machine (SVM) models to predict the outcome of driver injury severity in rollover crashes. Driver injury severity was aggregated as a three-level categorical variable.

Two common kernel functions, heterogeneous polynomials, and Gaussian RBF kernels were used to investigate the applicability and performance of SVM models in predicting driver injuries. The cubed SVM classifier outperformed the average Gaussian RBF SVM, scoring higher in the no-injury category than in the disabling and fatal injury categories. Further literature was reviewed that includes the algorithms that will be studied in this paper. Rezapour et al. (2020) investigated the feasibility of using parametric (binary logistic regression) and non-parametric (classification tree) methods in predicting motorcycle injury severity. K-fold validation was used to address the bias associated with test data selection, and the performance of the two models was similar in terms of predicting the severity of injury, but the parametric model outperforms slightly its counterpart. Shiran et al. (2021) adopted MLR, DT, and the ANN-MLP models to predict the severity of crashes on highways in California, USA. The results showed the MLR model revealed significant correlations between a set of factors (crash cause, weather conditions, road surface, number of vehicles) and the severity of the crash, and the decision tree revealed that the causes of the collision and the number of vehicles were among the most important variables influencing the severity of the crash.

In summary, machine learning algorithms produced more accurate results because they were not constrained by assumptions about data distribution. However, it can be argued that determining whether or not a model is accurate is primarily dependent on the data used, with some logistic regression models outperforming their peers. The majority of the previous research works were intended in studying and comparing the accuracy of the models used, and some of them were interested in demonstrating the effect of independent variables on the severity of the crash.

3. Methods

3.1 Logistic Regression

Logistic regression is a classification and predictive analytic method that estimates the likelihood of an event's occurrence based on a set of independent variables (IBM,2022). In the normal case, since the outcome is a probability, the dependent variable is restricted be-tween 0 and 1. However, a response variable with more than two classes is needed (Hastie,2021). For example, in this paper, crashes are categorized into four levels: (1) PDO, Minor Injury (3) Severe Injury (4) Death. The logistic regression equation can be represented as follows (Hastie ,2021):

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta 0 \ + \ \beta 1 X1 \ + \cdots \ + \beta p X p \qquad (1)$$

The probability of a crash occurring with a severity level can be calculated after estimating the model, and the severity level can be predicted based on the calculated probabilities.

3.2 K-nearest neighbour

The K-nearest neighbours' algorithm is a moderated, non-parametric learning classifier that uses proximity to make classifications or predictions about the aggregation of an individual data point. KNN is usually used as a classification algorithm, and it works on the assumption that similar points can be found close to each other (Elbaghdadi et al.,2021). Given a positive integer of K and a test observation (x0), the KNN classifier first selects the K neighbour points in the training data closest to X0. Then, it estimates the conditional probability as a fraction of the points in N0 that equal the response values (Cutler and Dickenson,2020). A Euclidean distance function is needed to implement the algorithm. The distance function used is as follows:

$$d_{ij}=\left(\sum_{k=1}^{p}(xik - xjk)^2\right)^{\frac{1}{2}} \qquad (2)$$

3.3 Decision Trees

Decision trees are widely used for classification and regression tasks. In theory, the decision is made through a hierarchy of "if/else" questions (Cutler and Dickenson,2020). The hierarchical structure of the tree consists of a root node, which does not contain any afferent branches. Outgoing branches from the root node then feed into internal nodes known as decision nodes. Both node types perform evaluations to form homogeneous subgroups referred to by their leaf nodes. The terminal nodes represent all possible outcomes within the data set (de Ville,2013).

3.4 Random Forests

A disadvantage of decision trees is that they tend to overfitting the data. Random Forests is one way to address this problem. It is basically a set of decision trees, where each tree is slightly different from the others, and each tree may do a relatively good job of forecasting, but it uses a part of the data (Cutler and Dickenson,2020). Each tree within the algorithm consists of a sample of data taken from a training set of replacement, called a bootstrap sample. One third of this sample is set aside as test data. Then, another example of randomization is injected, which adds more diversity and reduces the correlation between decision trees, depending on the type of problem the prediction selection will vary. The sample taken aside earlier is used for cross validation and finalization of this prediction.

