Predicting the occurrence of deer-vehicle collisions based on environmental factors

Maria Fernanda Pina-Mousseau

School of Business & Economics

State University of New York at Plattsburgh

Plattsburgh, NY

October 22nd, 2022

Abstract

Collisions between motor vehicles and large vertebrates are a severe problem in the state of New York and other states across the country. Specifically, deer-vehicle collisions are the most common accident among these types. Approximately 1.5 million collisions occur every year. The consequences are 1.3 million dead deer, 29,000 human injuries, 200 human fatalities, and one billion dollars in property damage nationwide. The objective of this study is to predict the occurrence of a deer-vehicle collision in the state of New York based on the deer-vehicle collision data from the NY Department of transportation. The model will predict whether a crash against a deer is more likely to occur using predictors such as weather conditions, light conditions, road surface, road characteristics, and the month of the year. Since the data contains only accidents (positive samples), negative samples have been generated based on the parameters of the positive ones. Additionally, the number of negative samples is three times the number of positive ones since the accident is a rare event. The target variable is categorical, so logistic regression, k-Nearest Neighbors, and classification trees will be evaluated. Ultimately, the information will be shared with the New York State Department of Transportation to determine if there are additional mitigation steps to be taken to prevent accidents of this nature. In the future, similar models can be developed for applications in Artificial Intelligence that can prevent accidents altogether.

Contents

1. Introduction

Collisions between motor vehicles and large vertebrates are a severe problem in the state of New York and other states across the country. This situation is one of the consequences of the increase in anthropogenic transformation of the landscape. After a series of studies, experts determined that 50% of the land area in the United States is around 300 meters away from a road (Bisonnette & Rosa, 2012). This statement means that animals' proximity to roads is minimal. Additionally, the population of *Odocoileus sp.* has increased significantly due to wildlife management practices and changes in habitat structure (Gonser et al., 2009).

Several species can cause these collisions: moose (Alces alces), elk (Cervus elaphus), and two species of deer (Odocoileus spp.), The latter is the one that causes most of these motor vehicle accidents. The issue has increased in the US causing ecological damage, economic loss, and injuries (Bisonette et. al, 2008). On average, 1.5 million collisions between motor vehicles and deer occur annually in the United States, causing 29,000 injuries, 200 fatalities, 1.3 million deer fatalities, and over a billion dollars' worth of property damage (Mastro et. al 2008). In New York State, there have been almost 600,000 collisions between motor vehicles and deer since 2007. There have been plenty of strategies to mitigate the problem such as deer fencing, sex distribution among deer, miles of roadway, reflectors, etc but these measures did not have a big impact (Gonser et al., 2009 & Reeve et al. 1993). Several circumstances increase the possibility of colliding against a deer (Mastro et.al 2008) such as:

- Lights conditions: 80-95% of the collisions occur between sunset and sunrise. The lack of light on the road diminishes the visibility of the driver, their immediate reaction (Mastro et.al 2008).
- Weather conditions: Some studies suggested that when the weather is clear, the possibility of an accident happening increases (Mastro et.al 2008).
- Speed limit: Most accidents occur when the speed limit increases since the time to stop the car could be longer (Mastro et.al 2008).

The objective of this study is to predict the occurrence of a deer-vehicle collision based on different weather, light, and road surface conditions. Ultimately, the purpose is to inform the population of the state of New York to create awareness of the causes of these unforeseen accidents.

2. Literature

As one of the main causes of accidents in the country, especially in rural areas, many studies have been conducted on Deer-Vehicle collisions. The topics range from predicting the event to different approaches to prevent it (Ujvari et al. 1998). Many studies such as the one made by Bisonnete & Kassar (2008) have analyzed the possible hotspots of the collisions in a way to explain the occurrence and potentially prevent it. The study found & Boyce 2011 built a model to predict the collisions in Edmonton, Canada based on the speed limit, forest vegetation, landscape heterogeneity, and distance to water. In a similar approach, we seek to predict the occurrence of these accidents in New York State by using a classification model.

This case study seeks to predict the occurrence of a collision against a deer, based on several predictors: light conditions, weather, road characteristics, and road surface. The dataset only contains positive samples (accidents that did occur). However, to build a proper classification model the dataset must also contain negative samples (non-accidents). These negative samples can be generated using a methodology proposed by *Yuan et. al* (2007). Using the positive samples from the current dataset it is possible to randomly change one of the following features of an observation to create a negative one:

- Hour: Picking a random time between 00:00 and 23:59 except for the time of the dataset. Changing the hour feature may cause the change of other features related to time in a day such as light conditions.
- Day: Picking a random day between 1 and 365 except for the day of the dataset. Changing the day feature will cause the change of features related to the day, including weather and time related features.
- Road Segment: Picking a random segment of the road except for the segment of the dataset.

