sustainability

MDPI

Article

Identifying Causes of Traffic Crashes Associated with Driver Behavior Using Supervised Machine Learning Methods: Case of Highway 15 in Saudi Arabia

Darcin Akin [1,*], Virginia P. Sisiopiku [2], Ali H. Alateah [1], Ali O. Almonbhi [3], Mohammed M. H. Al-Tholaia [1] and Khaled A. Alawi Al-Sodani [1]

[1] Department of Civil Engineering, University of Hafr Al Batin, Hafr Al Batin 39524, Saudi Arabia
[2] Department of Civil, Construction, and Environmental Engineering, The University of Alabama at Birmingham, Birmingham, AL 35294, USA
[3] The Ministry of Transport and Logistic Services (MOTLS), Riyadh 12628, Saudi Arabia
* Correspondence: darcin@uhb.edu.sa; Tel.: +966-13-720-5170

Abstract: Identifying the causes of road traffic crashes (RTCs) and contributing factors is of utmost importance for developing sustainable road network plans and urban transport management. Driver-related factors are the leading causes of RTCs, and speed is claimed to be a major contributor to crash occurrences. The results reported in the literature are mixed regarding speed-crash occurrence causality on rural and urban roads. Even though recent studies shed some light on factors and the direction of effects, knowledge is still insufficient to allow for specific quantifications. Thus, this paper aimed to contribute to the analysis of speed-crash occurrence causality by identifying the road features and traffic flow parameters leading to RTCs associated with driver errors along an access-controlled major highway (761.6 km of Highway 15 between Taif and Medina) in Saudi Arabia. Binomial logistic regression (BNLOGREG) was employed to predict the probability of RTCs associated with driver errors ($p < 0.001$), and its results were compared with other supervised machine learning (ML) models, such as random forest (RF) and k-nearest neighbor (kNN) to search for more accurate predictions. The highest classification accuracy (CA) yielded by RF and BNLOGREG was 0.787, compared to kNN's 0.750. Moreover, RF resulted in the largest area under the ROC (a receiver operating characteristic) curve (AUC for RF = 0.712, BLOGREG = 0.608, and kNN = 0.643). As a result, increases in the number of lanes (NL) and daily average speed of traffic flow (ASF) decreased the probability of driver error-related crashes. Conversely, an increase in annual average daily traffic (AADT) and the availability of straight and horizontal curve sections increased the probability of driver-related RTCs. The findings support previous studies in similar study contexts that looked at speed dispersion in crash occurrence and severity but disagreed with others that looked at absolute speed at individual vehicle or road segment levels. Thus, the paper contributes to insufficient knowledge of the factors in crash occurrences associated with driver errors on major roads within the context of this case study. Finally, crash prevention and mitigation strategies were recommended regarding the factors involved in RTCs and should be implemented when and where they are needed.

Keywords: traffic crash; driver error; binomial logistic regression; supervised machine learning; random forest (RF); k-nearest neighbor (kNN); Saudi Arabia

Citation: Akin, D.; Sisiopiku, V.P.; Alateah, A.H.; Almonbhi, A.O.; Al-Tholaia, M.M.H.; Al-Sodani, K.A.A. Identifying Causes of Traffic Crashes Associated with Driver Behavior Using Supervised Machine Learning Methods: Case of Highway 15 in Saudi Arabia. *Sustainability* **2022**, *14*, 16654. https://doi.org/10.3390/su142416654

Academic Editors: Jacek Oskarbski, Kyandoghere Kyamakya and Miroslava Mikušová

Received: 31 October 2022
Accepted: 8 December 2022
Published: 12 December 2022

1. Introduction

As road traffic crashes (RTCs) can have devastating impacts on human life, developing a sustainable transportation system is not an option but a requirement for achieving high quality of life and economic prosperity for a nation. Road safety needs a comprehensive plan and actions towards identifying the causes and effects of RTCs and developing necessary remedies to lower the risk of traffic-related injuries and deaths under national and regional plans, which aim to reduce the posed risks using systematic safety audits.

Understanding the causes and impacts of RTCs is critical for developing sustainable road network plans in a country. Analyzing how accidents occur and identifying the underlying factors of certain RTCs may increase the anticipation, prevention, and management of road safety plans and programs [1]. Road safety is a worldwide recognized public health issue obscuring the livelihood of citizens, sustainable development, and economic prosperity. The main findings of the UN Traffic Safety Committee's report show that there has been a significant improvement in road safety in Saudi Arabia. Road traffic fatalities fell by 35.4% between mid–2015 and mid–2019, and the fatality rate per 100,000 population declined by 40% through multiple initiatives under Saudi Arabia's 2030 vision strategic objective. Traffic injury is estimated to cost US$3.2 bn (SAR 12.3 bn) annually—1.7% of GDP. Urban crashes account for 70% of all injury collisions, with 60% of casualties in the 19–40 age group and males 5.6 times more likely than females to be killed in a road traffic crash [2]. Driver and vehicle-related factors, along with nonstandard road geometry, constitute the majority of the causes of RTCs. Driving behavior is a critical factor to consider when determining the causes of traffic accidents [1] because unsafe behaviors, such as texting, drinking, eating, speeding, tailgating or driving while being stressed and exhausted, are among the common factors of RTCs and road fatalities.

Saudi Arabia is a large and economically important country in its region as well as globally. As the largest country with a 2,149,690 km^2 land area (Figure 1) and a road network of 221,372 km [3] in the Gulf Cooperation Council (GCC), Saudi Arabia has reached 34.1 million as of 2021. The country's economic growth is expected to double to 4.9% in 2022, compared to 2.4% in 2021 [4], and its annual population growth was 1.2% in mid-2021 compared to the previous year [5]. Crash records show that traffic safety is an issue of major importance there. The country's mobility within and between the cities mainly depends on highways, leading to deaths and severe injuries, as road traffic injuries (RTIs) are cited as the third leading cause of death in Saudi Arabia [6]. Road traffic crashes (RTCs) have raised public health concerns due to the involvement of novice drivers and the fact that long-distance travel largely depends on private automobiles. Many road users become victims of RTCs and sustain serious injuries or deaths. Fast population growth and special conditions (including the overrepresentation of novice drivers, aggressive driving behaviors, environmental conditions, and so on) are expected to add to the problem. Thus, a need exists to understand the causes of crashes and contributing factors in Saudi Arabia in order to employ proper countermeasures to reduce them in the future.

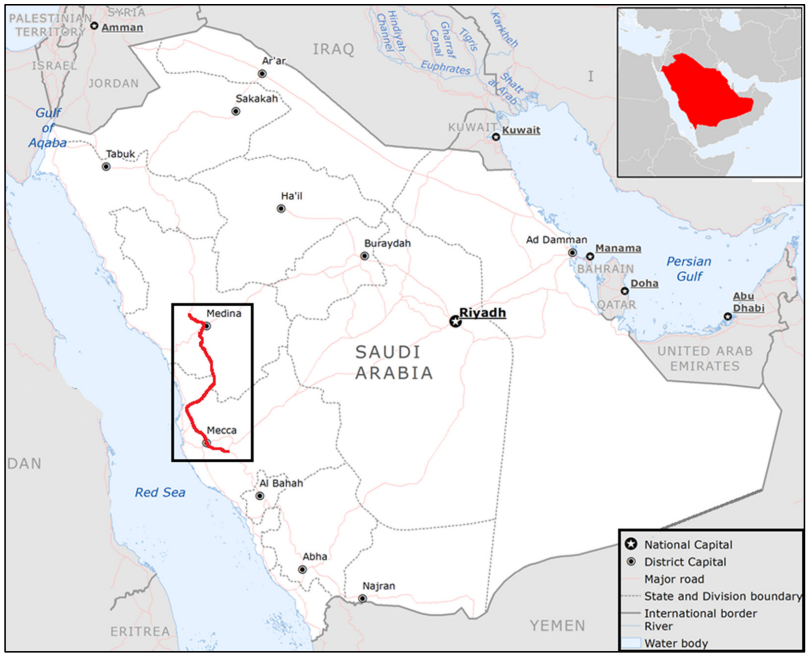

Figure 1. Map of Saudi Arabia with the study area (Highway 15).

This study aims to analyze and develop models to predict the probability of driver errors in the occurrence of RTCs using road features and traffic flow parameters based on historical crash records from 3-year data collected along an access-controlled major highway (761.6 km of Highway 15 between Taif and Medina) in Saudi Arabia. The rest of this paper is organized as follows: First, some background and a literature review are presented. Then, the materials and methods of the study are explained, followed by the results and a discussion of the findings. Finally, crash prevention and mitigation strategies are recommended regarding the factors determined in the occurrence of RTCs associated with driver errors. The paper closes with some concluding remarks and recommendations for future studies.

2. Background and Literature Review

RTCs are responsible for millions of deaths and injuries every year all over the world. Globally, an estimated 3400 people, including more than 500 children, are killed daily in RTCs. Approximately 1.3 million people die annually, and 20–50 million are injured in road crashes. Notably, only 54% of the world's vehicles are in developing countries, but 90% of the world's RTC deaths occur in them [7]. In the U.S., RTCs are a leading cause of death for people aged up to 54 years [8], and they are the leading cause of unnatural death for people [7]. In Saudi Arabia, 5771 persons died, and over 31,745 were injured in car crashes annually in 2018 and 2019 on average, as reported by the Ministry of Health (MOH) [9].

2.1. Causes of RTCs

Generally, driver-related factors (including weary, sleepy, and distracted drivers) and inadequately maintained vehicles are the major causes of RTCs. Among driver-related factors, not using seat belts, drink-driving, speeding, fatigue, and distracted driving are some leading causes of traffic crashes [10–12]. The study [12] identified the distracting activities of young drivers. Many reported frequent use of cell phones while driving and other activities, including adjusting audio devices, chatting with passengers, smoking, eating, and drinking. Their analysis showed that in-vehicle distractions greatly affected the crash likelihood and such dangerous driving behaviors directly increased the crash risk probability.

In the 1970s and 80s in Saudi Arabia, over 60% of RTCs occurred because of drivers traveling at excess speed and disobeying traffic signals. In the 1990s, it was reduced to a little above 40%. In addition, improper overtaking was found to be the highest contributing factor (65%) of all crashes in the 1970s. A comparison of the causes of RTCs between Saudi Arabia and the USA showed that an estimated 80% of crashes occurred due to the human factor. Road geometry or vehicle conditions contributed to only 20% of crashes [13]. Inadequate stopping distance, exceeding the critical speed on a curve, reduction in friction between tires and the road, and a diminished ability of a driver to see and respond to hazards due to being distracted are all factors that increase the likelihood of crash occurrence with increased driving speed [14]. Research shows that straight roads are of concern as they often make drivers fall asleep in front of the wheel [15].

Moreover, most crashes are the result of speeding [6,16–20], which is defined as driving significantly above the speed limit or driving too fast for the prevailing weather, light, traffic, and road conditions but within the speed limit [21–23]. Speed increases crash rates and severity [24], yielding about one-third of the accidents with fatalities [25]. Every km/h increase in speed is associated with an average increase of 3% in crash rate. And driving beyond the speed limit may increase the possibility of a traffic crash leading to injury or death. In the case of roads with a speed limit of 120 km/h, exceeding the speed limit by 1 km/h will increase injury and fatal crashes by about 1.7% and 3.3%, respectively [25,26].

Additionally, the literature findings confirm that geometric design consistency significantly affects the safety of rural motorways. An alignment that requires drivers to handle high-speed gradients and does not meet drivers' expectancy is considered inconsistent and produces higher crash frequency for multi-lane rural highways [16,17,22]. Curved

sections of roadways are more hazardous due to additional centripetal forces acting on a vehicle from the road curvature [27], driver expectations, and other variables. The safety of a horizontal curve is mainly determined by its geometric features [28,29]. Thus, on rural two-lane highways, horizontal curves pose a considerable safety problem, and such locations show higher crash rates than tangent sections. Statistical modeling indicates that more crashes are likely to occur on sharper, narrower curves that do not provide proper spiral transitions and have a higher superelevation deficiency. Higher traffic volumes and longer curves are also linked to more crashes occurring on curves, with all else being equal [30]. Some research quantified the safety effects of horizontal and vertical alignment combinations by developing crash modification factors representing safety performance relative to level tangents, and such modification factors were incorporated into the Highway Safety Manual [31].

In that regard, findings reported in the literature are mixed with respect to speed-crash occurrence causality on rural and urban roads. For example, contrary to the results of most other studies reviewed by Aarts and van Schagen [26], Garber and Gadiraju [32] reported a negative relationship between average speed and crash rate on the interstate, arterial, and major rural collector roads (with ≈90 km/h speed limit) in the USA over a 3-year study. Roads with a larger speed variance of observed vehicles had a higher crash rate than those with a smaller speed variance, and large speed variances in the traffic flow were associated with relatively low average traffic speeds. In conclusion, studies examining speed variance at a road section level found that a larger speed variance was related to a higher crash rate and that high average speeds were related to a low-speed variance. Finally, the exact relationship between speed and crash rate depends on many factors. Even though recent studies shed some light on these factors and the direction of the effects, knowledge is still insufficient to allow for specific quantifications. Researchers must be aware of the influence of external factors on the relationship between speed and crash rate and be explicit and precise about the external circumstances to which their results apply [26]. Thus, analyzing RTCs and their causes can provide valuable insights into identifying major contributors and the level at which they contribute to the probability of the occurrences of some types of crashes. The evaluation and analysis of important contributing factors affecting the number of vehicles involved in crashes play a key role in increasing the efficiency of road safety [33]. In conclusion, the literature review indicated a further need to examine the effect of crash parameters leading to severe crashes associated with driver errors under different environmental and cultural settings.

2.2. Machine Learning (ML) Models in RTC Analyses

Several studies recently explored the influence of risk factors on traffic crashes using machine learning methods [1,33–37]. Rahman et al. [1] established a Bayesian belief network (BBN) model by incorporating an expectation-maximization algorithm to examine the relationships between crash factors with driving behavior in Saudi Arabia in a northern city (Al-Ahsa). The model measured the uncertainty associated with accident outcomes by analyzing intentional and unintentional driving behaviors leading to different types of accidents and accident severities. When considering speeding alone, the results showed a 26% probability that a collision would occur, which was a 63% increase over the initial estimate. Guido et al. [33] employed ML algorithms to determine the number of vehicles involved in an accident. They used several factors affecting transport safety: daylight, weekday, type of accident, location, speed limit, average speed, and annual average daily traffic (AADT) of rural roads in Cosenza, Southern Italy. Their results showed that type of accident was ranked the highest, and the location variable had the lowest importance in their analyses. Farhangi et al. [34] used machine learning algorithms integrated with geographic information systems (GIS) to analyze the factors in the occurrence of RTCs. They employed bagged decision trees (BDTs), extra trees (ETs), and random forests (RFs) in their analyses. They concluded that including traffic volume in modeling could improve the model prediction. Tamakloe and Park [35] analyzed factors influencing the number

of vehicles and the number of casualties involved in fatal crashes at intersections and midblocks. Their time-of-day analysis revealed that large casualties were associated with nighttime at critical intersections. And reckless driving was related to single-vehicle crashes at intersections. Tamakloe et al. [36] employed the Association Rules Mining (ARM) algorithm to discover hidden groups of crash-risk factors leading to different crash severity levels in poor road conditions. They analyzed the crashes under different lighting conditions and determined the effect of factors on the severity of bus/minibus crashes in Ghana. Mirzahossein et al. [37] analyzed the severity of road traffic accidents (RTAs) on rural roads using statistical and intelligent models. Multiple Logistic Regression (MLR) was used to predict the probability of RTAs, and its results were compared with Multi-Layer Perceptron (MLP) and Radius Basis Function (RBF) neural networks to search for more accurate predictions.

Based on the review of the recent literature in the analysis of RTCS, this study aimed to identify the causes of RTCs associated with driver behavior and explore the significance of crash-related factors, including traffic volume. The study employed some popular supervised machine learning (ML) models within the context of a case study on a major highway in Saudi Arabia. Moreover, by identifying the crash-contributing factors leading to driver errors in serious RTCs, researchers and decision-makers can propose and implement crash prevention and mitigation strategies to reduce human and material losses from RTCs, which threaten the quality of life and economic prosperity and sustainability.

3. Materials and Methods

In this study, road traffic crashes (RTCs) were first analyzed using binomial logistic regression (BNLOGREG). Then its results were compared with those of other supervised machine learning (ML) models, i.e., random forest (RF) and k-nearest neighbor (kNN), to search for more accurate predictions, as ML algorithms have a higher prediction power than conventional logistic regression (LR) [38]. The ML methods were trained using the same database as the one used in regression modeling, and their results were compared to BNLOGREG since, for all data types and domains, no classification method is regarded as superior to all others [39].

3.1. Study Site

The study site is the 761.6 km section of Highway 15 between Taif and Madinah, passing through the cities of Makkah and Jeddah (Figure 1). The city of Makkah is one of the oldest continually inhabited cities in the world, located in the southwest of Saudi Arabia, inland from the Red Sea coast. The city underwent vast improvements in the last century to host approximately 3 million visitors during the peak season [40]. Jeddah is a major port city along the Red Sea, west of Makkah, that also serves as a major hub for pilgrims landing and traveling to the holy cities of Makkah and Madinah. Madinah is located about 160 km inland from the Red Sea and 442.5 km north of Mecca by road [41].

3.2. Materials

RTC data were procured from the Ministry of Transportation and Logistic Services (MOTLS) of Saudi Arabia for three years, from 2017–2019. The database included several unique features such as station no., road id, road type, weather and road conditions, number of deaths and injuries, accident types and causes, number of vehicles involved, and road geometry. Secondary data, including speed (speed limit, average, and 85th percentile speeds) and traffic volume (annual average daily traffic-AADT with the percent of heavy vehicle traffic volumes) information, were obtained from permanent traffic recorders (PTR) and were associated with traffic crashes records.