3.4 K-means Clustering

K-means is a clustering approach used to categorize various observations into particular groups by their internally homogeneous and externally heterogeneous characteristics (Li et al.,2018). The aggregation process begins by defining a number of points to create the initial centers of the desired groups, then linking each observation to the nearest center to create temporary groups (Jain,2010). Observations are assigned to clusters by calculating the distance between each object and all other centers based on the Euclidean distance, then the closest distance is chosen. The process of updating centers and resetting the cluster objects is repeated until the observations is stable (Aljofey,2018). Determining the best number of expected ensembles is one of the challenges of aggregation techniques. The k-mean algorithm requires the number of groups to be entered (k). Clusters are created to reduce intra-cluster variance. In this research work, the same characteristics entered in prediction models were used to classify crash severity into four categories.

4. Data Collection

Riyadh city is Saudi Arabia's capital and the Arabian Peninsula's largest urban area (Figure 1). The population of Riyadh's region was estimated to be 8.9 million in 2020, a 3 million increase from 2007 (KAPSARC,2022) Economic and population growth have both resulted in an increase in the need for mobility, which has resulted in an increase in car ownership and use (Aldalbahi,2015). Because of this increase, Riyadh in particular and Saudi Arabia in general are suffering from a rise in traffic crash rates, as the year of 2018 witnessed the occurring about more than 80,000 traffic crashes in the city of

Riyadh, causing 1091 deaths and 4,554 injuries (The Saudi General Authority for Statistics,2022). According to the report of the General Traffic Department, the causes of these crashes include wrong turning, wrong overtaking, sudden stopping, excessive speed, and not complying with the traffic signals.

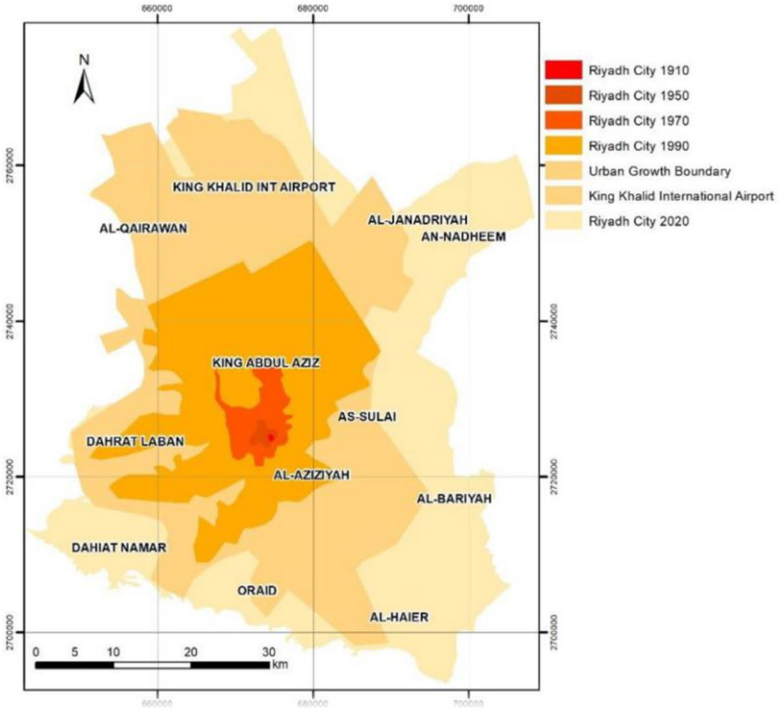
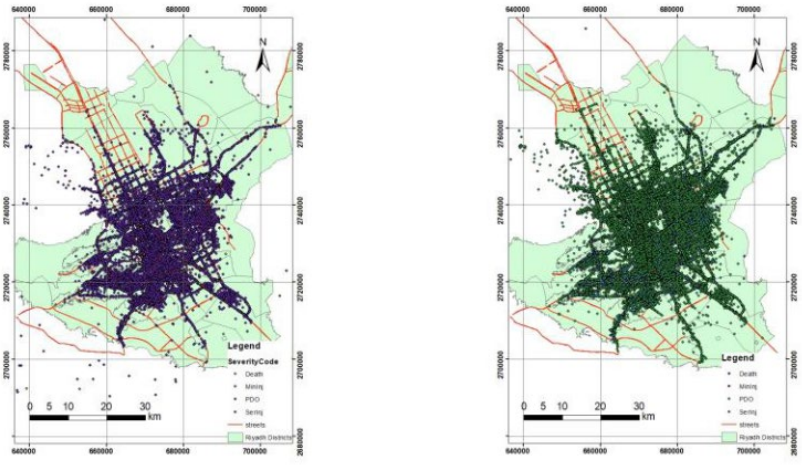