3. Exploratory Data Analysis

3.1 Data shape

Table 1. Description of each feature of the deer-vehicle collision dataset provided by the New York State department of transportation

Feature	Description	Values
Case Number	Case number assigned by the DMV	1
Case Year	Year of the accident	1
Crash Date	Date of the Accident	1
Crash Time	Time of the Accident	1
Crash Time Formatted	Time of the Accident	1
Posted Speed	Speed limit of the road	1
Number of fatalities	Number of fatalities in the accident	1
Number of injuries	Number of injuries in the accident	1
Number of injuries	Number of serious injuries in the accident	1
Number of other injuries	Number of other applicable injuries in the accident	1
Number of Vehicles	Number of Vehicles involved	1
UTM Easting	Coordinates of the occurrence of the accident	1
UTM Northing	Coordinates of the occurrence of the accident	1
Light Conditions	Light condition of the road at the time of the accident	7
Roadway Access Control Code	Type of access control	7
Weather Conditions	Weather condition of the road at the time of the accident	8
Roadway Characteristics	The road character at the location of the crash	9
Road Surface Conditions	Road surface condition at the time of the crash	10
DMV Crash Classification Code	Severity of the Accident	11
Collision Type	Classification of the Collision	13
Traffic Control Code	Traffic control present at the crash location.	23
Crash Type	Classification of the Crash	44

Table 1. describes the different features in the dataset, such as basic information like a case number, year, date, and time of each accident. Additionally, it describes the number of injured people involved in the accident and whether it was a fatality or a severe injury, the location in UTM coordinates, and different traffic parameters. The crash classification column is

the same for every observation since the study only focuses on deer-vehicle collisions. The dataset contains 593491 rows and 23 columns. The New York State Department of Transportation provided the dataset which includes collisions from 2007 to September 2022.

Table 1. also shows the light conditions feature contains the following categories: daylight, dawn, dusk, dark road-lighted, dark-road unlighted, and unknown. Similarly, weather conditions have the following categories: other, clear, cloudy, rain, snow, freezing rain, fog, and unknown. Road characteristics contain the following categories: straight and level, straight /grade, straight in hillcrest, curve and level, curve and grade, curve, and hillcrest and unknown. Finally, road conditions contain the following categories: other, dry, wet, muddy, snow/ice, slush, flooded, and unknown.