Table 1 presents ten unique cases of RTCs used in the analyses. The file includes 3439 cases, and the sample size is deemed adequate based on the discussions on the sample size requirement in the literature for logistic regression. A larger sample size is needed to estimate parameters involving categorical variables than numerical ones. A

sample size of 300–500 is deemed sufficient to estimate the parameters for the medium population [42]. Several studies recommended the consideration of 500 cases to increase the accuracy of the estimates, and their findings were statistically significant when compared to the parameters of the targeted population [42–45]. Another recommended rule of thumb is based on the rule of event per variable (EPV) of 50 and the formula; n = 100 + EPV × *i*, where *i* refers to the number of independent variables in the final model [44]. In our case, (*i* = 5, n = 100 + 50 × 5 = 350 << 3439) samples already satisfy the minimum sample size.

Table 1. Sample data of RTCs from Highway 15.

Station [1]	Road No.	Road Type [2]	Speed Limit	No. of Lanes	Weather Cond.	Road Cond.	No. of Deaths	No. of Injured	No. of Vehicles	Road Geometry
1568	15	fast	120	3	No rain	Dry	1	10	3	Straight
1917	15	double	110	2	No rain	Dry	0	0	2	Straight
1553	15	fast	120	3	No rain	Dry	0	0	1	Straight
1754	15	fast	110	4	No rain	Dry	0	0	2	Straight
1599	15	fast	120	3	No rain	Dry	0	6	2	Straight
1753	15	fast	110	4	No rain	Dry	0	1	2	Straight
1754	15	fast	110	4	No rain	Dry	0	3	1	Straight
1725	15	fast	120	3	No rain	Dry	0	0	2	Straight
1731	15	fast	120	3	No rain	Dry	0	5	1	Straight
1602	15	fast	120	3	No rain	Dry	0	0	1	Straight

[1] Station: kilometer post; [2] Double: 2 lanes in one direction, and Fast: 3 or more lanes in one direction.

Preprocessing Data

The RTC data (Table 1) were first reviewed and checked to ensure quality. The study database had 3439 cases (rows) and 46 features (columns), and 1.0% of the data had missing values. The following steps were applied for preprocessing data before the modeling step [39]:

1. Getting the data to know: This step studied the various attribute types, which included nominal, binary, ordinal, and numeric attributes. Basic descriptive statistics are used to learn more about each attribute's values. Knowing basic statistics makes it easier to fill in missing values, smooth noisy values, and spot outliers in the data preprocessing stage. Knowing attributes and their values can also help deal with inconsistencies incurred during data integration. Visualization of the RTC data provided information on the trend of the main attributes used in modeling.
2. Checking the completeness: This step was carried out by checking the completeness of the main attributes of crash occurrences, such as the crash type, road and weather conditions, number of casualties and injuries, crash reasons and remarks, and road geometry. Some missing data were imputed with information available from other attributes, but some could not. For example, 943 out of 3439 cases were missing "road geometry" attributes that could not be imputed and coded as 'unknown' or 'other' in the variables used to describe roadway geometry.
3. Imputing missing data: Data with missing values for some attributes are quite common. There are various methods for handling the problem of missing values in data. Here, using the most probable value to fill in the missing value was preferred and determined with decision tree induction using non-missing crash attributes in the data set [46].
4. Normalization: Data normalization gives all attributes an equal weight, where the values are scaled to a smaller range, such as 0.0 to 1.0. Normalization benefits classification algorithms such as neural networks or distance-based models. Normalizing the values for each attribute included in the training set helps speed up the learning phase when using the neural network backpropagation algorithm for classification. For distance-based methods, normalization prevents attributes with large ranges (e.g., AADT, min = 2083 and max = 60,244) from outweighing small-range attributes

(e.g., binary variables). There are several methods used for normalization. The min-max normalization was selected in this study. Min-max normalization transforms a value x of a numeric variable V to x' in the range (0, 1), as shown in Equation (1) below.

$$x' = \frac{x - \min(V)}{\max(V) - \min(V)} \quad (1)$$

3.3. Methods

The methodology part of the study is divided into two parts: (i) an analysis of crash data using descriptive statistics and (ii) the development of models to identify predictors of driver-error-related RTCs.

3.3.1. Analysis of RTCs

The RTC data set used in this study is first analyzed using descriptive statistics and visualization of the relationship between crashes and traffic and road infrastructure attributes. The findings are presented in the results section.

3.3.2. Modeling of RTCs

The ML methods were employed to classify the causes of RTCs into two groups (binomial outcome = (1) driver-related and (0) otherwise) using a set of possible attributes by estimating the chance that an observation belongs to a particular class based on its characteristics. Brief descriptions of the models are given in the following paragraphs. Details of models can be seen in Appendix A.

Binomial logistic regression (BNLOGREG) model: The model belongs to the family of Generalized Linear Models (GLM), which establish a relationship between the conditional expectation of the dependent variable (DV) and a linear combination of independent or explanatory variables (IV) using a suitable link function. The ability of BNLOGREG to provide probabilities and classify new samples using continuous and discrete measurements makes it a popular estimation tool. Predicting the probability of cases belonging to each of the two categories of the dependent variable (DV) is possible using the model's coefficients as well as the possibility of directly calculating the odds ratio [47]. BNLOGREG uses maximum likelihood estimation to evaluate the probability of categorical membership and does not make any assumptions of normality, linearity, or homoscedasticity of variance for independent variables. BNLOGREG necessitates careful consideration of the sample size and examination for outlying cases. Like other data analysis procedures, initial data analysis should be thorough and include careful univariate, bivariate, and multivariate assessments. Specifically, multicollinearity should be evaluated with simple correlations among the IVs. Also, multivariate diagnostics (i.e., standard multiple regression) can be used to assess multivariate outliers and the exclusion of outliers or influential cases.

First, the BNLOGREG model was created because logistic regression models are mostly used for data analysis and inference, where the objective is to understand the role of the input variables in explaining the outcome [48]. The analysis requires that the dependent variable be non-metric (i.e., dichotomous variable) to satisfy the level of measurement required. In this study, we decided not to apply the normalization to BNLOGREG variables because the prediction power was not affected by normalization. In addition, in BNLOGREG logistic regression, coefficients of variables indicate the effect of a one-unit change in DV on the log odds of "success." Transforming a variable by normalization changes the "unit" of the variable in the model context. Raw data for IVs vary across different units in the original metric. After the normalization, the data ranged from 0 to 1, i.e., changing one unit now means going from the lowest to the highest-valued observation. Lastly, the normalization did not affect the log odds of success change.

Random Forests (RF): RFs were introduced by Breiman [49] and further developed by Breiman and Cutler [50]. Random Forest builds a set of decision trees. Each tree is developed from a bootstrap sample from the training data. When developing individual

trees, an arbitrary subset of attributes is drawn, from which the best attribute for the split is selected. The final model is based on the majority vote from individually developed trees in the forest. RF works for both classification and regression tasks, and the response and predictor variables can be categorical or continuous [46]. Random Forests are computationally and statistically appealing, as well [51].

A combination of tree predictors known as a "random forest" depends on each tree being dependent on values from a random vector sampled randomly and uniformly across all trees in the forest. As a forest's tree population grows, the generalization error for forests constricts to a certain size. The strength of each individual tree in the forest and the correlation between them determine how accurate a forest of tree classifiers is at generalizing their results. Each node is split using a random selection of features, and the resulting error rates are better than Adaboost but more noise-resistant. Internal estimates track inaccuracy, strength, and correlation, which are used to demonstrate how the splitting process responds to an increase in the number of features. To gauge the significance of a variable, internal estimations are also used. Regression can benefit from these concepts [49].

Some random forests have consistently lower generalization errors than others, as described in the literature. For example, the random split selection is superior to bagging [52]. But none of these forests perform as well as Adaboost [53] or other algorithms that reweight the training set adaptively (arcing) [52–55]. To increase the precision, the injected randomness must decrease correlation while retaining strength. At each node of the examined forests, random inputs or input combinations are used to cultivate each tree. The generated forests have accuracy comparable to that of Adaboost. This class of procedures has many desirable characteristics: (i) RF's accuracy is comparable to Adaboost and sometimes better, (ii) RF is relatively robust to outliers and noise, (iii) RF is faster than bagging or boosting, (iv) RF provides useful internal estimates of error, strength, correlation, and variable importance, and (v) RF is simple and easily parallelizable [49].

Random forests are a significant variation of bagging in which many de-correlated trees are constructed and then averaged. On many tasks, random forests perform comparably to boosting, making them easier to train and modify [48]. Forests generated randomly are instances of ensemble methods. Imagine that each classifier in the ensemble is a decision tree classifier, such that the ensemble is a "forest" of classifiers. A random selection of qualities at each node determines the split when generating the individual decision trees. Formally, each tree depends on the values of an independently sampled random vector with the same distribution across the entire forest. Each tree votes during classification, and the most popular class is returned [39].

Random forests can be constructed by combining bagging with random feature selection. Each new training set is drawn from the original training set, with replacement. Then, a tree is constructed using random feature selection on the new training set. The cultivated trees are not pruned. There are two reasons why bagging is used. First, bagging appears to improve accuracy when random features are employed. Second, bagging can be used to provide ongoing estimates of the generalization error of the combined ensemble of trees, as well as estimates for the strength and correlation between the trees [49]. A training set of d data, D, is provided. The following is the general process for generating k decision trees for the ensemble. Each iteration, i (i = 1, 2,..., k), samples with replacement a training set, D_i, of d data from D. In other words, each D_i is a bootstrap sample of D, and so some data may appear multiple times in D_i, while others may be omitted. Let F be the number of attributes utilized to determine the split at each node, where F is significantly less than the total number of attributes. To create the decision tree classifier, M_i, randomly selects F attributes at each node as candidates for the node's split. The CART method is utilized to cultivate the trees. The trees are grown to their maximum size without pruning. This method of generating random forests with random input selection is called Forest-RI [39].

Forest-RC is a random forest variant that uses linear combinations of input attributes. Instead of selecting a subset of attributes randomly, it develops new attributes (or features) that are linear combinations of the current attributes. Thus, L, the number of original

attributes to be combined, is used to construct an attribute. At a particular node, L characteristics are selected randomly and inserted with coefficients consisting of uniform random values on the interval $(-1, 1)$. F-linear combinations are created, after which the optimal split is determined. This random forest variant is beneficial for reducing the correlation between individual classifiers when few attributes are available [39].

Random forests are comparable to AdaBoost in terms of accuracy but are more tolerant of errors and outliers. As long as there are a large number of trees in a forest, the generalization error will converge. Therefore, overfitting is not a concern. The precision of a random forest is determined by the quality of the individual classifiers and a measure of their interdependence. Maintaining the effectiveness of separate classifiers without increasing their correlation is optimal. The number of attributes selected for consideration at each split has no effect on random forests. Usually, up to $log_2 d + 1$ are selected. An intriguing empirical observation was that employing a single random input attribute can often result in greater accuracy than using multiple qualities. Random forests are effective on very big databases because they consider far fewer attributes for each split. They are sometimes faster than bagging or boosting. Random forests provide varied internal assessments of importance [39]. For the details of the algorithm, the reader is referred to Appendix A and other sources for further information on RFs.

k-nearest neighbor (kNN): kNN, a non-parametric supervised learning method, was first developed by Fix and Hodges [56] and later expanded by Cover and Hart [57]. It can be used for both classification and regression. In both cases, the input consists of the k closest training examples in a data set. Then, the output depends on whether it is used for classification or regression. In the classification problem, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor. The function is approximated only locally in the classification problem, and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, normalizing the training data can improve its accuracy significantly if the features represent different physical units with different scales. Both for classification and regression, a useful technique assigns weights to the contributions of the neighbors so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme gives each neighbor a weight of $1/d$, where d is the distance to the neighbor. The neighbors are taken from a set of objects for which the class is known. This can be thought of as the training set for the algorithm, though no explicit training step is required [48,58]. Appendix A includes the details of the method.

4. Results and Discussion

Results and discussion of the analyses and modeling of the crash data are presented in the following sections.

4.1. Descriptive Statistics and Visualization of RTC Factors

As the summary given in Table 2, three major factors are determined in the occurrence of RTCs: (i) driver-related, (ii) vehicle-related, and (iii) road and traffic conditions-related. Notably, driver-related factors and vehicle-related failures accounted for 74.7 and 14.2 percent of the total road crashes during the three-year period (2017–2019) in the case site. The effects of road and traffic conditions (object on the road, congestion, and road failure) and weather-related factors (low visibility due to sandstorms, rain/wet pavement, and extreme heat) on RTCs were ignorable, with only a meager percentage attributed to crashes (0.8%). The southwest and western regions (including Makkah, Jeddah, and Madinah) of Saudi Arabia exhibit a semi-arid climate, contrary to other northern hemisphere countries receiving substantial snow and heavy rain. Thus, only 22.4 rainfall days throughout the year yield 111.8 mm (<150 mm in most parts of the country) of precipita-

tion on average [59], except in the southwestern part, where the rainfall occurs between 400–600 mm annually [60]. The RTC data includes around 8.1% of crashes with unknown or unreported causes. Most of the crashes are driver related. In driver-related crashes, speeding, distracted driving/loss of control, reckless driving, and driver sleepiness were found to be the main contributors, with 29.9, 23.7, 10.3, and 7.9 percent, respectively. The analysis revealed that the deadliest crash contributing factor was drivers falling asleep at the wheel, with a rate of 0.195 deaths per crash, followed by distracted driving/loss of control, with a rate of 0.162. Driver's sleeping is also the highest injury-causing type of crash contributing factor, at a rate of 1.044 per crash. Among other factors, crashes caused by uncontrolled animal crossings were only 2.2 percent of all crashes (See Table 2).

Table 2. Summary of the RTCs (2017–2019).

Crash Causes	Crashes		Deaths		Injuries	
	Number	Percent	Number	Percent	Number	Percent
Driver-related crashes						
Speeding	1028	29.9	102	0.099	933	0.908
Distracted driving/loss of control	816	23.7	132	0.162	810	0.993
Reckless driving	353	10.3	29	0.082	341	0.966
Driver asleep	272	7.9	53	0.195	284	1.044
Other	99	2.9	15	0.269	88	1.129
Subtotal	2568	74.7	331	0.129	2456	0.956
Vehicle-related crashes						
Tire-blowout	361	10.5	51	0.141	343	0.950
Mechanical/electrical malfunction	114	3.3	10	0.088	56	0.491
Overloading/misloading	12	0.3	0	0.000	14	1.167
Subtotal	487	14.2	61	0.125	413	0.848
Other factors						
Road/traffic conditions-related	20	0.6	0	0.000	16	0.800
Weather-related	10	0.3	1	0.100	9	0.900
Animal crossing	74	2.2	10	0.135	53	0.716
Other/undetermined	280	8.0	27	0.096	270	0.964
Subtotal	384	11.1	38	0.099	348	0.906

The share of driver errors as the highest contributing factor clearly reveals that RTCs are the number one public safety concern, and they must be addressed diligently to develop mitigation and prevention strategies. Vehicle-related factors, mainly tire blowouts, mechanical or electrical failures of vehicles, and overloading/misloading, account for 14.2 percent of all cases. Driver-related and vehicle-related crashes have similar casualty rates (0.129 and 0.125 deaths per crash, respectively). An in-depth look at driver errors in RTCs shows that the top four crash causes contribute to 71.8% of all 3439 cases and are listed as speeding (29.9%), distracted driving/loss of control (23.7%), reckless driving (10.3%), and driver asleep at the wheel (7.9%). These findings also indicate that human error is the number one contributor to RTCs. In vehicle-related crashes, tire blowouts attract attention as they contribute to 0.141 deaths per crash and 0.950 injuries per crash.

Descriptive statistics of RTCs, traffic flow characteristics, and road geometry attributes used in the classification models are presented in Table 3. As seen, crashes caused by driver errors (0.1288 casualties per crash) are 13.1% deadlier than crashes caused by other factors (0.1139 casualties per crash). Similarly, crashes caused by driver errors (0.9911 injuries per crash) result in 22.2% more injuries than crashes caused by other factors (0.7710 injuries per crash). Overall, the number of RTCs was reduced by 19.9% from 2017 to 2019, and deaths and injuries by 6%, and 16.9%, respectively, for the road segment in this case study. This observation is aligned with the national statistics reported by WHO [2].

Table 3. Summary statistics of the variables for modeling (2017–2019).

Variable	Mean	Std. Dev.	Min.	Max.	Sum (No. of Cases) [2]
RTCs by causes and consequences					
Annual number of RTCs	1146.33	141.11	1048	1308	3439
RTCs caused by driver error	856.67	32.52	820	882	2570
RTCs caused by other factors	289.67	118.25	215	426	869
Annual number of total casualties	143.33	7.09	137	151	430
Casualties due to driver-errors	110.33	13.61	115	121	331
Casualties due to other factors	33	19.92	21	56	99
Number of casualties per all crashes	0.1258	0.01	0.1154	0.1355	0.1250
Casualties per driver-error crashes	0.1293	0.02	0.1077	0.1476	0.1288
Casualties per other factor crashes	0.1086	0.02	0.0921	0.1315	0.1139
Annual number of total injuries	1072.33	100.27	980	1179	3217
Injuries due to driver-errors	1018.33	131.98	855	1017	2800
Injuries due to other factors	139	20.07	125	162	417
Number of injuries per all crashes	0.9378	0.0378	0.9014	0.9769	0.9354
Injuries per driver-error crashes	0.9919	0.0739	0.9093	1.0518	0.9911
Injuries per other factor crashes	0.7362	0.1295	0.6491	0.8850	0.7710
Traffic flow characteristics					
Annual average daily traffic (AADT) [1]	13,756.08	14,377.92	2083	60,244	(3327) [2]
Daily average speed of traffic flow (ASF) in kph [1]	96.77	8.86	70.5	116	(3327) [2]
85th percentile speed in kph	117.56	10.17	101	151	(3320) [2]
Road geometry characteristics					
Number of lanes (NL) [1]	2.93	0.65	2	4	(3435) [2]
Speed limit (km/h)	116.91	4.63	100	120	(3435) [2]
Categorical variables					
Causes of RTCs (DV)	1 = driver-error (count = 2570)				(2852) [2]
	0 = otherwise (count = 869)				(587) [2]
RG1: Straight segment [1]	1 = straight (count = 2438)				(2438) [2]
	0 = otherwise (count = 1001)				(1001) [2]
RG2: Horizontal curve [1]	1 = horizontal curve (count =39)				(3400) [2]
	0 = otherwise (count = 3400)				(39) [2]

[1] IVs, [2] No. of cases.