Figure 1. Riyadh City and the urban growth boundaries (Riyadh Municipality,2021)

The details of the crashes, vehicles and persons involved in the crashes were obtained from the crash dataset collected by the Saudi General Directorate of Traffic. This data includes the location of the crash (latitude and longitude), the age of driver, causes to the crash (as registered by the policeman), the condition of the road surface and lighting conditions, and the weather condition. Around 467,273 traffic crashes were occurred between 2012-2016 in Riyadh. The data suffered from some deficiencies such as lack of geocoding of some crashes, and miscoding of others, and thus pre-processing phase was conducted by presenting crashes spatially using ArcGIS and getting rid of any miscoded or in-complete crashes (Figure 2). Data that included a lack of road surface condition, weather condition and lighting condition combined were eliminated, and the road surface was linked to weather condition assuming that the road surface was wet in the case of rainy weather, and dry in the other cases. Exploratory analysis of the data showed that 97% of crashes caused property damage only (PDO), 2.3% of crashes caused serious injury, and the rest was fatalities. Weekends recorded lower levels of crashes than on working days (Figure 3), and it was interesting to observe a remarkably huge number of crashes at 1 p.m., (Figure 4) and most of which were related to fatigue and over speeding. After the pre-processing phase, the number of crashes observed for the five years was reduced from 467,273 to 270,577. Table 1 shows the variables in the dataset that will be adopted into the model as depended variables to estimate the severity of the crash.

Table 1. Dependent and Independent variables used in the models.

Variables	Classes	Classes Count
Crash Point	Straight Road	7018
	Curve	6
	Non-Signalized Intersection	25
	Signalized Intersection	1
	(T) Intersection	7
	Other	263520
Land Surface	Dry	269983
	Wet	294
Weather	Clear	268004
	Cloudy	403
	Dust	1836
	Rainy	35
Light Conditions	Clear	269777
	Dark	500
Crash Type	Fire in the vehicle	252259
	Rollover	1142
	Run over	661
	Falling from a bridge	91
	Falling from a cliff	61
	Hitting a Traffic light	158
	Hitting Stationary object crash	348
	Hitting a Side Barrier Crash	2721
	Hitting an Animal	1045
	Motorcycle Crash	178
	Hitting a Road Fence	19
	Hitting an Electric Pole	51
	Hitting a Parked Vehicle	14
	Hitting a Tree	28
	Hitting a Plate	5992
Severity	PDO	262171
	Minor Injuries	354
	Serious Injuries	6227
	Death	1525
Day	Sunday	39780
	Monday	40771
	Tuesday	40391
	Wednesday	39677
	Thursday	39520
	Friday	33623
	Saturday	36515



(a) **(b)**

Figure 2. The spatial distribution and severity level of crashes in Riyadh City: (a) Between 2012-2013; (b) Between 2014-2015