3.2 Data Anomalies

The data set in this case study is big and there are several challenges to overcome to properly analyze the data and create a successful predictive model.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 593491 entries, 0 to 593490
Data columns (total 23 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   CaseNumber                  593491 non-null  object
 1   CaseYear                    593491 non-null  int64
 2   CrashDate                   593491 non-null  object
 3   CrashTime                   587333 non-null  float64
 4   CrashTimeFormatted          587333 non-null  object
 5   DmvCrashClassificationCode  593491 non-null  int64
 6   PostedSpeedLimit            0 non-null       float64
 7   NumberOfFatalities          593491 non-null  int64
 8   NumberOfInjuries            593491 non-null  int64
 9   NumberOfSeriousInjuries     593491 non-null  int64
 10  NumberOfOtherInjuries       593491 non-null  int64
 11  NumberOfVehicles            593491 non-null  int64
 12  UTMEasting                  588056 non-null  float64
 13  UTMNorthing                 588056 non-null  float64
 14  CrashTypeCde                593491 non-null  int64
 15  CollisionTypeCde            593491 non-null  int64
 16  LightConditionCde           593491 non-null  int64
 17  WeatherConditionCde         593491 non-null  int64
 18  RoadwayAccessControlCde     589785 non-null  float64
 19  RoadwayCharacteristicsCde   593491 non-null  int64
 20  RoadSurfaceConditionsCde    593491 non-null  int64
 21  TrafficControlCde           593491 non-null  int64
 22  Unnamed: 22                 0 non-null       float64
dtypes: float64(6), int64(14), object(3)
memory usage: 104.1+ MB
```