In the data set, some IVs have missing values, such as AADT and ASF (112 cases) and the number of lanes (4 cases). These missing values are to be imputed as described in the preceding sections. The continuous IVs (i.e., ASF and AADT) were investigated against outliers. The minimum and maximum values, first and third quartiles, interquartile range, and z-scores are given in Table 4. Figure 2 displays boxplots and marks outliers for continuous IVs considered in this study (the numbers on the graphs are case numbers marked with stars). When we look at them closely, for ASF, although there are 43 cases with z-score > 2.698 (Table 4), none of them are marked as outliers because, for them, ASF = 70.50 km/h, which cannot be considered as an outlier, corresponding to the minimum value of the variable. For AADT, there are 230 cases with z-score > 2.68, but they are considered legitimate outliers ranging between 55,443 and 60,244 veh/day. So, they are not excluded from the analyses.

Figure 3 depicts RTC frequency vs. AADT. The distribution follows a normal distribution, skewed to the right. In all intervals of AADT, RTCs associated with driver errors dominate the other factors. For AADT ranging between 5000 and 10,000, RTCs associated with driver errors increased by 2.5 folds compared to AADT of less than 5000. However, the same rate was 1.3 for the RTCs associated with other factors. The highest number of RTCs (1287 driver-error related, 370 other factors) was observed when AADT was between 5000–10,000, then continued to decline while AADT was increasing.

Table 4. Descriptive statistics of the continuous IVs (AADT and ASF).

Variable	Q1	Q3	Q3 − Q1	z-Scores [1]			
				Min.	No. of Cases	Max.	No. of Cases
ASF	88.5	102.2	13.7	−2.9658 > −2.698 *	43	2.1715 < 2.698 *	N/A [2]
AADT	6794	12,904	6110	−0.8119 < −2.698 *	N/A [1]	3.2333 > 2.698 *	230

[1] z (Q1 or Q3) ± 1.5 IQR = ±0.6745 ± (1.5 × 1.349) = ± 2.698 [61]; [2] N/A: not applicable; *: Far (true) outliers.

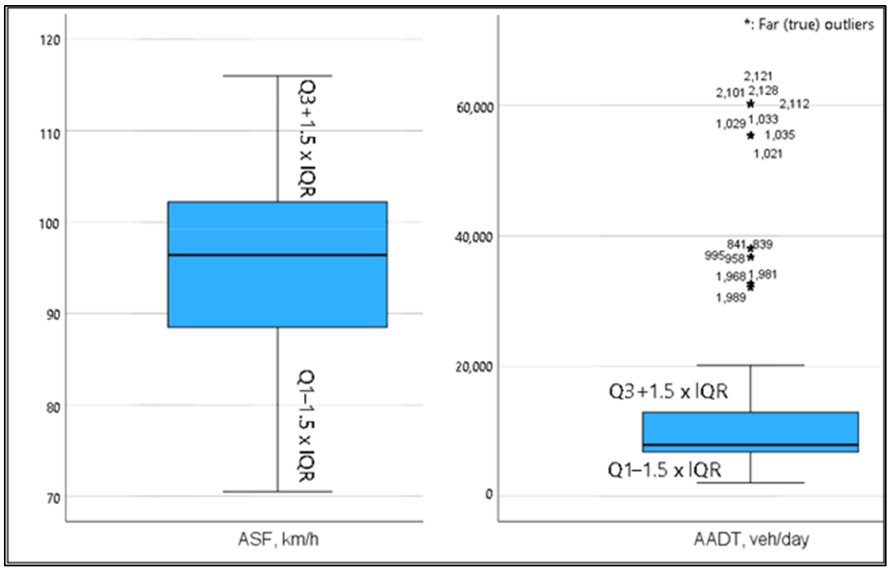

Figure 2. Q1, Q3, IQR, and outliers for ASF and AADT.

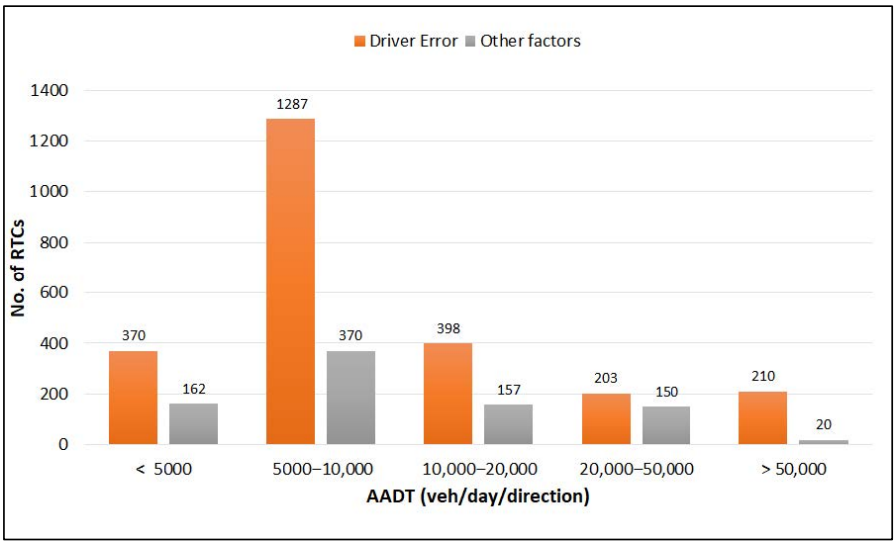

Figure 3. RTC vs. AADT.

Figure 4 shows that in four-lane (per direction) segments, driver-error-associated RTCs were considerably reduced compared to two- and three-lane segments by 27% (=(645 − 471)/645) and 67.5% (=(1451 − 471)/1451), respectively. Similar reductions were observed for other factor-related crashes (29.2% and 74.2% for N = 2 and 3 lanes, respectively). RTCs also substantially decreased by increasing the 85th percentile speed to 120 km/h (Figure 5). Speeds beyond 120 kph are illegal because that was the legal limit for the time period studied, so the RTCs associated with driving at such high speeds are not representative of legitimate behaviors in traffic. 1371 (72.3%) out of 1895 crashes

associated with driver errors occurred at speeds higher than 120 km/h. Driving beyond the speed limit is considered dangerous because it may increase the possibility of a traffic crash leading to injury or death. Exceeding the speed limit on 120 km/h speed limit roads by 1 km/h will increase injury and fatal crashes by about 1.7% and 3.3%, respectively [25,26]. As shown in Figure 6, when road geometry is considered, the dominant cause of RTCs is driver error, contributing to 79.0% of crashes occurring on straight road segments and 84.6% on horizontal curve segments. For non-straight and horizontal curve segments, the RTCs associated with driver errors were higher than half of the crashes (64.4 and 74.6%).

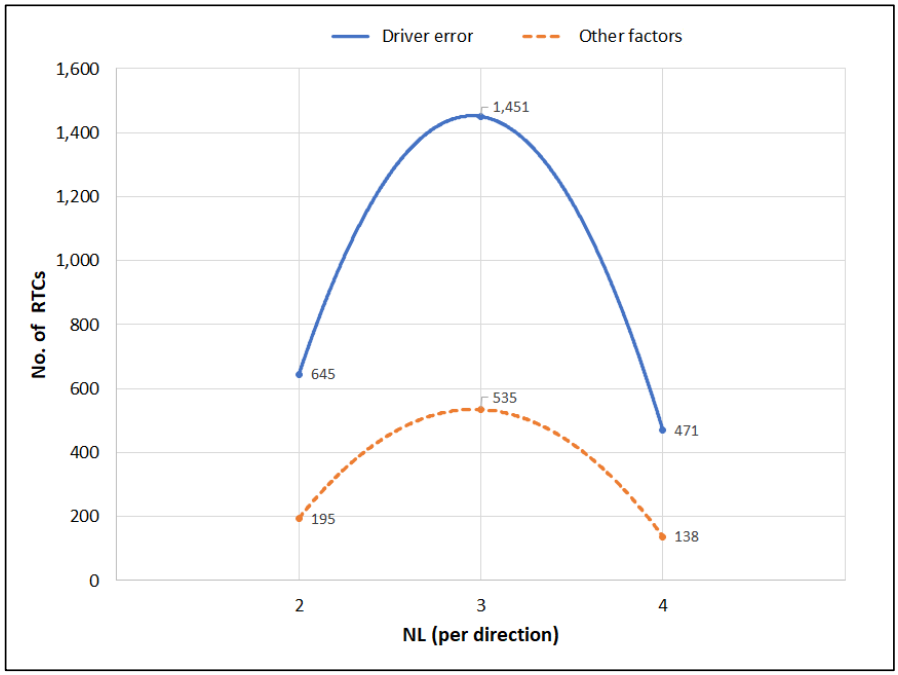

Figure 4. RTC vs. Number of Lanes (NL).

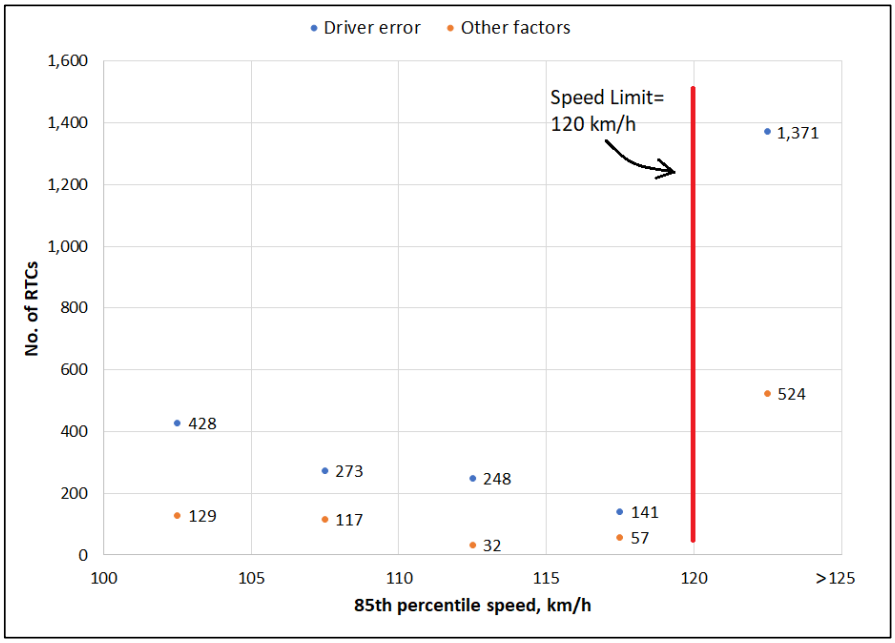

Figure 5. RTCs vs. 85th percentile speed.

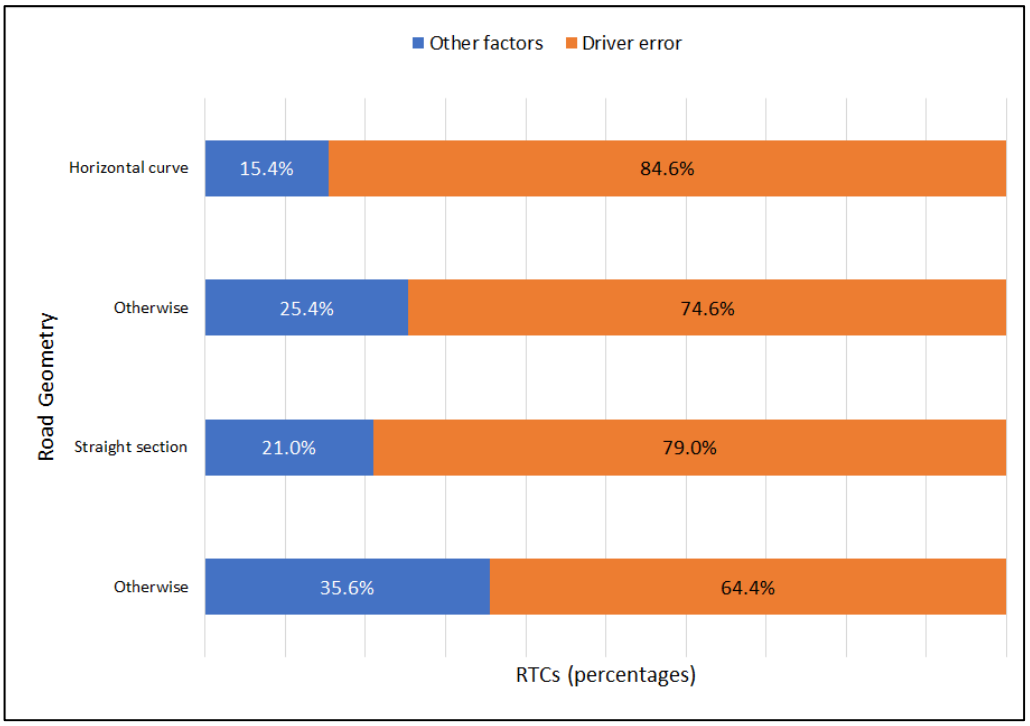

Figure 6. Road Geometry vs. RTC (%) by causing factor.

4.2. Modeling Results

The results of the classification analyses are presented and discussed here. The type of algorithm employed for classification plays a great role in affecting accuracy. BNLOGREG and other ML methods were compared since ML has received great attention due to its robustness in classification problems. The use of supervised ML algorithms in various applications can be found in the literature [1,62–64].

4.2.1. Results of BNLOGREG

The outputs are presented and discussed as follows. First, the case possessing summary is presented in Table 5. Due to the incompleteness of some attributes, there were 112 missing cases out of 3439, constituting 3.3% of all cases. Cases with missing values of AADT and ASF attributes were excluded from the analysis; thus, 3327 cases were analyzed. Categorical independent variables (RG1: Straight section, and RG2: Horizontal curve) are listed in Table 6. RTCs occurring on straight sections represent 70.5% of the cases, and horizontal curve-related RTCs are just 1.2%.

Table 5. BNLOGREG case processing summary.

Unweighted Cases		N	Percent
Selected Cases	Included in Analysis	3327	96.7
	Missing Cases	112	3.3
	Total	3439	100.0
Unselected cases		0	0.0
Total		3439	100.0

Table 6. Summary of categorical independent variables.

Unweighted Cases		N	Percent
RG1: Straight Section	Straight section (1)	2344	70.5
	Other (0)	983	29.5
RG2: Horizontal Curve	Horizontal curve (1)	39	1.2
	Other (0)	3288	98.8

Table 7 presents Model 0 (without IVs) and the prediction power of the model (74.2 percent) when IVs are not included but the constant. According to Wald statistics, the constant is significant at the level of 0.01 ($p < 0.001$). To measure how well the model performs, omnibus tests are performed. The chi-square statistic (Table 8), which represents the change in the −2 log-likelihood between Models 0 and 1, is significant (Chi-sqr. = 133.834, $p < 0.001$). The method used is stepwise forward LR (Likelihood Ratio) regression. The change from Model 0, where no variables are entered, makes a significant improvement in Model 1 (after all five IVs are included) because the significance of the change is small (sig. = 0.001 < 0.05), and the prediction power of the model is improved from 74.2 to 74.5 percent (Table 9).

Table 7. Prediction power of Model 0 (without IVs).

			Variable in the Equation					
Model			**B**	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**
Model 0	Constant [1]		1.055	0.040	709.766	1	<0.001	2.873
					Predicted			
					CRTC		Percentage	
	Observed				Otherwise = 0	Driver error = 1	Correct [2]	
Model 0	CRTC	Otherwise = 0			0	859	0.0	
		Driver error = 1			0	2468	100.0	
	Overall percentage						74.2	

[1] Constant is included in the model; [2] The cut value is 0.500.

Table 8. Omnibus Tests of Model Coefficients (Model 0 vs. Model 1).

Model		**Chi-Square**	**df**	**Sig.**
Step 5	Model 0	10.109	1	0.001
	Model 1	133.384	5	<0.001

Table 9. Prediction power of Model 1 (IVs included).

			Predicted		
			CRTC		**Percentage**
	Observed		Otherwise = 0	Driver Error = 1	Correct [2]
Model 1 [1]	CRTC	Otherwise = 0	16	843	1.9
		Driver error = 1	7	2461	99.7
	Overall percentage/accuracy rate				74.5

[1] Constant and five IVs included in the model, [2] The cut value is 0.500.

Though the correlation measures to estimate the strength of the relationship (pseudo-R square measures) are calculated as Cox and Snell = 0.039, and Nagelkerke = 0.058, these do not really tell much about the accuracy or errors associated with the model. The model summarizes the difference in the probability of driver-related errors for the causes of RTCs (CRTC) that are non-driver-related and CRTCs that are not. Non-metric IVs (RG1 and RG2) are included as a "factor," and dummy-coded (straight road section = 1, otherwise = 0, and horizontal curve = 1, otherwise = 0), and continuous IVs (NL, ASF, and AADT) are included as "covariates.". The estimated constant and the coefficients of IVs (NL, ASF, AADT, RG1, and RG2) are given in Table 10. The constant and all IVs are significant at a 0.01 level according to the Wald chi-square test (all $p < 0.001$).

Table 10. Coefficient of variables in BNLOGREG model.