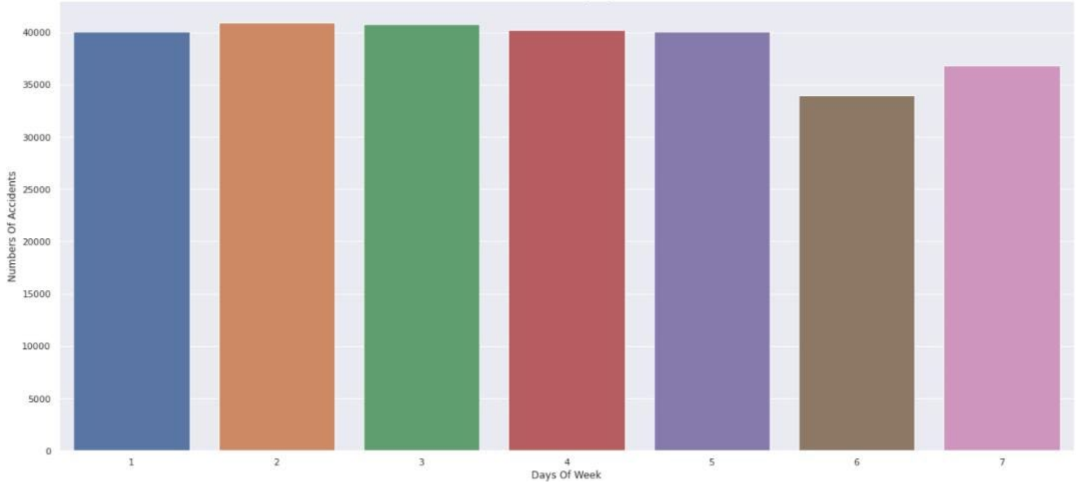


Figure 3. Number of crashes per day of week (1, sunday – 7, saturday)

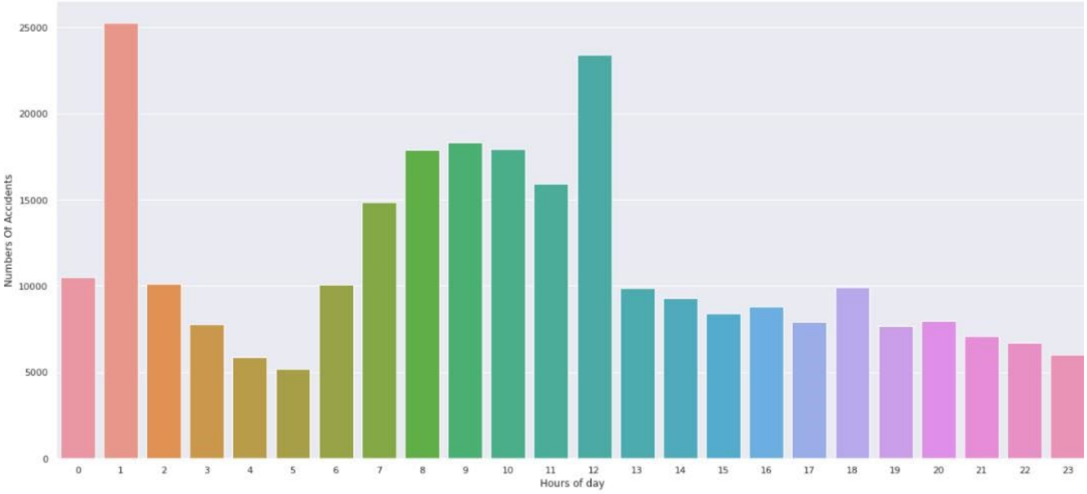


Figure 4. Number of crashes per time of day

5. Results and Discussion

5.1 Descriptive Analysis

Table 2 illustrates the connections between the independent factors and the severity of the accident. For example, when accidents with injuries and fatalities are considered, it is discovered that the ensuing accidents with significant injuries account for the majority of accidents. It was discovered that there was a link between the kind of accident and the degree of the injury, with accidents caused by a fire in the car and accidents caused by a collision with a permanent barrier accounting for the greatest number of accidents resulting in serious injuries. The independent and dependent variables were subjected to a chi-square test to ensure that there is a significant relationship between them, and thus confirm the validity of the used hypothesis. Table 3 shows the test results.

Table 2. Cross Table of crash severity and independent variables

		Crash Severity			
		PDO	Minor Injuries	Serious Injuries	Fatal
Crash Point	Other	255886	354	5860	1420
	Straight Road	6579	2	342	95
	Curve	2	0	3	1
	Non-Signalized Intersection	0	0	18	7
	Signalized Intersection	0	0	1	0
	(T) Intersection	2	0	3	2
Land Surface	Dry	262183	356	6209	1521
	Wet	286	0	18	4
Accident Type	Fire in the vehicle	248484	347	3103	625
	Rollover	1026	1	85	30
	Run over	286	1	319	55
	Falling from a bridge	56	0	29	6
	Falling from a cliff	56	0	2	3
	Hitting a Traffic light	30	0	106	22
	Hitting Stationary object crash	246	0	68	34
	Hitting a Side Barrier Crash	1082	5	1255	379
	Hitting an Animal	620	0	263	162
	Motorcycle Crash	170	0	4	4
	Hitting a Road Fence	7	0	10	2
	Hitting an Electric Pole	17	0	25	9
	Hitting a Parked Vehicle	10	0	1	3
	Hitting a Tree	24	0	1	3
	Hitting a Plate	4975	2	859	156
Light	Clear	262052	355	6167	1503
	Dark	417	1	60	22
Parties Involved	1	259184	349	4896	1172
	2	3269	6	1196	306
	3	16	1	124	41
	4	0	0	11	6