Figure 1. Output from Jupyter notebooks showing the missing values and the data types for each feature of the deer-vehicle collisions dataset.

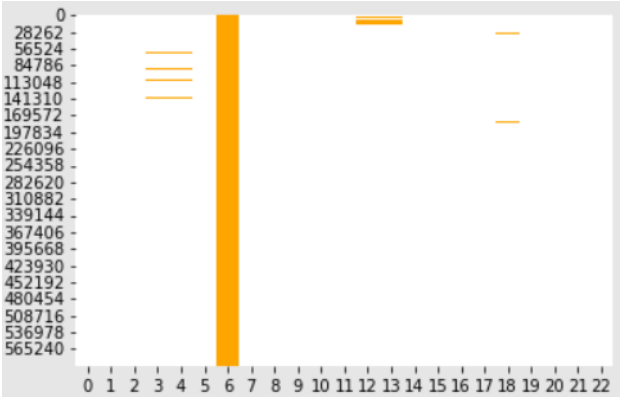

Figure 2. Graph of missing values where orange means null values for a particular feature

As we can see in **Fig. 1** and **Fig. 2** The data contains columns with null values or without any observations such as *Speed limit*. We could probably drop this variable since it does not contain any observations. Other variables contain missing values such as UTM Easting, UTM Northing, and Date in this case since the data is very robust, we can opt to eliminate those rows. Finally, most data types in the dataset are categorical, even though it says most of them are integers the department of transportation changed the category for a number for easier manipulation.

3.3 Data Description

Since most of the data we are using in this study is categorical and part of the outcome variable needs to be created, most of the graphics presented below account for the number of accidents per different conditions.

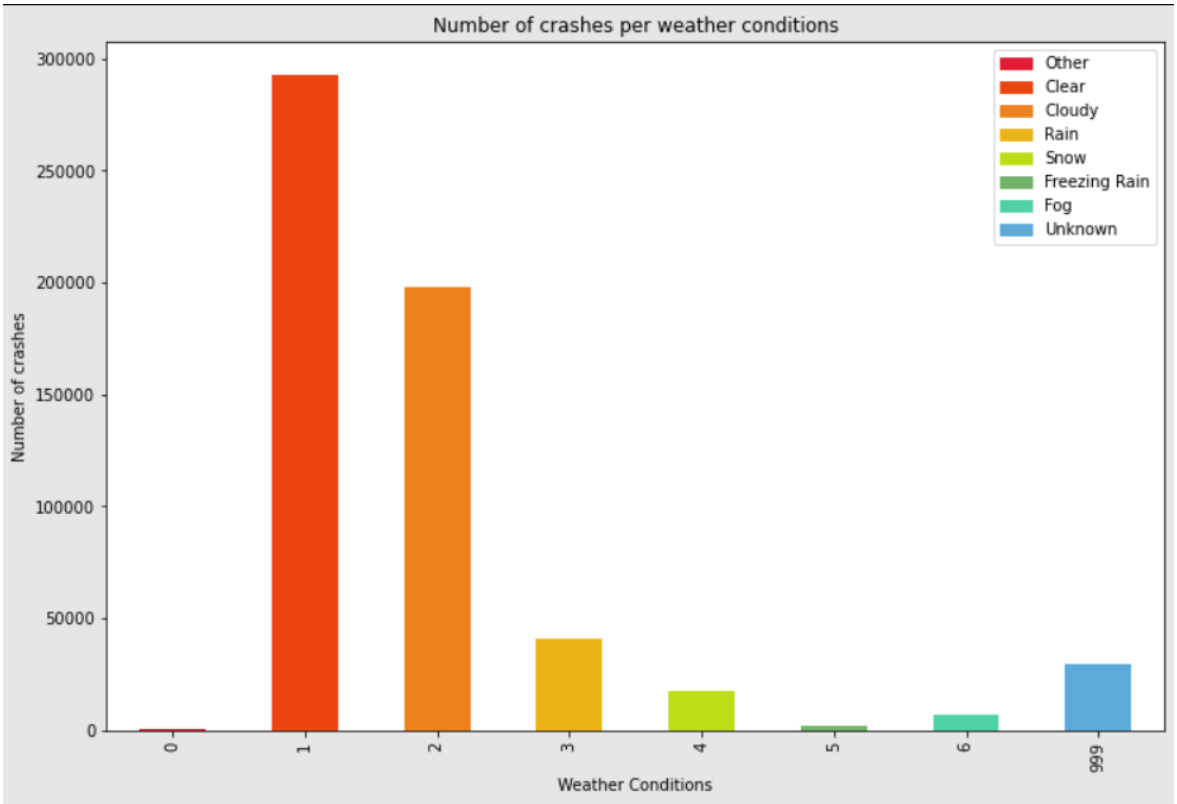

Figure 3. Number of collisions per weather conditions

In **Fig. 3.** We can see that under the weather condition of "Clear" there are significantly more collisions against deer than under any other weather conditions. Additionally, it is important to explain that the category "Other" is not defined in the data dictionary, however it could be a mix of other categories for example rainy and cloudy or snow and cloudy. As we can see there are not a lot of observations under this category since the bar is close to 0.

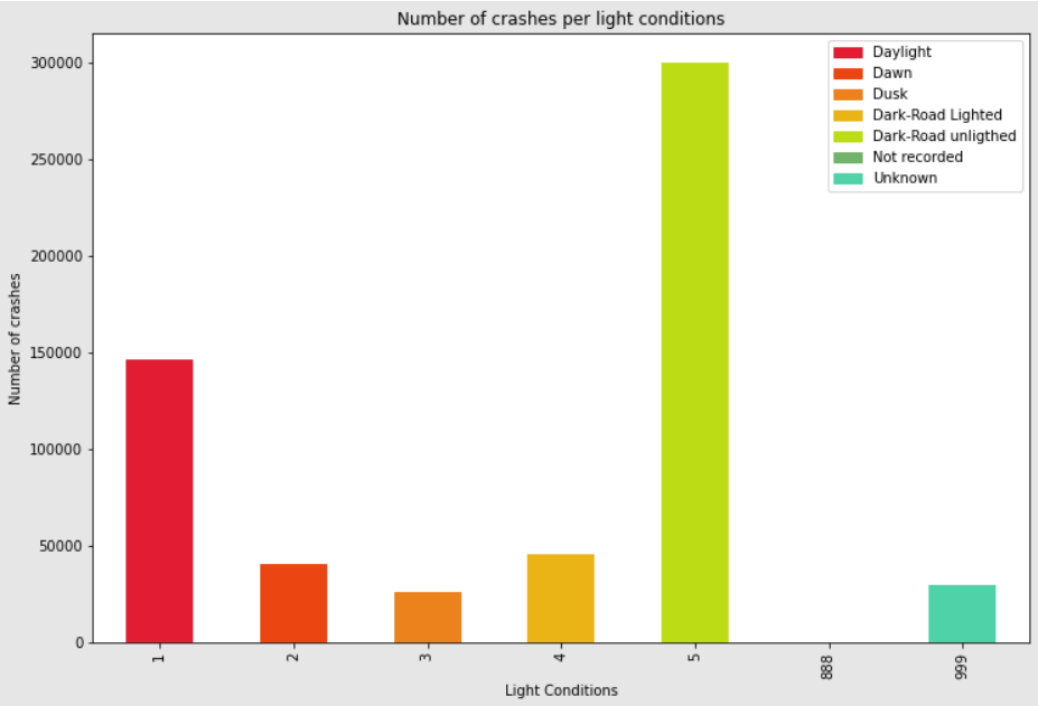

Figure 4. Number of collisions per light conditions