Coefficient	$\hat{\beta}$	Std. Error	Wald	df	Sig.	$e^{\hat{\beta}}$
Constant	4.375	0.622	49.495	1	<0.001	79.463
NL	−0.501	0.101	24.782	1	<0.001	0.606
ASF	−0.027	0.005	28.831	1	<0.001	0.973
AADT. 10^{-3}	0.016	0.005	12.580	1	<0.001	1.017
RG1 (Straight = 1)	0.824	0.086	92.068	1	<0.001	2.280
RG2 (Horz.curve = 1)	1.268	0.451	7.904	1	0.005	3.555

The outputs are presented and discussed as follows. The intercept $\hat{\beta_0}$ is the odds of a driver-related CRTC if it is driver-related. The weight $\hat{\beta_4}$ = 0.824 is the change of RG1 (straight section) in the log-odds ratio for a driver-related CRTC relative to a non-driver-related CRTC. The exponentiated value of $e^{0.824} = 2.280$ indicates that, on average, the odds that the CRTC will be 2.280 times driver-related if the RG1 is a straight section. This output is aligned with earlier findings confirming that straight road sections are one of several factors that can cause drivers to fall asleep while driving [26]. Similarly, horizontal curve (HC) sections might increase the CRTC to be driver related, as the odds ratio of the driver error is 3.555 times more for HC sections than for non-horizontal curve sections. AADT's effect on the increase in driver error is relatively small (0.016 per 1000 vehicles). On the other hand, one unit increase in NL and ASF will decrease the odds ratio of CRTC by 0.501 and 0.027, respectively. It is intuitive that the increase in the NL and ASF relieves road congestion and decreases driver-related errors in RTCs. The data visualization also supports this conclusion (Figures 4 and 5). Moreover, Elvik et al. [65] examined a more disaggregated level and reported that large traffic flows reduce speed and increase the crash risk [26].

Similar conclusions concerning the travel speed and the occurrence of RTCs are reported in the literature. As the average speed of traffic flow increases, the number of vehicle collisions may indeed be reduced; however, the injuries caused to pedestrians increase significantly [66]. The data from a federal report [67] also indicated that accident rates were reduced at sites where speed limits were raised and increased at sites where speed limits were lowered. Before and after data were collected simultaneously at comparison sites where speed limits were not changed to control the time trends. Repeated measurements of speed limit changes were observed at 14 locations to examine the short and long-term effects. The study showed that lowering speed limits by up to 20 miles per hour (32 km per hour) or increasing them by up to 15 miles per hour (24 km per hour) had little influence on the speed of motorists. The majority of drivers did not increase or decrease their speed by 5 or 10 miles per hour (8 or 16 km per hour) when posted speed restrictions were increased or decreased. According to data collected at the research locations, the majority of speed restrictions were set below the average traffic speed. Thus, dropping speed restrictions below the 50th percentile had no effect on accident rates but considerably impacted speeding. In contrast, increasing the listed speed limit did not increase speeds or collisions [67].

Using the cross-tabulation of observed and predicted values of CRTC in Table 9, sensitivity, specificity, false positive (FP) rate, and false negative (FN) rate are calculated. The percent occurrences vary with different cut-off values (here, it is 0.5), which may not be a good representation of goodness of fit unless an optimum value is determined. The accuracy assesses how precisely a model can predict the outcome. CRTC equal to zero was observed and predicted 16 times, though it was observed and predicted to be equal to one 2461 times. Therefore, the accuracy rate is calculated as 16 plus 2461, divided by the total valid sample size of 3327, equal to 74.5 percent. The misclassification rate is the percentage of observations predicted incorrectly. In the classification with BNLOGREG described above, the misclassification rate is calculated as (7 + 843) divided by 3327, resulting in 25.5 percent. The model needs improvement because it has a serious misclassification. After all, 843 crashes caused by non-driver errors were classified as

caused by driver-related factors. This error is called Type I, and the values are called False Positive (FP). Sensitivity or recall (percent of occurrences correctly predicted), specificity (percent of non-occurrences correctly predicted), false positive (FP) rate (percent of non-occurrences incorrectly predicted), and false negative (FN) rate (percent of occurrences incorrectly predicted) are used for observations in the classification table. The model's sensitivity is the ratio of occurrences or events correctly predicted. It is the probability that the predicted value of DV is equal to one, given the observed value of DV being one (2461/(7 + 2461) = 99.7%). On the other hand, specificity is the percentage of non-occurrences being correctly predicted, that is, the probability that the predicted value of DV is zero, given that the observed value of DV is also zero (16/(16 + 843) = 1.9%). The false positive (FP) rate is the percentage of non-occurrences that are mispredicted events (1 − Specificity = 1 − 1.9% = 98.1%). Similarly, the false negative (FN) rate is the percentage of occurrences that are predicted incorrectly (1 − Sensitivity = 1 − 99.7% = 0.3%). For different cut-off values, the model's predictive power is presented in Table 11. Additionally, Youden Index (Sensitivity + Specificity − 1) is proposed to select the best cut-off value, considering both true positive and true negative rates [68]. Based on the index, the best cut-off value for the model is between 0.58 and 0.59, as it yields the highest accuracy (74.8%).

Table 11. BNLOGREG predictive power for different cut-off values.

Cut-Off Value	Observed		Predicted		Percent Correct	Accuracy (%)	Sensitivity (TPR) (%)	Specificity (1-FPR) (%)	Youden Index
			0	1					
0.60	CRTC	0	127	732	14.8	74.5 [1]	95.3 [1]	14.8 [1]	10.1 [1]
		1	116	2352	95.3				
0.59/ 0.58	CRTC	0	124	735	14.4	74.8 ↑	95.8 ↑	14.4 ↓	**10.2** ↑
		1	104	2364	95.8				
0.55	CRTC	0	84	775	9.8	74.8 ↑	97.4 ↑	9.8 ↓	7.2 ↓
		1	65	2403	97.4				
0.50	CRTC	0	16	843	1.9	74.5 ↔	99.7 ↑	1.9 ↓	1.6 ↓
		1	7	2461	99.7				

[1]: base value, ↑: increase, ↓: decrease, and ↔: no change.

Variables not in the model are shown in Table 12. According to the Wald test used to predict whether an independent variable would be significant in the model, all *p*-values are higher than 0.05, so they are concluded as statistically not significant. With its coexistence of ASF in the model, the "speed limit" variable becomes statistically insignificant with a high correlation (Wald = 0.889, sig. = 0.343 > 0.05). Also, with the coexistence of the 85th percentile speed, AADT was dropped from the model (Wald = 3.397, sig. = 0.065 > 0.05) in our trials. Thus, it was decided to include ASF and AADT in the model because they are the fundamental indicators of highway traffic flow.

Table 12. Coefficient of variables not in BNLOGREG equation.

Coefficient	Wald	df	Sig.
Accident hour	20.477	23	0.613 > 0.05
Weekday	5.187	6	0.520 > 0.05
Month	5.271	11	0.219 > 0.05
Weather	3.853	5	0.247 > 0.05
Speed limit	0.899	1	0.343 > 0.05

The multicollinearity matrix is presented in Table 13. As seen, only a medium correlation (0.658 < 0.70) is available between NL and AADT. 10^{-3}. Other variables show no significant correlations at all. Multicollinearity can also be identified using the variance inflation factor (VIF), which evaluates the correlation and correlation strength between the predictor variables in a regression model. One is the least possible value for VIF, which implies no association among the variables in the model. A value between 1 and

5 is typically not severe enough to warrant special attention. VIFs larger than 5 indicate significant collinearity levels where the coefficient estimations cannot be relied upon, and the statistical significance is uncertain. The VIF for each variable can be calculated using the following formula as shown in Equation (2) [69]:

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j \,|X_{-j}}} \tag{2}$$

where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors. If $R^2_{X_j|X_{-j}}$ is close to one, collinearity is present so that the VIF will be large. In the BNLOGREG model, all VIFs are much smaller than 5, with only two values higher than 2.0.

Table 13. Multicollinearity matrix and VIF values of the variables in the model.

Variables	VIF	NL	ASF	AADT. 10^{-3}	RG1 = 1	RG2 = 1
NL	2.273	1.000				
ASF	1.349	0.245	1.000			
AADT. 10^{-3}	2.148	−0.658	0.133	1.000		
RG1 (Straight = 1)	1.049	−0.012	−0.132	0.023	1.000	
RG2 (Horz.crv = 1)	1.031	−0.036	−0.021	0.019	0.121	1.000

4.2.2. Comparison of BNLOGREG with Machine-Learning Algorithms

Orange software (an open-source data visualization, machine learning, and data mining toolkit) was used to perform the comparison analysis [46] with different supervised-learning classification models. Figure 7 presents the workflow of the modeling process. After importing the data file, data imputing and sampler steps are inserted in case missing values need to be imputed to avoid misclassifications. The data may be divided into two sets: training and testing. However, at first, both of the two steps are skipped. So, the data included 3439 cases and 45 attributes with 0.9% missing data. Later on, two more alternative data sets were created in addition to using the whole data: (i) cases with missing attributes were removed, and only 69.44% of data was used for model training, and (ii) secondly, using model-based imputer (simple tree), 100% of data is used to train models. The IVs of NL, ASF, AADT. 10^{-3}, RG1 (Straight = 1), and RG2 (Horz.curve = 1) were employed to determine the probability of CRTC whether or not it is driver-error related using the BNLOGREG and other supervised machine-learning classifiers, such as random forest (RF) and k-nearest neighbor (kNN). The models learned from the training data set and tested using 10-fold cross-validation.

The performance of the models, such as the area under the ROC curve (AUC), classification accuracy (CA and balanced CA), the harmonic mean of the precision and recall (F1), Precision, and LogLoss (a negative average of the log of corrected predicted probabilities for each instance) is given in Table 14. LogLoss is considered the most important classification metric when the data is imbalanced against one of the classes. For any given problem, a lower value means better predictions. Unlike accuracy, LogLoss is robust in the presence of imbalanced classes. It is observed that ML algorithms have a higher prediction power than conventional logistic regression [38]. The metrics confirmed this result, where RF yielded the highest classification accuracy (CA). BNLOGREG got third place in terms of accuracy. However, accuracy alone may not be a good metric for unbalanced data (CRTC = 1 cases are 74.7%), where 843 cases were classified as false-positive (FP) by BNLOGREG. In this case, balanced accuracy (BAC = (TPR + TNR)/2) is recommended as a better metric (highest BAC = 0.593 yielded by RF). F1 also shows how precise and robust the model is; however, the biased "recall" values affect the accuracy of F1.

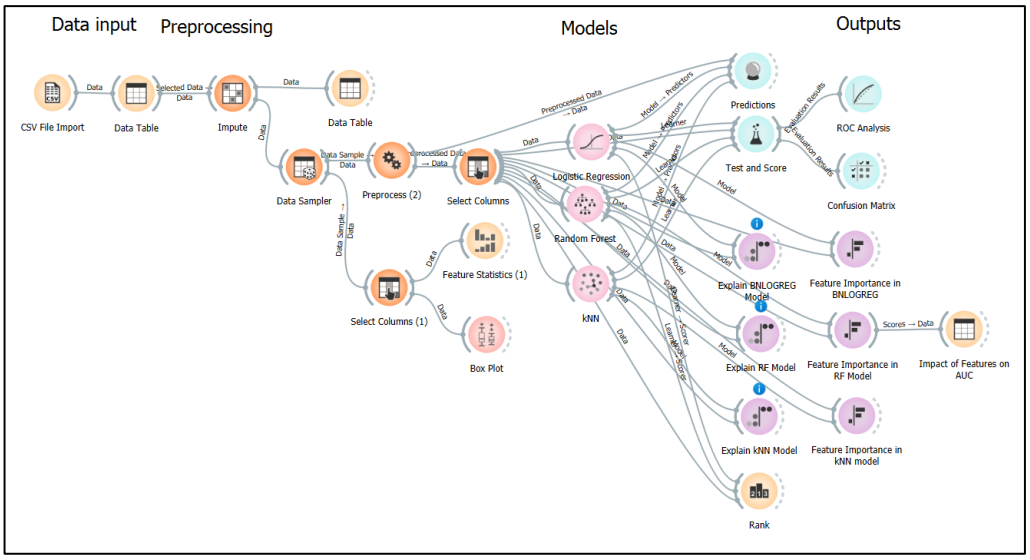

Figure 7. Flowchart of the model comparison.

Table 14. Performance of classification models with alternative data sets.

Model	AUC	Accuracy (CA)	Balanced Accuracy (BAC)	F1 [1]	Precision	LogLoss
Dataset (1): 3439 cases (no data imputing) and models were tested using 10-fold cross-validation.						
RF [2]	0.701	0.760	0.593	0.853	0.787	0.512
kNN [3]	0.612	0.745	0.600	0.839	0.792	4.557
BNLOGREG [4]	0.607	0.747	0.500	0.855	0.747	0.548
Dataset (2): Cases with missing values removed (69.44% data used) and models tested using 10-fold cross-validation.						
RF [2]	0.653	0.787	0.500	0.881	0.787	0.490
kNN [3]	0.595	0.724	0.556	0.829	0.809	2.365
BNLOGREG [4]	0.554	0.787	0.500	0.881	0.787	0.510
Dataset (3): Model-based imputer (simple tree) used to replace missing values (100% data used), and models tested using 10-fold cross-validation.						
RF [2]	0.712	0.762	0.598	0.854	0.789	0.503
kNN [3]	0.643	0.755	0.547	0.851	0.794	2.154
BNLOGREG [4]	0.608	0.747	0.500	0.855	0.747	0.547

[1] F1 = (2 × Precision × Recall)/(Precision + Recall); [2] No. of tress =10 without splitting subsets smaller than 5; [3] Lasso (L1) regularization with C = 1; [4] No. of neighbors = 5, distance = Euclidean, weight = distance.

ROC (Receiver Operating Characteristic) curve graphically illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied. It summarizes the performance by combining confusion matrices at all threshold values. AUC is the area under the ROC curve and measures the entire two-dimensional area underneath the ROC curve from (0, 0) to (1, 1). AUC turns the ROC curve into a numeric representation of performance for a binary classifier, indicating how successfully a model separates positive and negative classes. Using the models' outputs, the ROC curves are plotted for TP rate (Sensitivity) vs. FP rate (1 − Specificity) for the different models tested at different classification thresholds (Figure 8). ROC curves can provide insight when studying the performance of classifiers on class-imbalanced data. The greater the AUC, the more accurately the model classifies cases. The ROC curve should ideally extend to the upper left corner, where the AUC value would be 1, indicating that the model properly classifies every instance. The model is unusable if the AUC is 0.5 for a binary classification model [39]. Figure 8 shows that RF has the biggest AUC of all the models studied here (0.712 for

RF > 0.643 for kNN). Overall, the metrics for data set (3), using the model-based imputer, yielded superior results compared to alternative data sets (1) and (2) because including all cases with missing data and cases removed with missing data caused a great deal of loss of valuable information in the training set.

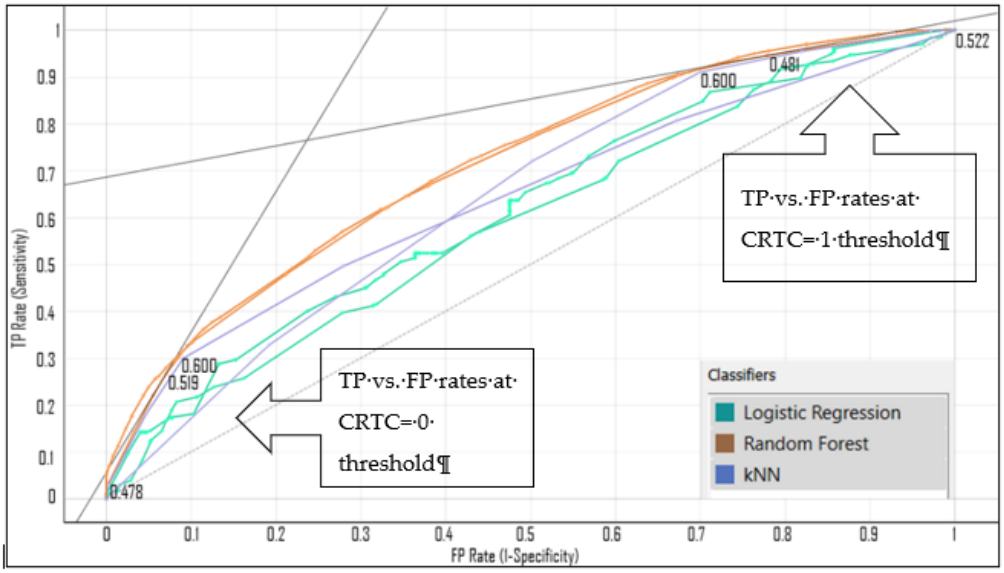

Figure 8. ROC curves for the models employed with data set (3).

The prediction results (the confusion matrix) of the RF model are given in Table 15. The accuracy rate is 76.2 percent. The misclassification rate is 23.8 percent. The model's sensitivity or recall (percentage of occurrences correctly predicted) is 93.1% (=2392/(178 + 2392). On the other hand, specificity (percentage of non-occurrences correctly predicted) is 26.4% (=229/(229 + 640). The false positive (FP) rate (the percentage of non-occurrences that are mispredicted events) is 73.6% (1 − Specificity = 1 − 26.4%). Similarly, the false negative (FN) rate (the percentage of occurrences that are predicted incorrectly is 6.9% (1 − Sensitivity = 1 − 93.1%). As seen, compared to the BNLOGREG model, although the RF model's accuracy is increased by only 1.9% from 74.8% to 76.2%, the specificity (ability to predict a true negative) is significantly improved by 83.3% from 14.4% to 26.4%.

Table 15. Prediction results (confusion matrix) of the RF model.

			Predicted		
			CRTC		Percentage
Observed			Otherwise = 0	Driver error = 1	Correct
RF model [1]	CRTC	Otherwise = 0	229	640	26.4
		Driver error = 1	178	2392	93.1
Overall percentage/accuracy rate					76.2

[1] No. of tress =10 without splitting subsets smaller than 5.