Table 3. Chi-square test results

variable	x-squared	degree of freedom	p-value
Accident Type	58,74	54	2.2e-16
Land Surface	20,725	3	0,00012
Weather	185,5	12	2.2e-16

Light Condition	345,29	3	2.2e-16

5.2 Predictive Analysis

The four prediction algorithms were adopted and trained using the crash data set collected in Riyadh, Saudi Arabia during 2016-2016. 5-fold Cross validation was conducted. The training data represents 80% of the data set, while the test data represented 20%. A correlation test was performed for all the variables. The correlation test showed a significant relationship between the number of parties involved in the crash and the type of crash, as this was explained by the number of reasons, in which it expresses a reason for each party. Figure 5 shows the correlation values for the 13 independent variables. The Logistic Regression model was tested using all 13 independent variables, and the accuracy of the model was calculated, which was 0.971. For this model, a confusion matrix was built. This demonstrated that 52465 of the outcomes in the test data are actual positive results, indicating that the model properly anticipated the positive outcome.

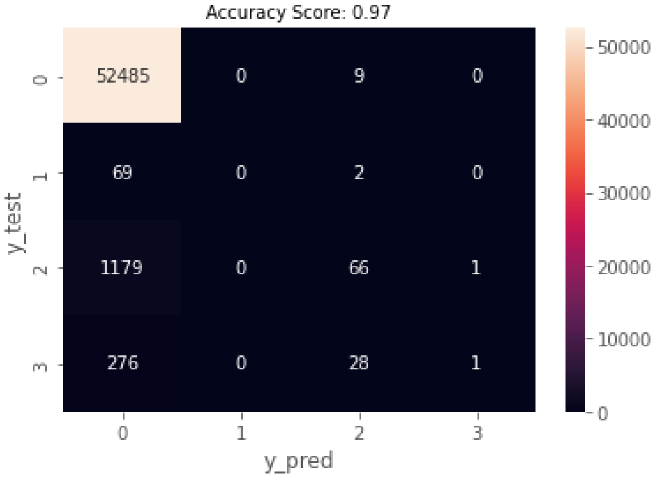


Figure 5. Correlation Test between Severity and Independent variables

The number of clusters for K-NN were determined by trying different values for both training and testing data sets to find the best prediction accuracy. Values from 3 to 8 were selected as shown in Figure 5. The figure shows the accuracy values for both the training and testing data. The test data recorded a relatively stable accuracy level from the fourth neighbor to the eighth neighbor, with a slight decrease at the fifth and sixth neighbors. The model recorded an accuracy of 0.974 using the crash dataset.