In **Fig. 4.** We can see that under the light condition of "Dark-Road Unlighted" a significant number of collisions against deer occur compared to any other conditions. The category not recorded exists, but it does not have any observation. We could probably drop this category since it does not add any information.

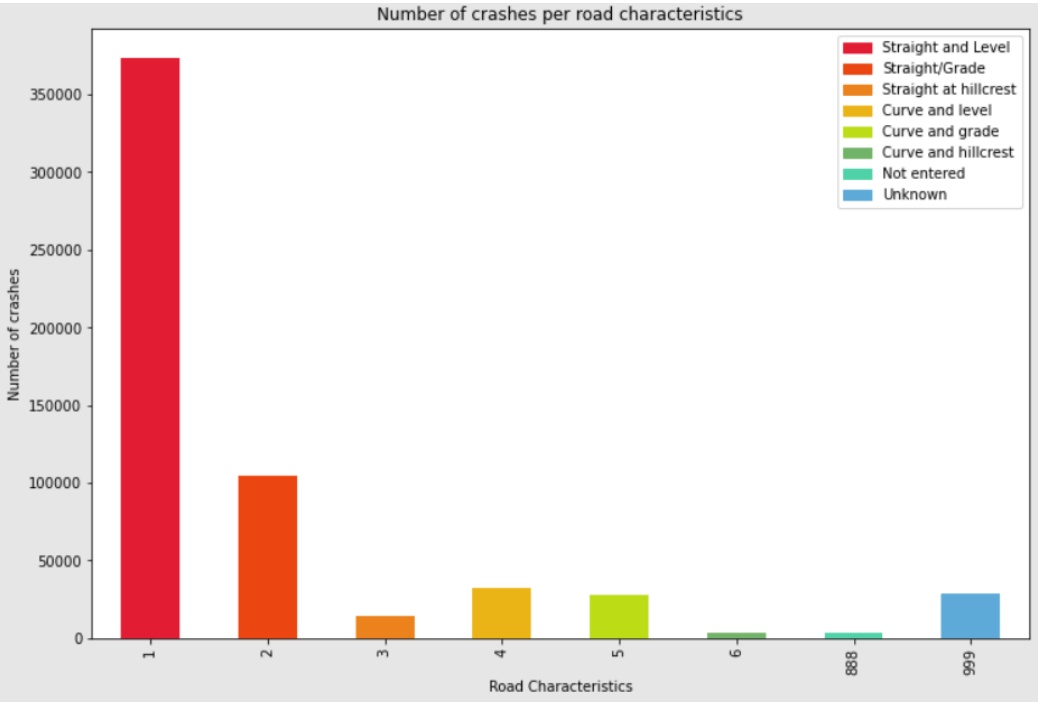

Figure 5. Number of collisions per road characteristics

Additionally in **Fig. 5.** we can observe that under “Straight and Level” most of the collisions occurred compared to the other categories. We also observe that there are values under categories “Not recorded” or “Unknown”.

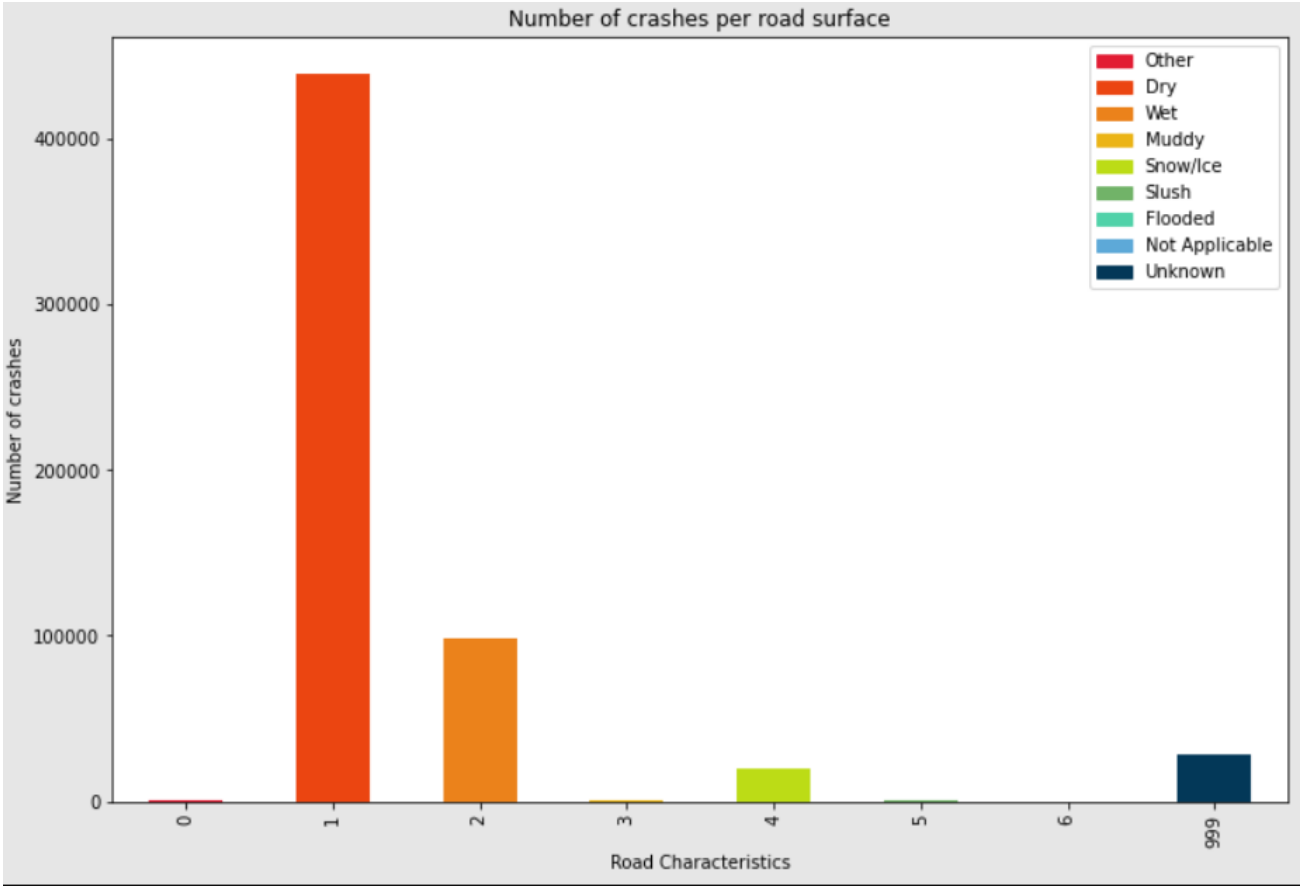

Figure 6. Number of collisions per road surface

On **Fig. 6.** we can observe that under “Dry” most of the collisions occurred compared to the other categories. There is also a category of “Other” (0 on the x-axis) which could be a combination of road surface conditions such as wet and snow. This category has a small number of observations.

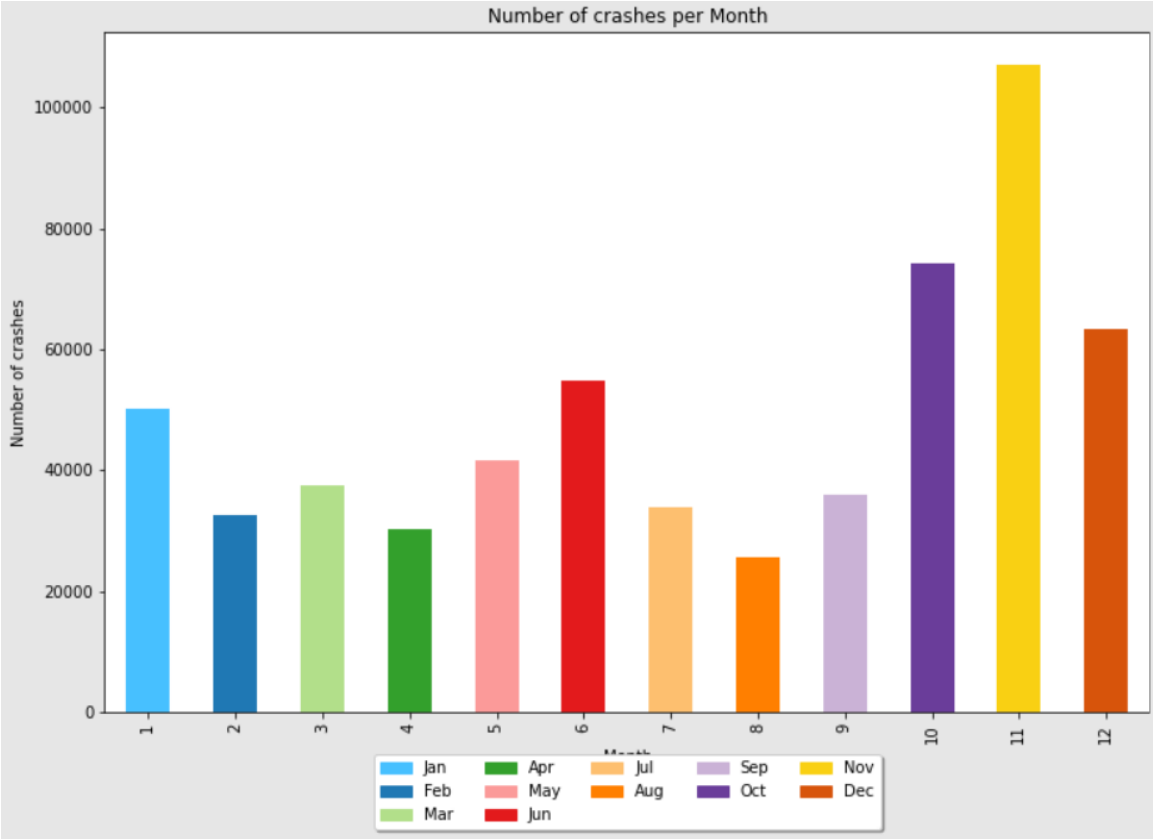

Figure 7. Number of collisions per month

Finally, on **Fig. 7.** We can observe that the number of accidents increases significantly between the months of October and December, this could be attributed to the start of the deer mating season (Parra *et al.*, 2014) or the decrease in daylight hours.

4. Data Cleaning

The data cleaning process started by eliminating unnecessary features like collision type since all the observations belonged to the deer-vehicle collision type. Additionally, we deleted the UTM fields due to the impossibility of converting the values to latitude and longitude. Finally, we deleted the features related to the number of injured and the number of severely injured because it makes no sense to include features that are consequences of accidents if we are predicting the occurrence of an accident.