Figure 9 visualizes the impact of variables on the RF model. As seen, the highest-ranked variable is ASF, with a score of 0.405 obtained from the "rank" widget (i.e., it has the highest impact on the model's prediction). The intensity of the red dots on the left means that an increase in ASF decreases the probability of CRTC = 1 (RTCs caused by driver errors), similar to NL (its score = 0.113). However, AADT in 1000, the second highest ranked variable with a score of 0.300, increases the probability of CRTC = 1 as the red dots are intensified on the right side of the graph. Regarding the horizontal alignment features, both RG = 1 (straight section) and RG = 2 (horz.curve) increase the probability of

CRTC = 1, with scores of 0.088 and 0.017, respectively. The impact of straight sections is higher than that of horizontal curves on the model. In conclusion, road geometry significantly influences the occurrence of RTCs associated with driver errors.

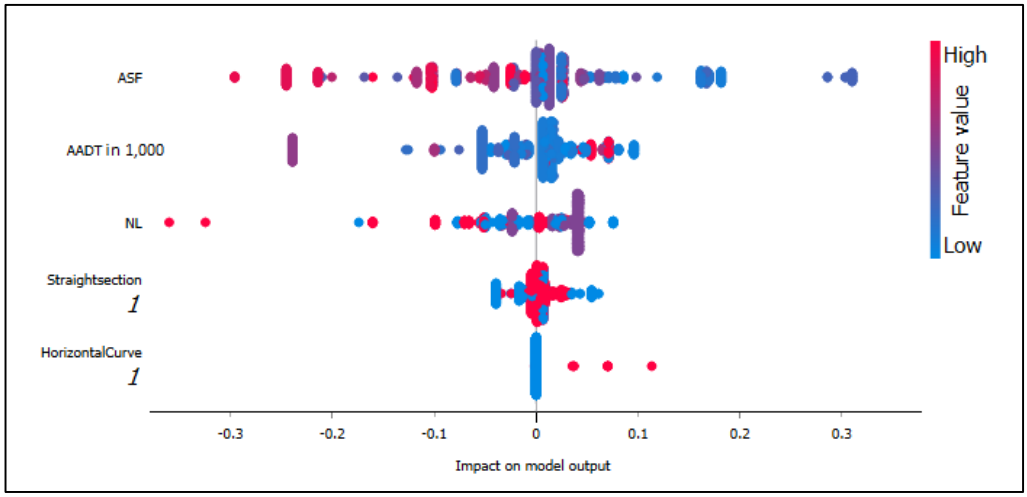

Figure 9. Impact of IVs on the RF model.

Figure 10 shows the importance of the IVs in the RF model. As seen, the variables of ASF, AADT in 1000, straight section, and NL have high impacts in decreasing the AUC by determining the probability of RTCs related to driver-errors, and ASF is the most influential feature with the highest mean value (0.195). Secondly, the impact of AADT in 1000 is the second highest, with a mean value of 0.107. The rest (straight section, NL, and horizontal curve) are ordered from high to low with mean values of 0.082, 0.069, and 0.005. The impact of the horizontal curve in reducing the AUC is the lowest in the RF model; similarly, it ranks 4th in order from high to low with its impact in the BNLOGREG model (mean value = 0.008), where the straight section variable scores the highest with a mean value of 0.079.

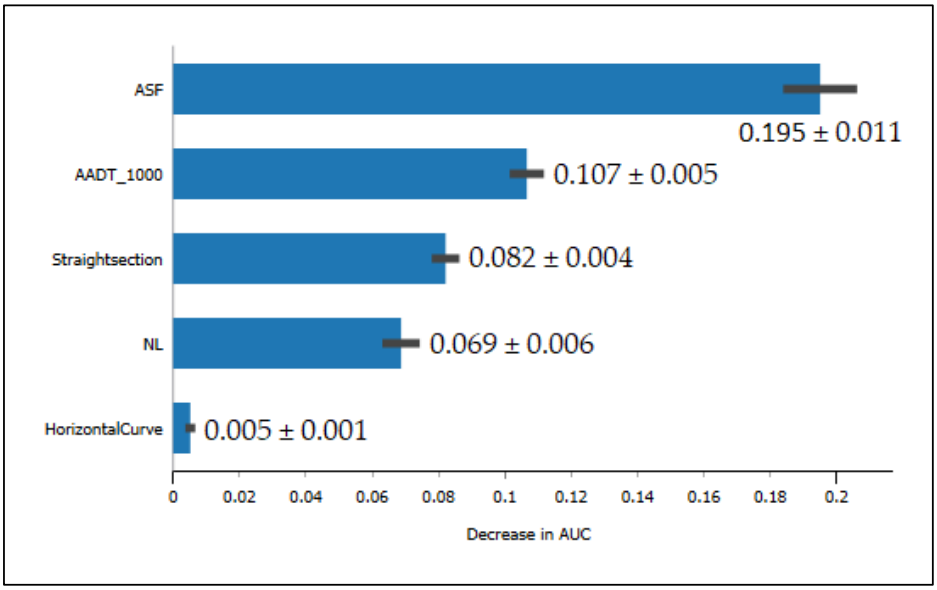

Figure 10. Effect of features in the RF model's performance in AUC (mean ± std.dev.).

4.2.3. Discussion of the Results and Research Limitations

Identifying the causes of road traffic crashes (RTCs) and contributing factors is of utmost importance for developing sustainable road network plans and urban transport

management. Driver-related factors are the leading causes of RTCs, and speed is claimed to be a major contributor to crash occurrences. The results reported in the literature are mixed regarding speed-crash occurrence causality on rural and urban roads. Even though recent studies shed some light on factors and the direction of effects, knowledge is still insufficient to allow for specific quantifications. Thus, this paper aims to discover the factors contributing to the occurrence of RTCs associated with driver errors using analytical models. In summary, while AADT and road geometry features (straight sections and horizontal curves) increase the probability of RTCs caused by driver-related errors (such as speeding or losing control), an increase in NL and ASF reduce that probability, confirming similar findings by previous studies and literature reviews as reported in [26,32,66]. The findings support previous studies (very few, in fact) in similar study contexts that looked at speed dispersion in crash occurrence and severity but disagreed with others that looked at absolute speed at individual vehicle or road segment levels. Thus, the paper contributes to insufficient knowledge of the factors in crash occurrences associated with driver errors on major roads. It is of foremost importance that roadway designers and planners use the results of this study as guidance in their efforts to implement countermeasures to reduce the occurrence and severity of RTCs in the future.

The study has some limitations due to the data set analyzed; for example, different environmental factors such as weather, pavement condition, visibility, and horizontal and vertical geometric design features were not considered in the analyses. Weather-related crashes were only 0.3% of all cases (see Table 2). Moreover, the gender, age, and level of education of drivers were not included in the data set to analyze them. Their impact on crash occurrences should be considered. For future work, the inclusion of additional variables concerning the incidence of road crashes, such as driving environment, safety treatments and speed control methods, weather conditions, and road alignment design elements, is to be considered to achieve improved performance in crash classification. Next, crash prevention and mitigation strategies are recommended regarding the factors determined in the occurrence of RTCs associated with driver errors.

4.3. Crash Prevention and Mitigation Strategies

Cultural and behavioral factors associated with dangerous driving habits in different countries must be examined to develop empirically driven strategies to prevent traffic crashes and injuries [70–76]. Compared to the general population, teen drivers are involved in a disproportionately high number of fatal and injury-incurred motor vehicle crashes [71,77–80]. Based on the summary of crash statistics presented in the preceding sections, it is of utmost significance to propose strategies that prevent and mitigate the impacts of RTCs. Such strategies can be named as developing target-oriented driver education and awareness programs and implementing technology-driven traffic control and management approaches to increase road safety in the case study area and other places in Saudi Arabia. RTCs cause notable deaths and injuries yearly in Saudi Arabia despite the severe penalties imposed on violators [81]. Due to some common contributing factors, general prevention and mitigation strategies can be applied to reduce the impact of such RTCs. However, some unique types of crashes may require in-depth analysis to identify the root causes. Al-Wathinani et al. [82] reported the voluntary responses of 316 participants in a cross-sectional study that men between 20 and 39 years old generally drove safely; however, they exceeded the legal speed limit, drove aggressively around slow drivers, and became distracted while driving at some frequency. That being said, similar driving behaviors (speeding, distracted and reckless driving, sleeping at the wheel, violating right-of-way, improper passing maneuvers, driving in the wrong direction, sudden acceleration or deceleration, mobile phone use, and taking a wrong exit) constituted 74.7% of the RTCs in this case study. The most complicated factors are erratic driver behaviors; it is usually not easy to manage because the driver population presents a wide range of behaviors with several underlying reasons. Though several studies have already offered remedies to

prevent and mitigate the impact of RTCs [62,83–89], some are found to be noteworthy to recommend to the engineers and planners within the scope of this case study.

4.3.1. Designing Safe Roads and Maintaining Work Zone Safety

Road design and construction greatly affect road safety. While some RTCs are attributed to driver errors, mechanical failures, and environmental conditions, many results from the failure to meet certain design criteria in geometric design [90]. A good design and taking necessary safety precautions eliminate conflicts by proposing the physical separation of conflicting movements and by taming excessive speed and aberrant driving behaviors that can significantly increase the accident risk on roads. Based on the findings of this study, it is evident that an increase in traffic volume increases driver errors leading to RTCs, and improved traffic flow with higher average speed and an increased number of traffic lanes reduce driver errors in the occurrence of RTCs. High-speed roads must be divided with wide medians and guard rails. On the right-hand side of such roads, emergency lanes must be available to accommodate wide vehicles in case of vehicle breakdown and a wide obstacle-free zone to prevent fixed-object crashes. Intersections with and between motorways should always be grade separated [91].

Road work zones are dangerous and crash-prone areas posing many health hazards that can cause RTCs resulting in injuries and deaths of road users or site workers. US Department of Transportation reported that work zone fatalities were 845 and 857 in 2019 and 2020 in all territories of the USA [92]. Road construction or maintenance areas can potentially contribute to increasing rear-end, fixed-object, and head-on crashes by slowing or stopping vehicles. Driver behavior was regarded as the highest risk factor in crash occurrence by road users and experts in Saudi Arabia, such as reckless or aggressive driving through the work zones. Secondly, lack of lighting and work-zone road signs were reported as other contributory factors. Suggestions to improve road safety in work zones were identified as (1) taking strict actions against contractors or consultants who create safety violations, (2) creating a stronger collaboration between government agencies to improve road safety, and (3) employing certificated safety engineers or professionals on the project for road risk assessment [93]. In that respect, Road Safety Audit (RAS) should become mandatory for all highways [90]. RSA is the formal safety performance examination of an existing or future road or a junction by an independent, multidisciplinary team. It qualitatively estimates and reports on potential road safety issues and identifies opportunities for improvements in safety for all road users [94,95]

In road design, self-explaining roads (SER) and forgiving roadside (FRS) concepts were developed to make roads safer and more user-friendly for all road users. The idea of SER encouraging drivers to naturally adopt behavior consistent with design and function was introduced by Theeuwes and Godthelp [96] and Theeuwes [97]. SER was first applied in the Netherlands and other parts of Europe. The design is aimed to be distinctive for different classes of roads regarding the features such as road width, pavement markings, signing, and street lighting that would be consistent throughout the route. Drivers would perceive the type of road and instinctively recognize how to behave. The environment effectively provides a "label" for the particular type of road, and there would thus be less need for separate traffic control devices, such as additional traffic signs, to regulate traffic behavior [98]. The notion of SER has gained great popularity and is now considered the main design principle for road authorities and departments of transportation worldwide. In many countries, roads were redesigned and adapted to be consistent with the SER principles. The EU Mobility and Transport committee also adopted this principle and funded research projects on this issue [99]. The SER approach uses simplicity and consistency of design to reduce driver stress and error. It is already used for high-speed, high-volume roads. However, on low-class roads, consistency in design is often compromised by other objectives such as high access levels, variable alignment, mixed-use and varied roadside development, which result in a lack of consistency and differentiation between road classes [98]. Similarly, the concept of FRS targets minimizing the consequences

of driving errors rather than preventing them. Safer roads and roadsides aim to reduce the risk of vehicles leaving the road, provide adequate recovery space when vehicles run off the road, and ensure that any collision occurring on the roadside will not cause severe or fatal injury to vehicle occupants) [100,101].

4.3.2. Driver Education and Awareness

Novice drivers (most in their late teens or early twenties) are overrepresented in crash statistics, and there is a clear need for remedial measures [102]. Young drivers (aged 15–25) constitute 11.8% of the male population of Saudi Arabia in 2021 and are active in daily traffic. Most importantly, they see driving quite differently from adults. Taking erratic behaviors in traffic and speeding are quite normative for them. To target young drivers, the agonizing results of RTCs must be delivered to them through diverse visual, aural, and written media, as well as traffic awareness programs in formal education institutes and driving schools. Driving schools currently present adult-oriented information and mostly ignore young drivers informing them about their erratic actions in traffic. It is recommended that first-time drivers, especially teens between 16 and 18 years who wish to apply for a driving license, complete a target-oriented driver education and awareness program. The education and awareness program must not only include information and statistics like it is conducted in regular class sitting but also should include visits to hospitals, accident sites, post-accident rehabilitation centers, and police departments. Those first-timers need to witness the implications and side effects of breaking traffic laws. And those who wish to renew their driving license may be required to attend a refresher course about erratic driver behaviors and consequences based on their driving records. In summary, increasing the level of awareness and improving the curricula of driving schools might have a higher impact on the safety of young drivers and others in road traffic, as aggressive and speedy driving behavior among young drivers (aged 18–24) is one of the most common causes of road accidents in Saudi Arabia [103]. Another hypothesis is that children inherit their parents' driving habits through genetic disposition and model learning. A series of regression models indicated that parents' self-reported driving behavior explains their children's respective self-reported behavior, even when exposure and demographic and lifestyle factors are controlled [104]. Distracted driving behavior among young drivers is quite common, and developing enforcement and educational strategies to reduce this behavior might directly affect the probability of crash risk [12]. Therefore, it is necessary to improve driver education for all ages regardless of cultural background, individual behaviors, and personal attitudes and enroll them in proper training without compromising because children are inclined to present similar behaviors to their parents. Training programs actively engaging young drivers can reduce their tendency to speed, and such programs may efficiently reduce young drivers' speeding and other aberrant driving behaviors [86]. It was found that the effectiveness of such programs was significantly increased if they were strengthened by further communication campaigns targeting key segments of the population such as the female, young, with lesser educational levels, and non-driving population, which was found to benefit less from them because not all population segments in terms of gender, education, and income have the same remembrance level, and gaps in accessing to information channels and sources may possibly vary [105]. In fact, it was shown that carefully designed and well-executed positive messages in mass-media campaigns can successfully contribute to the reduction in alcohol-impaired driving, according to U.S.-based studies published between 2007 and 2014 [106].

4.3.3. Application of Advance Technologies

Intelligent transportation and vehicle/highway systems (ITS) are offering safer driving vehicles and environments. Intelligent transportation is indeed a prominent aspect of smart city development. Intelligent transportation targets handle several issues by considering traffic or human mobility, sensory data, and geographical data generated in cities. It combines the concepts of urban sensing, data management and analytics, and various

service-providing mechanisms into a recurrent process for a discreet and continuous improvement of an individual's transit experience and operations of the city transport authority [107]. The main goal of creating intelligent transportation systems architecture is to design and implement human-focused, sustainable transportation systems with cutting-edge technologies such as industry 4.0 technologies, mobile applications, augmented reality, and the internet of things [108]. ITS technologies and monitoring systems are quite popular and reasonably well deployed in developed countries, particularly the roadways and airways [109].

Moreover, real-time highway traffic and performance monitoring of black spots and construction zones are also very promising. The technology is available and cheaper compared to previous times for tracking and classification of vehicles with the computation of traffic flow parameters. This provides information to identify the underlying reasons for certain types of RTCs. The system can process video continuously over long periods, accumulating large volumes of tracking data to build daily highway models consisting of the traffic flow parameters, density, flow, and speed. These daily models are used to categorize the speed profile of live traffic [110].

Smart vehicle technologies have developed rapidly in recent years, improving mobility and safety across transportation systems. Drivers' behavior is pivotal in developing new transportation technologies, such as connected and automated vehicle technology, and planning for future transportation systems. Connected and automated vehicle technologies are expected to improve mobility and safety significantly. As connected and autonomous vehicles have not been used in practice at a large scale, there are still some uncertainties about their applications. Therefore, researchers utilize traffic simulation tools to model the presence of these vehicles. Several studies have shown the impacts of vehicle connectivity and automation at the segment level. With the advent of connected vehicle (CV) technology, driver–vehicle behavior is expected to change significantly, as these vehicles can communicate with each other and also traffic management centers on a real-time basis [111].

A vehicle navigation system guides the vehicle along the optimal path from starting to destination. A reliable vehicle navigation system can reduce traffic chaos in the city and improve the level of service [112]. Drivers who use a navigation system can travel with less stress and more confidence behind the wheel. According to those who think that driving with a navigation system that delivers traffic information improves the quality of the chosen route, this positively impacts traffic safety. As a result, the journey time and navigation errors are reduced. However, more studies on human behavior need to be pursued concerning using cellular telephones and route guidance in-vehicle navigation systems [113–115]. Using an intelligent speed adaptation and safety system (ISASS), if speed limits were strictly followed, road fatalities and hospitalized injuries are estimated to be reduced by 20% and 15%, and fuel usage and carbon dioxide emissions lowered by 11% [116].

Finally, it should be underlined that potential problems are needed to be addressed in implementing advanced technologies and user willingness related to shared data privacy and user perceptions of insecurity. Vehicles equipped with such technologies continuously broadcast data while using certain ITS services, including their speed and location. Some interviewed stakeholders raised their concerns about data privacy and protection breaches because the data are personal, and the driver must consent to public authorities to use the data according to the European General Data Protection Regulation. There can be only a few exceptions to this when the use of the data is of vital interest to the driver himself or the public in general [117].