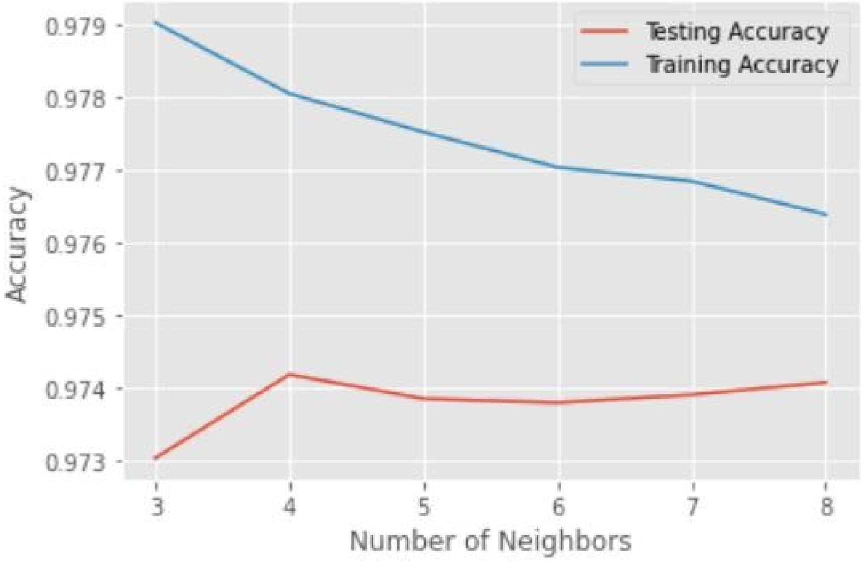


Figure 6. Number of Neighbors in KNN and their accuracies

In the decision tree, two metrics were used: entropy and Gini. The Gini metric is intended to measure how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset. The probability of misclassification is inversely proportional to the value of the Gini metric. The measure of entropy describes the state of randomness or perturbation in a node, so nodes with more diverse composition are considered to have higher entropy. The results of the accuracy test showed that the decision tree was recorded in both metrics, with an average accuracy of 0.976.

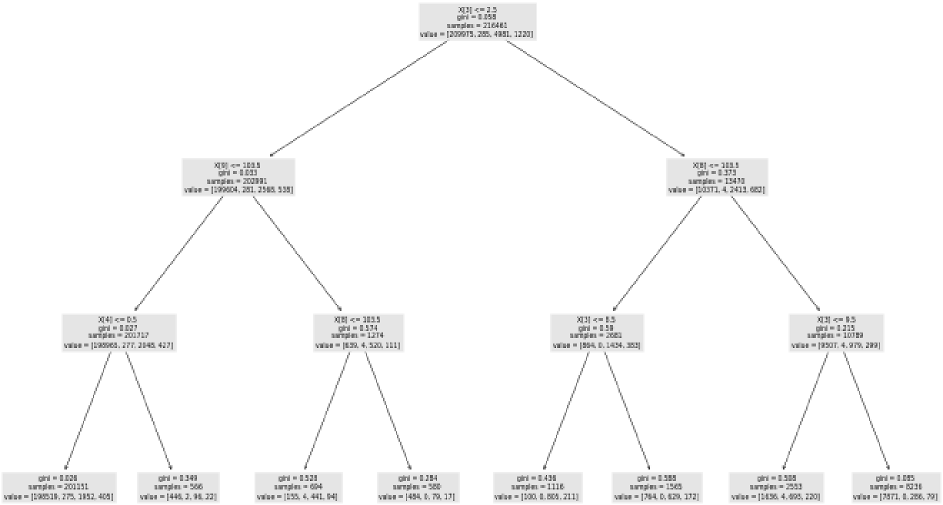


Figure 7. Decision Tree with maximum depth 3

The training of Random Forests needs to adjust the number of trees to grow and the number of variables from which samples are taken randomly to ensure that each insertion row is expected at least several times and that the results obtained using random seeds are not usually different (Strobl,2008). Using the crash dataset, the Random Forest recorded an accuracy of 0.976. Figure 6 shows the most important variables affecting the crash severity, in which the type of crash is in the top of the most important features, followed by the time of crash. On the other hand, the number of parties involved in the crash and the lighting condition were the lowest.

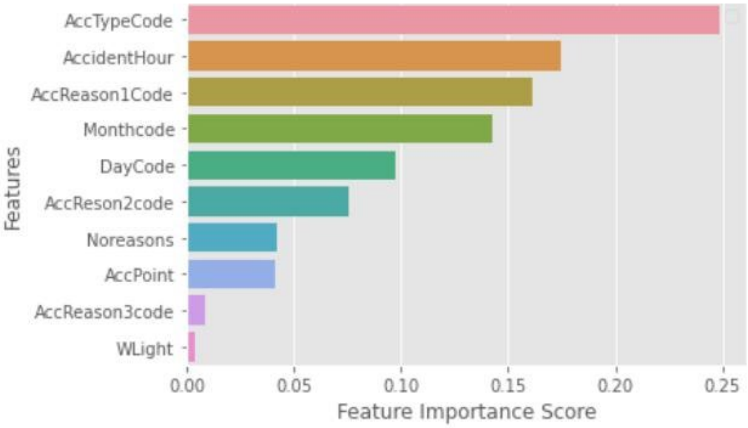


Figure 8. Feature important scores

As can be seen, both Decision Tree and Random Forest algorithms scored the highest accuracy at all levels of crash severity, compared to the other two algorithms. The K-nearest neighbor ranked second in the levels of injury and death, while it was last in the level of PDO. The Logistic Regression ranked last in the levels of injuries and deaths while it was ranked second in the level of PDO. The most important depended variables were found for the type, timing, and cause of the crash.