As mentioned above, the dataset only contained accidents or positive samples, it was necessary to create the negative samples to apply the classification models. The negative samples were generated via www.mockaroo.com based on the features of the positive samples. The dataset is purposely imbalanced since accidents are rare events. Based on the methodology of Yuan *et al.*(2017), there are three times more negative samples (non-accidents) than positive samples(accidents). The result of the cleaned dataset is shown in **Figure 8.**

	LightConditionCde	WeatherConditionCde	RoadwayCharacteristicsCde	RoadSurfaceConditionsCde	TrafficControlCde	Month	Day_of_Week	Hour_day	day_of_year	is_accident
0	2	1	1	1	1	5	2	6	123	1
1	1	1	2	1	1	6	6	19	172	1
2	3	6	2	1	1	6	7	4	159	1
3	5	2	5	1	7	11	7	17	320	1
4	5	1	1	1	1	5	1	1	145	1
...	...	...	...	...	...	...	...	...	...	...
19995	1	7	3	6	17	9	7	6	244	0
19996	3	3	6	7	10	12	4	9	343	0
19997	3	2	1	2	9	10	7	5	295	0
19998	6	7	5	2	9	5	6	11	137	0
19999	6	0	2	7	5	1	7	5	6	0

Figure 8. Quick view of the cleaned dataset

4.1 Feature importance

Up to this point in the project, the dataset was cleaned, and we kept the six environmental factors that determine the accidents. The scope of the investigation was narrowed to only external factors on the road. For this reason, the feature selection was not a crucial step since all the features were important as shown in the outputs below.

According to the *Forward elimination* method the most important variables are Traffic Control, Road Surface Conditions, Weather Conditions, Roadway Characteristics, and Month and Light Conditions, also shown in **Fig. 9.**