4.3.4. In-Vehicle Technologies and Autonomous Driving

In-vehicle technologies for safe driving are included in ITS; however, many car manufacturers have already developed and employed several of them successfully in certain vehicle types and models independently of the integrated intelligent vehicle-highway systems. Autonomous vehicles equipped with advanced in-vehicle technologies (AVTs)

are expected to improve road traffic safety and reduce accidents by replacing the driver's role and introducing new capabilities [118]. This includes vehicle control technology to mitigate harm due to collisions, leveraging automation to react quicker and more consistently than human drivers. Among researchers, the most discussed AVT helps to avoid obstacles, i.e., to plan the motions of vehicles to eliminate collisions [119,120]. However, due to uncertainty created by the behavior of other road users, collisions can never be eliminated [121]. Increased use of AVTs has generated excitement and concern among researchers, policymakers, and the public. An increasing number of driver assistance systems are already available in today's automobiles, and many of them are expected to become standard. Such technologies must ensure to meet the needs of drivers, particularly younger and older age groups, who are known to have a higher crash risk [118]. Such existing and new technologies must focus on improving road safety for all users with both manual and autonomous driving.

In-vehicle (plug-in) monitoring driving devices [122,123] must be considered for new drivers, traffic law violators, or a specific-age target group. Such devices would record speed, excessive breaking or accelerating, hard turns, going over or below the speed limits, changing lanes without signaling, losing vehicle control, and so on. The owner or user of such devices would get a performance report of these parameters. Some private companies in the US are using plug-in monitoring driving devices to monitor the driving behaviors of their employees, especially those who work in the field [124]. Some insurance companies in the US offer insurance policy discounts for their policyholders if they agree to plug in those monitoring driving devices. A growing number of people are ready to give it a try—in exchange for lower insurance premiums, according to recent research from Nationwide, one of the largest insurance companies in the U.S. The same study shows two-thirds of consumers said they would allow a device to monitor their driving behavior if it provided a discount [125]. Among the obvious cost-saving benefits, the driver will be more vigilant and careful since his driving is monitored consistently. In addition, authorities and researchers would have access to real and accurate behavioral driving data to study and analyze. Issue of data privacy might be an issue for some individuals or countries; however, the same concerns might be said about personal health data.

4.3.5. Legislation and Enforcement of Traffic Regulations

Saudi traffic laws focus on correcting behavior. Strict traffic regulations are expected to deter anyone who lacks a sense of responsibility when using roads and penalize those who fail to correct their behavior. Traffic laws are created and applied to correct improper behavior and benefit road users. A road traffic crash is deemed to be a liability if it is due to driver negligence. The law mechanism in Saudi Arabia ensures road safety for all users [126]. So, when an integrated traffic system is in place, it establishes fundamental rules that contribute to reducing traffic accidents and achieving the necessary security and goals established by authorities [1]. For example, speed regulation policies towards high-level speeding can be highly effective. Viallon and Laumon [24] reported that the fraction of fatal crashes attributable to high-level speeding (>20 kph over the speed limit) decreased from 25% to 6% and that attributable to medium-level speeding (10–20 kph over the speed limit) decreased from 13% to 9%, whereas that attributable to low-level speeding progressively increased from 7% to 13%. Similar trends were observed on main roads. These results demonstrate the effectiveness of the speed regulation policies introduced during the study period with respect to high-level speeding. They also suggest that future policy should focus on low and medium-level speeding to significantly reduce road deaths since these levels correspond to the major fraction of fatal crashes [24].

In order to achieve the targeted goals, current legislation and enforcement of traffic regulations must be continuously reviewed and updated. Most of the cities of Saudi Arabia have already accommodated the enforcement of traffic safety regulations and added smart cameras to detect overspeeding, crossing at the red light, mobile phone use while driving, and not wearing seatbelts. These safety management steps are expected to improve road

safety in Saudi Arabia; however, human behaviors are quite complex and often hard to modify; thus, continuous data collection and analyses are desired for the betterment of traffic safety.

Finally, developing a reward system for law-obeying drivers can be implemented to encourage drivers to further obey the laws by offering tax cuts, lower insurance costs, reduced renewal driving license fees, gas coupons, or any other incentive-driven programs. For example, the North Carolina Safe Driver Incentive Plan (SDIP) was created by the state law to give drivers a financial incentive to practice safe driving habits. SDIP points are charged for convictions and at-fault accidents occurring during the experience period (a three–year period preceding either the date an individual applies for coverage or the insurance company prepares to renew an existing policy) [127].

Through multiple road safety initiatives (i.e., speed limit and red light violation cameras, geometric improvements of highway sections and intersections, implementing warning vibrations on shoulders along highspeed highways, placing high-tensile fences and guardrails, and so on), and cooperation among different governmental agencies, Saudi Arabia 2030 Vision already achieved 13.5 deaths per 100,000 in 2021, and it is targeting to reduce the number of traffic fatalities on its road network to 8 deaths per 100,000 by 2030 [128]. The country observed a 33% decrease in traffic accident deaths and 25% in injuries and accidents during 2018 compared to 2017 [129]. It is recommended that transportation authorities, traffic and road engineers, drivers, and other stakeholders continue to work together to address this critical issue. Furthermore, an integrated approach that combines engineering interventions, education initiatives, and enforcement actions is recommended to increase awareness and compliance among drivers as a means to support crash prevention in the future.

4.3.6. Benefits of Academic Studies and Research on Traffic Safety

Oftentimes transferability of findings from academic studies and research on traffic safety from other places can be quite limited if the cultural background and personal attitudes are rather different in local cases. For this reason, research and scientific studies on traffic safety should be supported and encouraged by universities and research institutions at local, regional, and national levels. The desired level of collaboration among related authorities, such as the Ministries of Transportation, Interior, and Health, must be achieved to foster traffic safety studies and reduce the frequency and severity of traffic crashes. Data regarding traffic crashes, such as location, type, involvements, hospitalization, cost of damage, and so on, must be systematically collected, stored, and processed in a national database with authorized accesses and with open and free access to the public. It is recommended that government and private funds be used to create a "national traffic study fund initiative" that will be used to support research initiatives related to the betterment of traffic safety in the future.

5. Summary and Conclusions

As many developed and developing, nations are dealing with the crisis of RTCs causing many lives to be lost and tremendous costs associated with them in terms of monetary value and human assets. Studies show that driver-related factors are the leading causes of road traffic crashes (RTCs). This paper aimed to determine the road features and traffic flow parameters leading to RTCs associated with driver errors along a 761.6 km long section of an access-controlled major road (Highway 15 between Taif and Medina) in Saudi Arabia using supervised ML models. BNLOGREG, RF, and kNN models were developed, and several standard metrics were used to measure their performances, including the AUC of the ROC curve, F1, classification accuracy (CA), and the Youden index. First, the BNLOGREG equation was developed to understand the role of the input variables in explaining the outcome. The model's accuracy is 74.8%. The five independent variables (IV) found statistically significant at 0.01 level by the regression are NL ($\beta = -0.501$,

sig. < 0.001), ASF (β = −0.027, sig. < 0.001), AADT (β = 0.016, sig. < 0.001), the road geometry features (RG1, β = 0.824, sig. < 0.001; RG2, β = 1.268, sig. = 0.005).

The performance of the BNLOGREG model was compared with those of RF and kNN. The RF model yielded the highest CA of 76.2%. For such unbalanced data with a high percentage of RTCs (CRTC = 1, 74.7% driver-related cases), the metric of balanced accuracy (BAC) was calculated to see the actual accuracy. Compared to CAs, BACs were significantly reduced, ranging from 19.5 to 36.5%. They decreased the most (36.5%) for the RF and BNLOGREG models with data set (2), i.e., cases with valuable information of the predictors were removed from the data set. The reduction was less (21.5 to 33.1%) with the imputed data set (3) used in training. All models performed well regarding the BAC. The RF model outperformed the rest in all metrics except F1 with data sets (1) and (3). The BNLOGREG model ranked first regarding the F1 metric with all data sets. Overall, RF is the best model with its highest AUC (0.712), followed by kNN (0.643).

The following conclusions are reached:

- The performance of all models is comparable, so they are found to be suitable for predicting the probability of driver errors in the occurrence of RTCs and understanding the role of the input variables in explaining the model outcomes.
- The two most influential variables are ASF and AADT in the RF model. In line with the findings of previous research conducted in a similar study context, an increase in the number of lanes (NL) and daily average speed of traffic flow (ASF) reduces the likelihood of the RTCs caused by driver errors. This finding is also supported by the results of previous studies [32,67]. In contrast, an increase in traffic volume (AADT) and the road geometry features (straight sections and horizontal curves) significantly contributed to driver errors leading to RTCs.
- Straight road sections and the sections with horizontal curves increase the probability of driver-error-related RTCs by more than two and three folds (odds ratios = 2.280 and 3.555 yielded by BNLOGREG).
- The impact of geometric elements is significant on the RF model's output (Figures 9 and 10). Thus, it is concluded that road geometry substantially influences the occurrence of RTCs associated with driver errors.
- The inferences concerning the effects of crash attributes are in agreement with the findings in the literature. Thus, the paper sufficiently contributes to insufficient knowledge of the factors in RTCs on major roads within the context of this case study.

Finally, it is important to acknowledge that analyses of RTCs in different cultural and environmental settings may yield unique results in determining the contributing factors and their weights in calculating the probabilities of outcomes that may be affected by local conditions. For this reason, other data sets should be obtained and analyzed to examine the transferability of the results of this study to other cases and situations. Detailed analyses of RTCs concerning the features of road geometry, especially at curved road sections with the combination of high-grade straight sections and vertical curves, are recommended for future study.

Author Contributions: Conceptualization, D.A. and V.P.S.; methodology, D.A. and V.P.S.; formal analysis, D.A.; data curation, A.H.A., A.O.A. and M.M.H.A.-T.; writing—original draft preparation, D.A.; writing—review and editing, D.A., V.P.S., A.H.A., A.O.A., M.M.H.A.-T. and K.A.A.A.-S.; supervision, K.A.A.A.-S.; funding acquisition, D.A., V.P.S. and A.H.A.; project administration, D.A., A.H.A. and A.O.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was funded by institutional fund projects under no. IFP-A-2022-2-1-07. Therefore, the authors gratefully acknowledge technical and financial support from the Ministry of Education (MOE) and the University of Hafr Al Batin (UHB), Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of study data. Data were obtained from the Ministry of Transportation and Logistic Services (MOTLS) of Saudi Arabia and are available with its permission.

Acknowledgments: The RTC data procured from the Ministry of Transportation and Logistic Services (MOTLS) of Saudi Arabia is greatly acknowledged. Special thanks go to Mohammed Mesfer Al-Abood (General Director of Road Safety, MOTLS) and Ahmed Hassan Hasib (Principal Engineer-Traffic Engineering). Finally, we thank the anonymous reviewers for their valuable comments and questions, which helped improve the original submission.

Conflicts of Interest: The authors declare no conflict of interest. The funder had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

BNLOGREG model: The typical setup for logistic regression is as follows: there is an outcome y that falls into one of two categories (say 0 or 1), and Equation (A1) below is used to estimate the probability that y belongs to a particular category given inputs, $X = (x_1, x_2, \ldots, x_k)$ [69]:

$$P(y = 1|X) = sigmoid(z) = \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z} \qquad where\ x \in [\![0,\ 1]\!] \tag{A1}$$

and

$$z = \hat{\beta_0} + \hat{\beta_1} x_1 + \hat{\beta_2} x_2 + \ldots + \hat{\beta_k} x_k \tag{A2}$$

The equation for z is called a linear predictor, and it is transformed by the sigmoid function so that the values fall between 0 and 1, and can therefore be interpreted as probabilities. $\hat{\beta_i}$'s are the estimated linear coefficients. The resulting probability is then compared to a threshold to predict a class for y based on X. In BNLOGREG used in this study, the probability of category membership on a dichotomous dependent variable (DV: cause of the road traffic crash (CRTC), driver-related = 1, and otherwise = 0) was based on multiple independent variables (IVs: number of lanes (NL), the daily average speed of traffic flow (ASF), annual average daily traffic (AADT), and road geometry (RG: binary, straight section = 1, otherwise = 0, and horizontal curve = 1, otherwise = 0)). Three of the IVs (NL, ASF, and AADT) are continuous (i.e., interval or ratio in scale), and the remaining two (RG1 = straight section and RG2 = horizontal curve) are categorical (binary). Non-metric IVs (RG1 and RG2) are included as a "factor," and dummy-coded (straight road section = 1, otherwise = 0, and horizontal curve = 1, otherwise = 0), and continuous IVs (NL, ASF, and AADT) are included as "covariates." The same DV and IVs are used in other supervised classification models used in this study.

While binomial logistic regression does compute correlation measures to estimate the strength of the relationship (pseudo-R square measures), these correlation measures do not really tell much about the accuracy or errors associated with the model. The regression coefficients for logistic regression are calculated using maximum likelihood estimation (MLE). The natural logarithm of the odds ratio is equivalent to a linear function of the independent variables. The antilog of the logit function permits the following estimation of the regression equation, Equation (A3):

$$logit\ (\hat{p}) = \ln\left(\frac{\hat{p}}{1 - \hat{p}}\right) = \hat{\beta_0} + \hat{\beta_1} \cdot \mathrm{NL} + \hat{\beta_2} \cdot \mathrm{AADT} + \hat{\beta_3} \cdot \mathrm{ASF} + \hat{\beta_4} \cdot RG1 + \hat{\beta_5} \cdot RG2 \tag{A3}$$

We isolate p by taking the antilog of Equation (A3) and get:

$$\frac{\hat{p}}{1 - \hat{p}} = e^{\hat{\beta_0} + \hat{\beta_1} \cdot \mathrm{NL} + \hat{\beta_2} \cdot \mathrm{AADT} + \hat{\beta_3} \cdot \mathrm{ASF} + \hat{\beta_4} \cdot RG1 + \hat{\beta_5} \cdot RG2} \tag{A4}$$

We continue to solve for $\hat{p}$ as shown below:

$$\hat{p} = e^{\hat{\beta_0}+\hat{\beta_1}\cdot \mathrm{NL}+\hat{\beta_2}\cdot \mathrm{AADT}+\hat{\beta_3}\cdot \mathrm{ASF}+\hat{\beta_4}\cdot RG1+\hat{\beta_5}\cdot RG2}(1-\hat{p}) \tag{A5}$$

Then, by skipping the intermediate steps, we estimate the probability $\hat{p}$ for the logistic regression equation:

$$\hat{p} = \frac{e^{\hat{\beta_0}+\hat{\beta_1}\cdot \mathrm{NL}+\hat{\beta_2}\cdot \mathrm{AADT}+\hat{\beta_3}\cdot \mathrm{ASF}+\hat{\beta_4}\cdot RG1+\hat{\beta_5}\cdot RG2}}{1+e^{\hat{\beta_0}+\hat{\beta_1}\cdot \mathrm{NL}+\hat{\beta_2}\cdot \mathrm{AADT}+\hat{\beta_3}\cdot \mathrm{ASF}+\hat{\beta_4}\cdot RG1+\hat{\beta_5}\cdot RG2}} \tag{A6}$$

and the BNLOGREG equation becomes:

$$P(CRTC = Driver\ error = 1|0) = \frac{1}{1+e^{-(\hat{\beta_0}+\hat{\beta_1}\cdot \mathrm{NL}+\hat{\beta_2}\cdot \mathrm{AADT}+\hat{\beta_3}\cdot \mathrm{ASF}+\hat{\beta_4}\cdot RG1+\hat{\beta_5}\cdot RG2}} \tag{A7}$$

As two common problems, overfitting and underfitting, are observed in regression models, regularization is used to train the model of BNLOGREG in order to avoid them. In the case of overfitting, the model performs well on the training data set but not so well on the test data set. In underfitting, the model neither performs well on the training data nor on the test data sets. Simplifying the model as in Equation (A2) and using fewer parameters overcomes the overfitting issue. In addition, other approaches can be considered to overcome the overfitting problem, such as (i) regularizing the model (putting constraints on the model with Lasso-Least Absolute Shrinkage and Selection Operator, and Ridge regression techniques), (ii) gathering more training data, (iii) removing variables that do not improve the model, and (iv) removing the noise in the training data (e.g., fixing data errors and removing outliers).

Underfitting occurs when the model is too simple to understand the underlying structure of the data properly. This can be overcome by (i) building a better model with more parameters, (ii) feeding better features to the learning algorithm, and/or (iii) reducing the constraints on the model. It must be noted that increasing the sample size for the training data set will not help. The regularization technique adds an extra term—the regularization term– to the error function used in the training stage. The Ridge regularization term is the sum of the squares of all parameters, which is known as weight decay and drives parameters toward zero. The Lasso regularization term is the sum of the absolute values of the parameters [130]. Equations (A8) and (A9) show the Ridge and Lasso terms added to the minimization of the residual sum of squares (RSS), respectively [69]:

$$\lambda \sum_{i=1}^{p} |\beta_j| \tag{A8}$$

$$\lambda \sum_{j=1}^{p} \beta_j^2 \tag{A9}$$

where p is the number of variables included in the model, and $\lambda \geq 0$ is a tuning parameter to be determined separately. Ridge and Lasso implicitly function as their own form of feature selection, as shown in Equations (A8) and (A9). The attributes that do not contribute to the predictive power of the regression have their coefficients lowered, but the more predictive features have higher coefficients, despite the extra penalty. Because Ridge regression squares the coefficients in the penalty term, coefficients on less valuable features tend to approach zero but not quite reach it. On the other hand, Lasso will shrink certain parameters (β_j) towards zero, but it will not set any of them exactly to zero (unless $\lambda = \infty$). This may not hurt prediction accuracy, but it can make model interpretation difficult in the circumstances with a large number of variables, p. However, as $\lambda \to \infty$, the shrinkage penalty's weight grows, and regression coefficient estimates will approach zero. Unlike the least squares method, which generates only one set of coefficient estimates, Lasso and

Ridge regression will produce a different set of coefficient estimates, $\hat{\beta}_{\lambda}^{L}$ and $\hat{\beta}_{\lambda}^{R}$, for each value of λ. So, selecting a good value for λ is critical [60].