6. Conclusion

Classification of crashes based on their severity is critical in order to use appropriate data in a crash cost analysis. Moreover, in crash severity analysis, accurate prediction models are necessary to classify crashes based on their severity. In this work, statistical and machine learning algorithms are used to build models that predict crash severity at four different levels such as PDO, minor injury, severe injury, and fatality. Four models were implemented: logistic regression, K-nearest neighbor, Decision Tree, Random Forests on crash dataset for the city of Riyadh covering four years 2012-2016. 13 independent variables were used in the four models. Each mode gives different results. The findings can be summarized in the following:

1- Based on statistical testing, we may deduce that accidents with significant injuries are the most common patterns of accident severity, and that most accidents with serious injuries occur on straight highways. Fires in vehicles are the greatest cause of fatalities and serious injuries. In exchange for a minimal impact of these factors on accidents resulting in mild injuries.
2- Although the logistic regression model had the lowest accuracy rate of the four models, the confusion matrix test revealed a high number of positive correct hypotheses, indicating that the model's forecast of the selected hypothesis was right.
3- K-nearest neighbor was the second lowest model in term of accuracy with 0.974.The test data recorded a relatively stable accuracy level from the fourth neighbor to the eighth neighbor, with a slight decrease at the fifth and sixth neighbors.
4- Both the decision tree and the random forest recorded the best accuracy of 0.976. Entropy and Gini measures were used to construct the decision trees. Using the significance scale, the type and timing of the incident ranked first.

In conclusion, this research provides valuable insights into the application of machine learning algorithms in crash severity predictive models. The findings of this research demonstrate that the proposed models outperform traditional statistical methods and achieve high accuracy in predicting crash severity. Furthermore, this research provides a novel approach to predicting crash severity by incorporating multiple factors, including weather, road conditions, and vehicle type, into the prediction process. This makes the proposed models a valuable contribution to the field of road safety and offers a potential solution for reducing the number of fatal crashes on our roads. The results of this research will serve as a reference for future studies and can be applied in real-world situations to reduce the risk of crashes and improve road safety. In comparison to previous related work, this research provides a more comprehensive and sophisticated approach to predicting crash severity and makes an additive value by offering a data-driven solution for mitigating road crashes.

References

Aldalbahi, M.; Walker, G. Riyadh Transportation History and Developing Vision. *Procedia Soc Behav Sci*, vol. 216, pp. 163–171, 2016.

Aljofey, A.M.; Alwagih, K. Analysis of Crash Times for Highway Locations Using K-Means Clustering and Decision Rules Extracted from Decision Trees. *International Journal of Computer Applications Technology and Research*, vol. 7, pp. 1–11, 2018.

Al-Mistarehi, B.W.; Alomari, A.H.; Imam, R.; Mashaqba, M. Using Machine Learning Models to Forecast Severity Level of Traffic Crashes by R Studio and ArcGIS. *Front Built Environ*, *8*, 2022.

Chakraborty, M.; Gates, T.; Sinha, S. Causal Analysis and Classification of Traffic Crash Injury Severity Using Machine Learning Algorithms.; 2021.

Chen, C.; Zhang, G.; Qian, Z.; Tarefder, R.A.; Tian, Z. Investigating Driver Injury Severity Patterns in Rollover Crashes Using Support Vector Machine Models. *Accidents Analysis and Prevention*, Vol. *90*, pp. 128–139, 2016.

Cutler, J.; Dickenson, M. *Introduction to Machine Learning with Python*, OREILLY, 2020.

de Groot, K. Global Status Report on Road Safety 2018- Un. *World Dev*, no. 1, pp. 1–15, 2018.

de Ville, B. Decision Trees: Decision Trees. *WIREs Comput Stat*, 5, No. 448–455, 2013.