```
Variables: LightConditionCde, WeatherConditionCde, RoadwayCharacteristicsCde, RoadSurfaceConditionsCde, TrafficControlCde, Month
Start: score=11637.67, constant
Step: score=9526.27, add TrafficControlCde
Step: score=7905.34, add RoadSurfaceConditionsCde
Step: score=2575.00, add WeatherConditionCde
Step: score=1806.45, add RoadwayCharacteristicsCde
Step: score=1749.22, add Month
Step: score=1734.86, add LightConditionCde
Step: score=1734.86, add None
['TrafficControlCde', 'RoadSurfaceConditionsCde', 'WeatherConditionCde', 'RoadwayCharacteristicsCde', 'Month', 'LightConditionCde']
```

Figure 9. Output of the forward elimination

Showing the same results as forward elimination, according to the *stepwise elimination* method the most important variables are Traffic Control, Road Surface Conditions, Weather Conditions, Roadway Characteristics, and Month and Light Conditions, also shown in **Fig. 10.**

```
Variables: LightConditionCde, WeatherConditionCde, RoadwayCharacteristicsCde, RoadSurfaceConditionsCde, TrafficControlCde, Month
Start: score=11637.67, constant
Step: score=9526.27, add TrafficControlCde
Step: score=7905.34, add RoadSurfaceConditionsCde
Step: score=2575.00, add WeatherConditionCde
Step: score=1806.45, add RoadwayCharacteristicsCde
Step: score=1749.22, add Month
Step: score=1734.86, add LightConditionCde
Step: score=1734.86, unchanged None
['TrafficControlCde', 'RoadSurfaceConditionsCde', 'WeatherConditionCde', 'RoadwayCharacteristicsCde', 'Month', 'LightConditionCde']
```

Figure 10. Output of the stepwise elimination

Showing the same results as forward elimination and stepwise elimination, according to the *backward elimination* method the most important variables are Traffic Control, Road Surface Conditions, Weather Conditions, Roadway Characteristics, and Month and Light Conditions, also shown in **Fig. 11.**