Random forest (RF) algorithm: A Random Forest is a tree-based ensemble with each tree depending on a collection of random variables. More formally, for a p-dimensional random vector $X = (X_1, \ldots, X_p)^T$ representing the real-valued input or predictor variables and a random variable Y representing the real-valued response, we assume an unknown joint distribution $P_{XY}(X, Y)$. The goal is to find a prediction function $f(x)$ for predicting Y. The prediction function is determined by a loss function $L(Y, f(X))$ and defined to minimize the expected value of the loss [51]:

$$E_{XY}(L(Y, f(X))) \tag{A10}$$

where the subscripts denote expectation with respect to the joint distribution of X and Y.

Intuitively, $L(Y, f(X))$ is a measure of how close $f(X)$ is to Y; it penalizes values of $f(X)$ that are a long way from Y. Typical choices of L are squared error loss $L(Y, f(X)) = (Y - f(X))^2$ for regression and zero-one loss for classification [51]:

$$L(Y, f(X)) = I(Y \neq f(X)) = \begin{cases} 0 \text{ if } Y = f(X) \\ 1 \text{ otherwise.} \end{cases} \tag{A11}$$

In the classification situation, if the set of possible values of Y is denoted by $\mathcal{Y}$, minimizing $E_{XY}(L(Y, f(X)))$ for zero one loss gives

$$f(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} P(Y = y | X = x), \tag{A12}$$

otherwise known as the Bayes rule [51].

Ensembles construct f in terms of a collection of so-called "base learners" $h_1(x), \ldots, h_j(x)$ and these base learners are combined to give the "ensemble predictor" $f(x)$. In classification, $f(x)$ is the most frequently predicted class ("voting")

$$f(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \sum_{j=1}^{J} I\big(y = h_j(x)\big) \cdot \tag{A13}$$

In Random Forests, the jth base learner is a tree denoted $h_j(X, \Theta_j)$, where Θ_j is a collection of random variables and the Θ_j's are independent for $j = 1, \ldots, J$. To understand the Random Forest algorithm, it is important to know the type of trees used as base learners [51].

k-nearest neighbor (kNN): kNN is commonly employed in pattern recognition. Nearest-neighbor classifiers learn by analogy or by comparing a set of test data with similar training data. The training dataset is characterized by n attributes. Each data point represents a location in a space with n dimensions. This way, all the training data are stored in n-dimensional pattern space. When given unknown data, a kNN classifier searches the pattern space for the k training data closest to the unknown data. These k training data are the k "nearest neighbors" of the unknown data, which is the core decision factor in the model's accuracy. k is usually an odd number if the number of classes is even. The model's performance must be checked by calculating the prediction on different values of k and comparing their performance. "Closeness" is defined as a distance metric, such as Euclidean distance [39]. The Euclidean distance between two points, say, $X_1 = (x_{11}, x_{12}, \ldots, x_{1n})$ and $X_2 = (x_{21}, x_{22}, \ldots, x_{2n})$, is described by Equation (A14):

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \tag{A14}$$

There are other distance measures, like the Manhattan distance (Equation (A15)), which is also very commonly used for continuous variables:

$$dist(X_1, X_2) = \sum_{i=1}^{n} |x_{1i} - x_{2i}| \quad \text{(A15)}$$

Typically, the values of continuous attributes are normalized before using Equation (A14) or Equation (A15). This prevents attributes with initially large ranges (e.g., AADT) from outweighing the ones with smaller ranges (e.g., NL and binary attributes such as RG1 and RG2). The algorithm has the following three steps: (i) calculate distance using Equation (A14) or Equation (A15), (ii) find k-nearest neighbors, and (iii) assign a class containing the maximum number of nearest neighbors. In the case of very small values of k, the algorithm is too sensitive to noise. Larger values of k make the class boundaries smoother, which might not be desirable as the points of other classes may get included in the neighborhood. When the training data points are scattered, the value of k is difficult to determine [131].

References

1. Rahman, M.M.; Islam, M.K.; Al-Shayeb, A.; Arifuzzaman, M. Towards Sustainable Road Safety in Saudi Arabia: Exploring Traffic Accident Causes Associated with Driving Behavior Using a Bayesian Belief Network. *Sustainability* **2022**, *14*, 6315. [CrossRef]
2. United Nations | Saudi Arabia: World Health Organization Together with the Ministry of Health and Ministry of Interior. Join Efforts to Reduce Mortality from Road Traffic Accidents. Available online: https://saudiarabia.un.org/en/105869-world-health-organization-together-ministry-health-and-ministry-interior-join-efforts-reduce (accessed on 9 September 2022).
3. Transport and Infrastructure in Saudi Arabia. Available online: https://www.worlddata.info/asia/saudi-arabia/transport.php (accessed on 29 October 2022).
4. Worldbank. Saudi Arabia | Overview. Available online: https://data.worldbank.org/country/SA (accessed on 7 October 2022).
5. General Authority for Statistics (GSTAT), 2021. Population Estimates in the Midyear of 2021. Available online: https://www.stats.gov.sa/sites/default/files/POP%20SEM2021E.pdf (accessed on 27 August 2022).
6. Alghnam, S.; Alkelya, M.; Alfraidy, M.; Al-Bedah, K.; Albabtain, I.T.; Alshenqeety, O. Outcomes of road traffic injuries before and after the implementation of a camera ticketing system: A retrospective study from a large trauma center in Saudi Arabia. *Ann. Saudi Med.* **2017**, *37*, 1–9. [CrossRef] [PubMed]
7. Sauber-Schatz, E.K.; Parker, E.M.; Sleet, D.A.; Ballesteros, M.F. Road & Traffic Safety—Chapter 8—Yellow Book | Travelers' Health | CDC, 2020. Available online: https://wwwnc.cdc.gov/travel/yellowbook/2020/travel-by-air-land-sea/road-and-traffic-safety (accessed on 7 October 2022).
8. WISQARS (Web-based Injury Statistics Query and Reporting System) | Injury Center | CDC, 2021. Available online: https://www.cdc.gov/injury/wisqars/index.html (accessed on 7 October 2022).
9. Ministry of Health Saudi Arabia (MOH). Available online: https://www.moh.gov.sa/en/Pages/Default.aspx (accessed on 28 October 2022).
10. Harbeck, E.L.; Glendon, A.I.; Hine, T.J. Young driver perceived risk and risky driving: A theoretical approach to the "fatal five". *Transp. Res. Part F Traffic Psychol. Behav.* **2018**, *58*, 392–404. [CrossRef]
11. Shen, S.; Neyens, D.M. Factors affecting teen drivers' crash-related length of stay in the hospital. *J. Transp. Health* **2016**, *4*, 162–170. [CrossRef]
12. Shaaban, K.; Gaweesh, S.; Ahmed, M.M. Investigating in-vehicle distracting activities and crash risks for young drivers using structural equation modeling. *PLoS ONE* **2020**, *15*, e0235325. [CrossRef] [PubMed]
13. Ansari, S.; Akhdar, F.; Mandoorah, M.; Moutaery, K. Causes and effects of road traffic accidents in Saudi Arabia. *Public Health* **2000**, *114*, 37–39. [CrossRef] [PubMed]
14. Patterson, T.L.; Frith, W.J.; Small, M.W. *Down with Speed: A Review of the Literature, and the Impact of Speed on New Zealanders*; Accident Compensation Corporation and Land Transport Safety Authority: Wellington, New Zealand, 2000.
15. Ladegaard, I. For Drivers with Heavy Eyelids, Good Roads can Kill, 2012. Available online: https://sciencenorway.no/cars-and-traffic-forskningno-norway/for-drivers-with-heavy-eyelids-good-roads-can-kill/1373723 (accessed on 11 August 2022).
16. Hauer, E. Speed and safety. *Transport. Res. Rec.* **2009**, *2103*, 10–17. [CrossRef]
17. Montella, A.; Imbriani, L.L. Safety performance functions incorporating design consistency variables. *Accid. Anal. Prev.* **2015**, *74*, 133–144. [CrossRef] [PubMed]
18. Wang, C.; Quddus, M.A.; Ison, S.G. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. *Accid. Anal. Prev.* **2011**, *43*, 1979–1990. [CrossRef]
19. Yau, K.K.W. Risk factors affecting the severity of single-vehicle traffic accidents in Hong Kong. *Accid. Anal. Prev.* **2004**, *36*, 333–340. [CrossRef] [PubMed]
20. Yu, R.; Abdel-Aty, M. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. *Saf. Sci.* **2014**, *63*, 50–56. [CrossRef]

21. Montella, A.; Andreassen, D.; Tarko, A.; Turner, S.; Mauriello, F.; Imbriani, L.L.; Romero, M. Crash databases in Australasia, the European Union, and the United States: Review and prospects for improvement. *Transport. Res. Rec.* **2013**, *2386*, 128–136. [CrossRef]
22. Montella, A.; Imbriani, L.L.; Marzano, V.; Mauriello, F. Effects on speed and safety of point-to-point speed enforcement systems: Evaluation on the urban motorway A56 Tangenziale di Napoli. *Accid. Anal. Prev.* **2015**, *74*, 133–144. [CrossRef] [PubMed]
23. NHTSA. MMUCC Guideline: Model Minimum Uniform Crash Criteria—5th ed. *Rep. DOT HS* **2017**, *812*, 433.
24. Viallon, V.; Laumon, B. Fractions of fatal crashes attributable to speeding: Evolution for the period 2001–2010 in France. *Accid. Anal. Prev.* **2013**, *52*, 250–256. [CrossRef]
25. Ratanavaraha, V.; Suangka, S. Impacts of accident severity factors and loss values of crashes on expressways in Thailand. *IATSS Res.* **2014**, *37*, 130–136. [CrossRef]
26. Aarts, L.; van Schagen, I. Driving speed and the risk of road crashes: A review. *Accid. Anal. Prev.* **2006**, *38*, 215–224. [CrossRef]
27. Hummer, J.E.; Rasdorf, W.; Findley, D.J.; Zegeer, C.V.; Sundstrom, C.A. *Procedure for Curve Warning Signing, Delineation, and Advisory Speeds for Horizontal Curves*; NCDOT Report, No. FHWA/NC/2009-07; North Carolina Department of Transportation: Raleigh, NC, USA, 2010.
28. Findley, D.J.; Hummer, J.E.; Rasdorf, W.; Zegeer, C.V.; Fowler, T.J. Modeling the impact of spatial relationships on horizontal curve safety. *Accid. Anal. Prev.* **2012**, *45*, 296–304. [CrossRef]
29. Glennon, J.C. *Effect of Alignment on Highway Safety. State of the Art Report 6*; Transportation Research Board. National Research Council: Washington, DC, USA, 1987.
30. Zegeer, C.V.; Stewart, J.R.; Council, F.M.; Reinfurt, D.W.; Hamilton, E. Safety Effects of Geometric Improvements on Horizontal Curves. *Transport. Res. Rec.* **1992**, *1356*, 11–19.
31. Bauer, K.M.; Harwood, D.W. Safety Effects of Horizontal Curve and Grade Combinations on Rural Two-Lane Highways. *Transport. Res. Rec.* **2013**, *2398*, 37–49. [CrossRef]
32. Garber, N.J.; Gadiraju, R. Factors affecting speed variance and its influence on accidents. *Transport. Res. Rec.* **1989**, *1213*, 64–71.
33. Guido, G.; Shaffiee Haghshenas, S.; Vitale, A.; Astarita, V.; Park, Y.; Geem, Z.W. Evaluation of Contributing Factors Affecting Number of Vehicles Involved in Crashes Using Machine Learning Techniques in Rural Roads of Cosenza, Italy. *Safety* **2022**, *8*, 28. [CrossRef]
34. Farhangi, F.; Sadeghi-Niaraki, A.; Razavi-Termeh, S.V.; Choi, S.-M. Evaluation of Tree-Based Machine Learning Algorithms for Accident Risk Mapping Caused by Driver Lack of Alertness at a National Scale. *Sustainability* **2021**, *13*, 10239. [CrossRef]
35. Tamakloe, R.; Park, D. Factors Influencing Fatal Vehicle-Involved Crash Consequence Metrics at Spatio-Temporal Hotspots in South Korea: Application of GIS and Machine Learning Techniques. *Int. J. Urban Sci.* **2022**, 1–35. [CrossRef]
36. Tamakloe, R.; Sam, E.F.; Bencekri, M.; Das, S.; Park, D. Mining groups of factors influencing bus/minibus crash severities on poor pavement condition roads considering different lighting status. *Traffic inj. Prev.* **2022**, *23*, 308–314. [CrossRef]
37. Mirzahossein, H.; Sashurpour, M.; Hosseinian, S.M.; Gilani, V.N.M. Presentation of machine learning methods to determine the most important factors affecting road traffic accidents on rural roads. *Front. Struct. Civ. Eng.* **2022**, *16*, 657–666. [CrossRef]
38. Hu, P.; Li, Y.; Liu, Y.; Guo, G.; Gao, X.; Su, Z.; Wang, L.; Deng, G.; Yang, S.; Qi, Y.; et al. Comparison of Conventional Logistic Regression and Machine Learning Methods for Predicting Delayed Cerebral Ischemia After Aneurysmal Subarachnoid Hemorrhage: A Multicentric Observational Cohort Study. *Front. Aging Neurosci.* **2022**, *14*, 857521. [CrossRef]
39. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*, 3rd ed.; Morgan Kaufmann: Burlington, MA, USA, 2011; p. 391. [CrossRef]
40. Glubb, J.B. Mecca | Definition, History, Pilgrimage, Population, Kaaba, City, & Facts | Britannica, 2022. Available online: https://www.britannica.com/place/Mecca (accessed on 7 October 2022).
41. Glubb, J.B.; Abdo, A.S. Madinah | Britannica, 2022. Available online: https://www.britannica.com/place/Medina-Saudi-Arabia (accessed on 7 October 2022).
42. Bujang, M.A.; Sa'at, N.; Joys, A.R.; Ali, M.M. An audit of the statistics and the comparison with the parameter in the population. In Proceedings of the AIP Conference Proceedings 1682, Selangor, Malaysia, 22 October 2015. [CrossRef]
43. Bujang, M.A.; Ghani, P.A.; Bujang, M.A.; Zolkepali, N.A.; Adnan, T.H.; Ali, M.M.; Selvarajah, S.; Haniff, J. A comparison between convenience sampling versus systematic sampling in getting the true parameter in a population: Explore from a clinical database: The Audit Diabetes Control Management (ADCM) registry in 2009. In Proceedings of the 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE), Langkawi, Kedah, Malaysia, 10–12 September 2012; pp. 1–5. [CrossRef]
44. Bujang, M.A.; Sa'at, N.; Sidik, T.M.I.T.A.; Joo, L.C. Sample Size Guidelines for Logistic Regression from Observational Studies with Large Population: Emphasis on the Accuracy Between Statistics and Parameters Based on Real Life Clinical Data. *Malays. J. Med. Sci.* **2018**, *25*, 122–130. [CrossRef]
45. Nemes, S.; Jonasson, J.M.; Genell, A.; Steineck, G. Bias in odds ratios by logistic regression modelling and sample size. *BMC Med. Res. Methodol.* **2009**, *9*, 56. [CrossRef]
46. Demsar, J.; Curk, T.; Erjavec, A.; Gorup, C.; Hocevar, T.; Milutinovic, M.; Mozina, M.; Polajnar, M.; Toplak, M.; Staric, A.; et al. Orange: Data Mining Toolbox in Python. *J. Mach. Learn Res.* **2013**, *14*, 2349–2353.
47. Ghazvini, K.; Yousefi, M.; Firoozeh, F.; Mansouri, S. Predictors of tuberculosis: Application of a logistic regression model. *Gene Rep.* **2019**, *17*, 100527. [CrossRef]

48. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009; pp. 120–124.
49. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
50. Breiman, L.; Cutler, A. *Random Forests-Classification Description*; Department of Statistics, University of California: Berkeley, CA, USA, 2007.
51. Cutler, A.; Cutler, D.R.; Stevens, J.R. Random forests. In *Book Title Ensemble Machine Learning: Methods and Applications, Chapter 5*, 1st ed.; Zhang, C., Ma, Y., Eds.; Springer: New York, NY, USA, 2011; Volume 45, pp. 157–175.
52. Dietterich, T.G. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Mach. Learn.* **1999**, *40*, 139–157. [CrossRef]
53. Freund, Y.; Schapire, R.E. Experiments with a new boosting algorithm. In Proceedings of the Thirteenth International Conference on Machine Learning, Bari, Italy, 3–6 July 1996; pp. 148–156.
54. Breiman, L. Randomizing Outputs to Increase Prediction Accuracy. *Mach. Learn.* **2000**, *40*, 229–242. [CrossRef]
55. Bauer, E.; Kohavi, R. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Mach. Learn.* **1999**, *36*, 105–139. [CrossRef]
56. Fix, E.; Hodges, J.L. *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties*; USAF School of Aviation Medicine: Randolph Field, TX, USA, 1951; available online: https://apps.dtic.mil/dtic/tr/fulltext/u2/a800276.pdf (accessed on 28 November 2022).
57. Cover, T.M.; Hart, P.E. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
58. Piryonesi, S.M.; El-Diraby, T.E. 2020. Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems. *J. Transp. Eng. Part B Pavements* **2020**, *146*, 04020022. [CrossRef]
59. Weather Atlas. Mecca, Saudi Arabia—Climate & Monthly Weather Forecast. Available online: https://www.weather-atlas.com/en/saudi-arabia/mecca-climate (accessed on 6 September 2022).
60. Saudi Arabia | Climate Change Knowledge Portal for Development Practitioners and Policy Makers. Available online: https://climateknowledgeportal.worldbank.org/country/saudi-arabia/climate-data-historical (accessed on 6 September 2022).
61. Dekking, F.M.; Kraaikamp, C.; Lopuhaä, H.P.; Meester, L.E. *A Modern Introduction to Probability and Statistics*, 1st ed.; Springer: London, UK, 2005; pp. 234–240. [CrossRef]
62. Jamal, A.; Rahman, M.T.; Al-Ahmadi, H.M.; Mansoor, U. The Dilemma of Road Safety in the Eastern Province of Saudi Arabia: Consequences and Prevention Strategies. *Int. J. Environ. Res. Public Health* **2020**, *17*, 157. [CrossRef] [PubMed]
63. Speiser, J.L.; Wolf, B.J.; Chung, D.; Karvellas, C.J.; Koch, D.G.; Durkalski, V.L. BiMM forest: A random forest method for modeling clustered and longitudinal binary outcomes. *Chemom. Intell. Lab. Syst.* **2019**, *185*, 122–134. [CrossRef] [PubMed]
64. Zeng, S.; Li, L.; Hu, Y.; Luo, L.; Fang, Y. Machine learning approaches for the prediction of postoperative complication risk in liver resection patients. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 371. [CrossRef] [PubMed]
65. Elvik, R.; Christensen, P.; Amundsen, A. *Speed and Road Accidents. An Evaluation of the Power Model. TØI Report 740/2004*; Institute of Transport Economics (TØI): Oslo, Norway, 2004.
66. Hill, C. The Pros and Cons of Increasing Speed Limits. The National, 2010. Available online: https://www.thenationalnews.com/opinion/feedback/the-pros-and-cons-of-increasing-speed-limits-1.508324 (accessed on 15 August 2022).
67. U.S. DOT/FHWA. *Effects of Raising and Lowering Speed Limits*; Report No. FHWA-RD-92-084; U.S. Department of Transportation, Federal Highway Administration: Washington, DC, USA, 1992; Available online: https://www.ibiblio.org/rdu/sl-irrel/index.html (accessed on 15 August 2022).
68. Zhang, Z. Estimating the Optimal Cutoff Point for Logistic Regression. University of Texas at Al Paso. 2018. Open Access Theses & Dissertations. 1565. Available online: https://digitalcommons.utep.edu/open_etd/1565 (accessed on 15 August 2022).
69. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*, 1st ed.; Springer Science and Business Media, Springer: New York, NY, USA, 2021; pp. 133–139. ISBN 978-1-4614-7137-7.
70. Bener, A.; Özkan, T.; Lajunen, T. The Driver Behaviour Questionnaire in Arab Gulf countries: Qatar and United Arab Emirates. *Accid. Anal. Prev.* **2008**, *40*, 1411–1417. [CrossRef] [PubMed]
71. Blows, S.; Ameratunga, S.; Ivers, R.Q.; Lo, S.K.; Norton, R. Risky driving habits and motor vehicle driver injury. *Accid. Anal. Prev.* **2005**, *37*, 619–624. [CrossRef] [PubMed]
72. El Bcheraoui, C.; Basulaiman, M.; Tuffaha, M.; Daoud, F.; Robinson, M.; Jaber, S.; Mikhitarian, S.; Wilson, S.; Memish, Z.A.; Al Saeedi, M.; et al. Get a License, Buckle up, and Slow down: Risky Driving Patterns among Saudis. *Traffic Inj. Prev.* **2015**, *16*, 587–592. [CrossRef] [PubMed]
73. Fergusson, D.; Swain-Campbell, N.; Horwood, J. Risky Driving Behaviour in Young People: Prevalence, Personal Characteristics and Traffic Accidents. *Aust. N. Z. J. Public Health* **2003**, *27*, 337–342. [CrossRef]
74. Lajunen, T.; Gaygisiz, E. Born to Be a Risky Driver? The Relationship Between Cloninger's Temperament and Character Traits and Risky Driving. *Front. Psychol.* **2022**, *113*, 867396. [CrossRef]
75. Sohrabivafa, M.; Tosang, M.A.; Zadeh, S.Z.M.; Goodarzi, E.; Asadi, Z.S.; Alikhani, A.; Khazaei, S.; Dehghani, S.L.; Beiranvand, R.; Khazaei, Z. Prevalence of Risky Behaviors and Related Factors among Students of Dezful. *Iran. J. Psychiatry* **2017**, *12*, 188–193.
76. Tarlochan, F.; Izham, M.; Ibrahim, M.; Gaben, B. Understanding Traffic Accidents among Young Drivers in Qatar. *Int. J. Environ. Res. Public Health* **2022**, *19*, 514. [CrossRef]

77. Al-Tit, A.A.; Dhaou, I.B.; Albejaidi, F.M.; Alshitawi, M.S. Traffic Safety Factors in the Qassim Region of Saudi Arabia. *SAGE Open* **2020**, *10*, 2158244020919500. [CrossRef]
78. Compton, R.P.; Ellison-Potter, P. *Teen Driver Crashes: A Report to Congress (No. DOT-HS-811-005)*; The United States, National Highway Traffic Safety Administration: Washington, DC, USA, 2008.
79. Hassan, H. Examining the factors associated with the involvement of the Saudi' young drivers in at-fault crashes: A survey-based study. In Proceedings of the Transportation Research Board 93rd Annual Meeting, Washington, DC, USA, 12–16 January 2014; Available online: https://trid.trb.org/view/1287489 (accessed on 15 August 2022).
80. Miniño, A. Mortality among Teenagers Aged 12–19 Years: United States, 1999–2006. *NCHS Data Brief.* **2010**, *37*, 1–8.
81. Ministry of Interior (MOI), 2021. Traffic Violations and Penalties. Available online: https://www.moi.gov.sa (accessed on 23 October 2022).
82. Al-Wathinani, A.M.; Schwebel, D.C.; Al-Nasser, A.H.; Alrugaib, A.K.; Al-Suwaidan, H.I.; Al-Rowais, S.S.; AlZahrani, A.N.; Abushryei, R.H.; Mobrad, A.M.; Alhazmi, R.A.; et al. The Prevalence of Risky Driving Habits in Riyadh, Saudi Arabia. *Sustainability* **2021**, *13*, 7338. [CrossRef]
83. Goniewicz, K.M.; Goniewicz, W.P.; Fiedor, P. Road Accident Rates: Strategies and Programmes for Improving Road Traffic Safety. *Eur. J. Trauma Emerg. Surg.* **2016**, *42*, 433–438. [CrossRef] [PubMed]
84. Jadaan, K.; Almatawah, J. A Review of Strategies to Promote Road Safety in Rich Developing Countries: The GC Countries Experience. *Int. J. Eng. Res. App.* **2016**, *6*, 12–17.
85. Ponnaluri, R.V. Road Traffic Crashes and Risk Groups in India: Analysis, Interpretations, and Prevention Strategies. *IATSS Res.* **2012**, *35*, 104–110. [CrossRef]
86. Prabhakharan, P.; Molesworth, B.R.C. Repairing Faulty Scripts to Reduce Speeding Behaviour in Young Drivers. *Accid. Anal. Prev.* **2011**, *43*, 1696–1702. [CrossRef]
87. Shaaban, K. Assessment of Drivers' Perceptions of Various Police Enforcement Strategies and Associated Penalties and Rewards. *J. Adv. Transp.* **2017**, *2017*, 5169176. [CrossRef]
88. Sharma, B.R. Road traffic injuries: A major global public health crisis. *Public Health* **2008**, *122*, 1399–1406. [CrossRef]
89. Zwetsloot, G.I.J.M.; Kines, P.; Wybo, J.-L.; Ruotsala, R.; Drupsteen, L.; Bezemer, R.A. Zero Accident Vision based strategies in organisations: Innovative perspectives. *Saf. Sci.* **2017**, *91*, 260–268. [CrossRef]
90. Tougwa, F.N. A Review of Highways Geometric Design to Ensure Road Health and Safety. *Int. J. Sci. Eng. Sci.* **2021**, *5*, 16–26.
91. Principles for Safe Road Design. Institute for Road Safety Research (SWOV) Fact Sheet. Available online: https://swov.nl/en/fact-sheet/principles-safe-road-design (accessed on 25 November 2022).
92. Work Zone Data: At A Glance. National Workzone Safety Information Clearing House. Available online: https://workzonesafety.org/work-zone-data/ (accessed on 26 November 2020).
93. Alsultan, S.M.; Alqahtani, F.K.; Alkahtani, K.F. Health and Safety in Temporary Work Zone Road Construction Project in Saudi Arabia: Risks and Solutions. *Int. J. Environ. Res. Public Health* **2022**, *19*, 10627. [CrossRef]
94. Road Safety Audits (RSA). US Department of Transportation, Federal Highway Administration. Available online: https://highways.dot.gov/safety/data-analysis-tools/systemic/road-safety-audits-rsa (accessed on 26 November 2022).
95. Anandraj, A.; Vijayabaskaran, S. Evaluation of Road Safety Audit on Existing Highway by EmpiricalBabkov's Method. *Saudi J. Civ. Eng.* **2020**, *46*, 85–91. [CrossRef]
96. Theeuwes, J.; Godthelp, H. Self-explaining Roads. *Saf. Sci.* **1995**, *19*, 217–225. [CrossRef]
97. Theeuwes, J. Self-explaining roads: Subjective categorization of road environments. In *Vision in Vehicle VI*; Gale, A., Ed.; Elsevier Science: Amsterdam, The Netherlands, 1998; pp. 279–288.
98. Self Explaining Roads. Mobility & Transport—Road Safety. Available online: https://road-safety.transport.ec.europa.eu/statistics-and-analysis/statistics-and-analysis-archive/roads/self-explaining-roads_en (accessed on 27 November 2022).
99. Theeuwes, J. Self-explaining roads: What does visual cognition tell us about designing safer roads? *Cogn. Res.* **2021**, *6*, 15. [CrossRef] [PubMed]
100. La Torre, F. Forgiving Roadsides Design Guide Report. CEDR, Conference of European Directors of Roads. 2012. Available online: https://www.cedr.eu/download/Publications/2013/T10_Forgiving_roadsides.pdf (accessed on 27 November 2022).
101. Saleh, P.; La Torre, F.; Nitsche, P.; Helfert, M. A Guidance Document for the Implementation of the CEDR Forgiving Roadsides Report. National Roads Authority, 2013. Available online: https://www.tii.ie/tii-library/road-safety/Road%20Safety%20Research/Forgiving-Roadsides.pdf (accessed on 27 November 2022).
102. De Winter, J.C.F. Predicting Self-Reported Violations among Novice License Drivers Using Pre-License Simulator Measures. *Accid. Anal. Prev.* **2013**, *52*, 71–79. [CrossRef]
103. Mohamed, M.; Bromfield, N.F. Attitudes, Driving Behavior, and Accident Involvement among Young Male Drivers in Saudi Arabia. *Transp. Res. Part F Traffic Psychol. Behav.* **2017**, *47*, 59–71. [CrossRef]
104. Bianchi, A.; Summala, H. The 'Genetics' of Driving Behavior: Parents' Driving Style Predicts Their Children's Driving Style. *Accid. Anal. Prev.* **2004**, *36*, 655–659. [CrossRef] [PubMed]
105. Alonso, F.; Faus, M.; Fernández, C.; Useche, S.A. "Where Have I Heard It?" Assessing the Recall of Traffic Safety Campaigns in the Dominican Republic. *Energies* **2021**, *14*, 5792. [CrossRef]
106. Zatoński, M.; Herbeć, A. Are mass media campaigns effective in reducing drinking and driving? Systematic review—An update. *J. Health Inequal.* **2016**, *2*, 52–60. [CrossRef]

107. Mandal, R.; Mandal, A.; Dutta, S.; Alam, M.; Saha, S.; Nandi, S. Framework of intelligent transportation system: A survey. In *Proceedings of International Conference on Frontiers in Computing and Systems*; Basu, S., Kole, D.K., Maji, A.K., Plewczynski, D., Bhattacharjee, D., Eds.; Springer: Berlin/Heidelberg, Germany, 2022; Volume 404, pp. 93–108. [CrossRef]
108. Capali, B. Intelligent Transportation Systems Architecture: Recommendation for K-AUS. *J. El-Cezeri Sci. Eng.* **2022**. [CrossRef]
109. Narayanaswami, S. Intelligent transportation systems: The state of the art in railways. In *Handbook of Research on Emerging Innovations in Rail Transportation Engineering*; Rai, B., Ed.; IGI Global: Hershey, PA, USA, 2016; pp. 387–404. [CrossRef]
110. Morris, B.; Trivedi, M. Real-time video based highway traffic measurement and performance monitoring. In Proceedings of the 2007 IEEE Intelligent Transportation Systems Conference, Bellevue, WA, USA, 30 September–3 October 2007; pp. 59–64. [CrossRef]
111. Fakhrmoosavi, F.; Saedi, R.; Talebpour, A.; Zockaie, A. Impacts of Connected and Autonomous Vehicles on Traffic Flow with Heterogeneous Drivers Spatially Distributed over Large-Scale Networks. *Trans. Res. Rec.* **2020**, *2674*, 817–830. [CrossRef]
112. Chen, M.-X. Design and implementation of vehicle navigation systems. In *Telematics Communication Technologies and Vehicular Networks: Wireless Architectures and Applications*; Huang, C.-M., Chen, Y.-S., Eds.; IGI Global: Hershey, PA, USA, 2010; pp. 131–143. [CrossRef]
113. Tijerina, L.; Johnston, S.; Parmer, E.; Winterbottom, M. *Driver Distraction with Wireless Telecommunications and Route Guidance Systems, DOT HS 809-069*; National Highway Traffic Safety Administration: Washington, DC, USA, 2000; pp. 1–90. Available online: https://rosap.ntl.bts.gov/view/dot/14090 (accessed on 23 October 2022).
114. Uang, S.; Hwang, S. Effects on driving behavior of congestion information and of scale of in-vehicle navigation systems. *Transp. Res. Part C Emerg. Technol.* **2003**, *11*, 423–438. [CrossRef]
115. Eby, D.W.; Kostyniuk, L.P. An on-the-road comparison of in-vehicle navigation assistance systems. *Hum. Factors* **1999**, *41*, 295–311. [CrossRef]
116. Oei, H.L.; Polak, P.H. Intelligent Speed Adaptation (ISA) and Road Safety. *IATSS Res.* **2002**, *26*, 45–51. [CrossRef]
117. Schroten, A.; van Grinsven, A.; Tol, E.; Leestemaker, L.; Schackmann, P.P.; Vonk-Noordegraaf, D.; Van Meijeren, J.; Kalisvaart, S. *The Impact of Emerging Technologies on the Transport System*; Research for TRAN Committee; European Parliament, Policy Department for Structural and Cohesion Policies: Brussels, Belgium, 2020.
118. Furlan, A.D.; Kajaks, T.; Tiong, M.; Lavallière, M.; Campos, J.L.; Babineau, J.; Haghzare, S.; Ma, T.; Vrkljan, N. Advanced vehicle technologies and road safety: A scoping review of the evidence. *Accid. Anal. Prev.* **2020**, *147*, 105741. [CrossRef] [PubMed]
119. Liu, J.; Jayakumar, P.; Stein, J.L.; Ersal, T. Combined speed and steering control in high-speed autonomous ground vehicles for obstacle avoidance using model predictive control. *IEEE Trans. Veh. Technol.* **2017**, *66*, 8746–8763. [CrossRef]
120. Lin, F.; Wang, K.; Zhao, Y.; Wang, S. Integrated Avoid Collision Control of Autonomous Vehicle Based on Trajectory Re-Planning and V2V Information Interaction. *Sensors* **2020**, *20*, 1079. [CrossRef]
121. Parseh, M.; Asplund, F. New needs to consider during accident analysis: Implications of autonomous vehicles with collision reconfiguration systems. *Accid. Anal. Prev.* **2022**, *173*, 106704. [CrossRef]
122. Plug & Play. Available online: https://www.mastrack.com/plugNplay (accessed on 29 November 2022).
123. In-Vehicle Monitoring Systems (IVMS). Available online: https://www.digitalmatter.com/applications/features/in-vehicle-monitoring-systems/ (accessed on 29 November 2022).
124. In-Vehicle Monitoring Systems Improve Driving Skills. Available online: https://www.shell.com/business-customers/shell-fleet-solutions/health-security-safety-and-the-environment/in-vehicle-monitoring-systems-can-help-everyone-to-improve-their-driving-skills.html (accessed on 29 November 2022).
125. Survey: Consumers Are More Ready to Use Telematics than in Years Past. Available online: https://news.nationwide.com/120920-survey-consumers-are-more-ready-to-use-telematics-than-in-years-past/ (accessed on 29 November 2022).
126. Al-Sharif, D.T. Saudi Traffic Laws Focus on Correcting Behavior. Arab News. Available online: https://arab.news/6y9bc (accessed on 24 October 2022).
127. Safe Driver Incentive Plan. Available online: https://www.ncdoi.gov/consumers/auto-and-vehicle-insurance/safe-driver-incentive-plan (accessed on 29 November 2022).
128. 5 Years after Its Launch Summary of the Achievements of the Kingdom's Vision 2030. Available online: https://www.media.gov.sa/en/news/5001 (accessed on 29 October 2022).
129. 33% Decrease in the Number of Traffic Accident Deaths and 25% in Injuries and Accidents during 2018. Available online: https://mot.gov.sa/ar/MediaCenter/News/Pages/new986.aspx (accessed on 29 October 2022).
130. Musumeci, F.; Rottondi, C.; Nag, A.; Macaluso, I.; Zibar, D.; Ruffini, M.; Tornatore, M. An Overview on Application of Machine Learning Techniques in Optical Networks. *IEEE Commun. Surv. Tutor.* **2018**, *21*, 1383–1408. [CrossRef]
131. Taunk, K.; De, S.; Verma, S.; Swetapadma, A. A brief review of nearest neighbor algorithm for learning and classification. In Proceedings of the 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 15–17 May 2019; pp. 1255–1260. [CrossRef]