Elbaghdadi, A.; Mezroui, S.; el Oualkadi, A. K-Nearest Neighbors Algorithm (KNN), pp. 161–178, 2020.

Hastie, T.; Tibshirani, R.; James, G.; Witten, D. *An Introduction to Statistical Learning* ,2nd Ed.,.Vol. 102, Springer, 2021.

Huang, H.; Chin, H.C.; Haque, M.M. Severity of Driver Injury and Vehicle Damage in Traffic Crashes at Intersections: A Bayesian Hierarchical Analysis. *Accident Analysis and Prevention*, No, *40*, pp. 45–54, 2008.

IBM What Is Logistic Regression? | IBM, IBM, 2022.

Iranitalab, A.; Khattak, A. Comparison of Four Statistical and Machine Learning Methods for Crash Severity Prediction. *Accident Analysis and Prevention*, No. *108*, pp. 27–36, 2017.

Jain, A.K. Data Clustering: 50 Years beyond K-Means. *Pattern Recognit Lett*, No. 31, pp. 651–666, 2010.

Li, Z.; Chen, C.; Ci, Y.; Zhang, G.; Wu, Q.; Liu, C.; Qian, Z. (Sean) Examining Driver Injury Severity in Intersection-Related Crashes Using Cluster Analysis and Hierarchical Bayesian Models. *Accident Analysis and Prevention*, No. 120, pp. 139–151, 2018.

Milton, J.C.; Shankar, V.N.; Mannering, F.L. Highway Crash Severities and the Mixed Logit Model: An Exploratory Empirical Analysis. *Accident Analysis and Prevention*, No. *40*, pp. 260–266,2018.

Ouyang, Y.; Shankar, V.; Yamamoto, T. Modeling the Simultaneity in Injury Causation in Multivehicle Collisions. *Transportation Research Record*, pp. 143–152, 2002.

Population by Administrative Region and Gender — KAPSARC Data Portal.

Rahim, M.A.; Hassan, H.M. A Deep Learning Based Traffic Crash Severity Prediction Framework. *Accident Analysis and Prevention*, No. *154*,2021.

Rezapour, M.; Mehrara Molan, A.; Ksaibati, K. Analyzing Injury Severity of Motorcycle At-Fault Crashes Using Machine Learning Techniques, Decision Tree and Logistic Regression Models. *International Journal of Transportation Science and Technology*, No. *9*, pp. 89–99, 2020**.**

Riyadh Municipality. Available online: https://www.alriyadh.gov.sa/ar/riyadh/popudev (accessed on 5 May 2021).

Santos, K.; Dias, J.P.; Amado, C. A Literature Review of Machine Learning Algorithms for Crash Injury Severity Prediction. *Journal of Safety Research*, No. *80*, pp. 254–269, 2022.

Shiran, G.; Imaninasab, R.; Khayamim, R. Crash Severity Analysis of Highways Based on Multinomial Logistic Regression Model, Decision Tree Techniques and Artificial Neural Network: A Modeling Comparison. *Sustainability (Switzerland)* 13, 2021.

Strobl, C.; Boulesteix, A.L.; Kneib, T.; Augustin, T.; Zeileis, A. Conditional Variable Importance for Random Forests. *BMC Bioinformatics*, 9, 2008.

The Saudi General Authority for Statistics, the Fifty-fifty issue of the Statistical Yearbook, https://www.stats.gov.sa/sites/default/files/14-12_1.xlsx ,2022.

Total Injuries and deaths 1439, https://data.gov.sa/Data/ar/dataset/traffic-accident-statistics-as-of-1438-h, 2022-

Winston, C.; Maheshri, V.; Mannering, F. An Exploration of the Offset Hypothesis Using Disaggregate Data: The Case of Airbags and Antilock Brakes. *J Risk Uncertain* **2006**, *32*, 83–99, doi:10.1007/s11166-006-8288-7.

Biographies

Hussien Bachir is a Master student at the Civil Engineering Department, College of Engineering, King Saud University, Riyadh, Saudi Arabia.

Mohammed Almannaa is an Assistant Professor at the Civil Engineering Department at King Saud University, Riyadh, Saudi Arabia.