```
Variables: LightConditionCde, WeatherConditionCde, RoadwayCharacteristicsCde, RoadSurfaceConditionsCde, TrafficControlCde, Month
Start: score=1734.86
Step: score=1734.86, remove None
['LightConditionCde', 'WeatherConditionCde', 'RoadwayCharacteristicsCde', 'RoadSurfaceConditionsCde', 'TrafficControlCde', 'Month']
```

Figure 11. Output of the backward elimination

4.2 Data split

The data was split into three sub-datasets: Training, validation, and test. Fifty percent of the data was assigned to the training split, twenty-five percent of the data was assigned to validation, and the other twenty-five percent of the data was allocated to testing, as shown in **Fig 12**. The importance of the testing dataset is that it will allow us to compare the different models with new data.

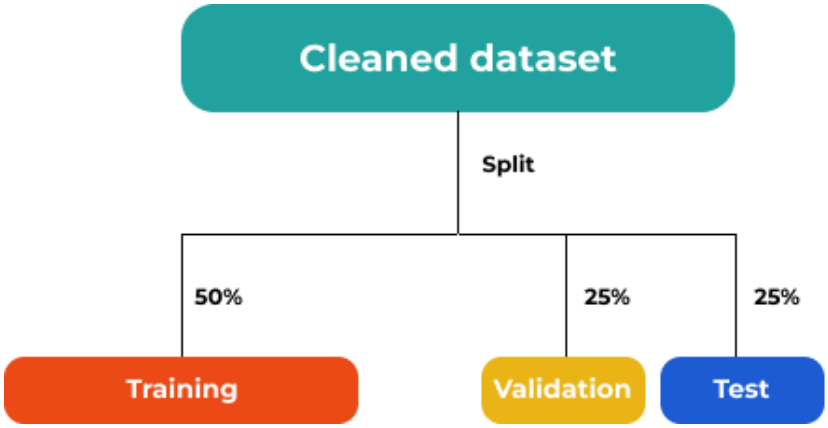

Figure 12. Data Split

5. Classification Models

Since the target variable was *is_accident* and the values were 1 and 0. This would be approached as a classification problem. Hence the following models were implemented:

- Logistic Regression
- K Nearest Neighbors
- Classification trees
- Random Forest

5.1 Logistic Regression

The logistic regression model applied to the dataset yielded very good results. As we can see, in the validation and confusion matrix **Fig 13,** the model obtained 93% accuracy. Additionally, in **Fig 14.** We can see the gain and lift chart where it becomes apparent the improvement of the model against the random selection. It is important to evaluate other performance measures such as recall, precision, and F1 score, as shown in **Fig. 15.**

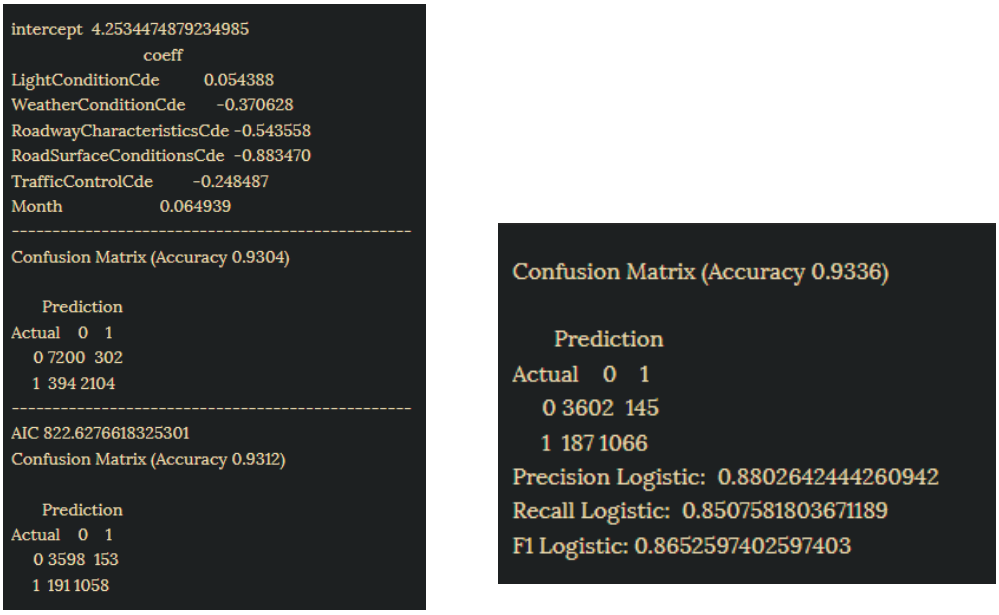

Figure 13. Confusion matrix of logistic regression. A. Output of the logistic regression B. Confusion matrix of the testing dataset.

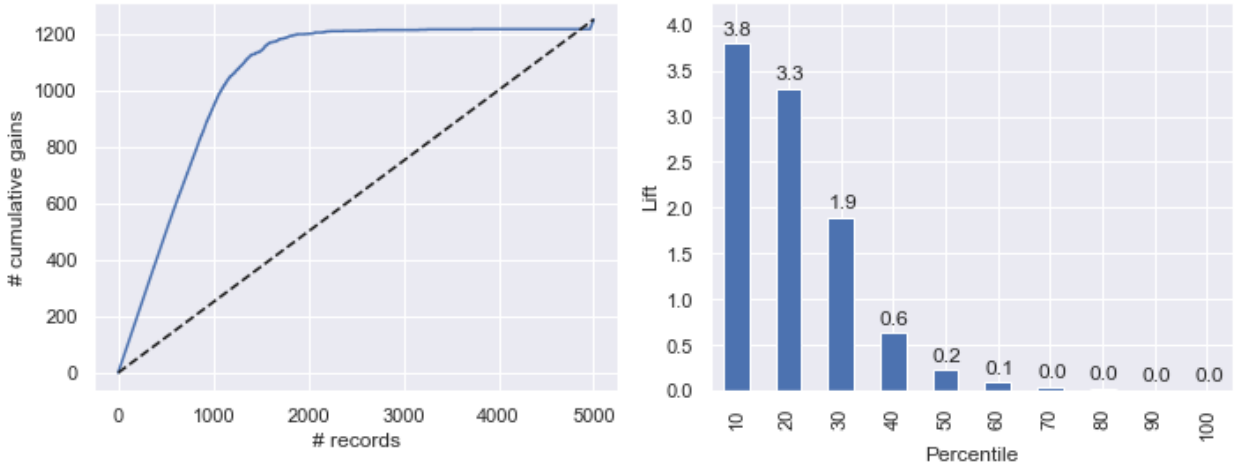

Figure 14. Gains and lift charts for logistic regression

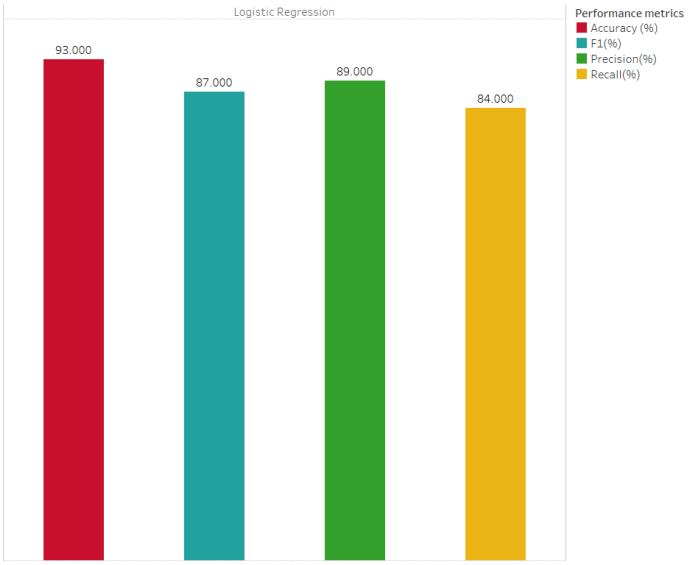

Figure 14. Performance metrics for logistic regression

5.2 K nearest neighbors

KNN is a data drive method. The model consists of classifying the observations based on the similarities with existing observations. To implement this model, it is necessary to normalize the data and in this case, we dummified the variables for easier manipulation. Initially, we selected the number of neighbors to be five, but the optimal number of neighbors is three as shown in **Fig. 16.** The accuracy went up by around 0.5% when selecting the optimal k. We can see that the model obtained 96% accuracy, as shown in **Fig. 17.**

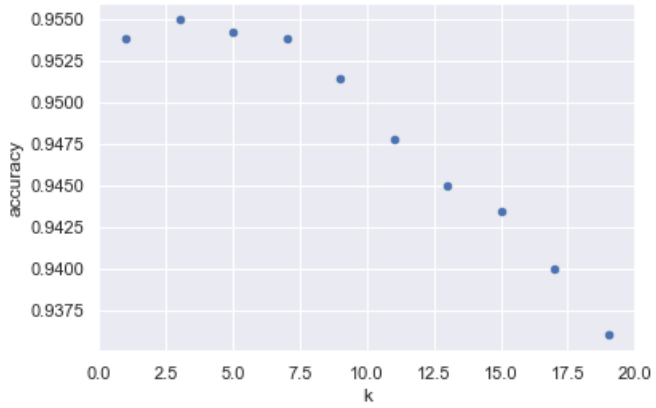

Figure 16. Optimal number k neighbors based on accuracy

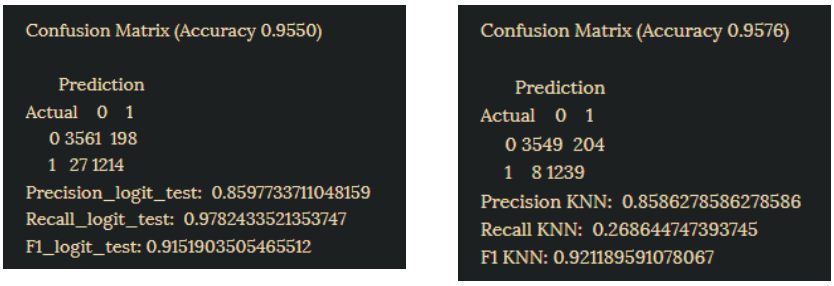

Figure 17. Confusion matrix for KNN

Finally, we can see the performance metrics of the KNN model. As shown in **Fig, 18** the recall was only 29% which means that the model is classifying many accidents as non-accidents (Japkowicz, 2006).

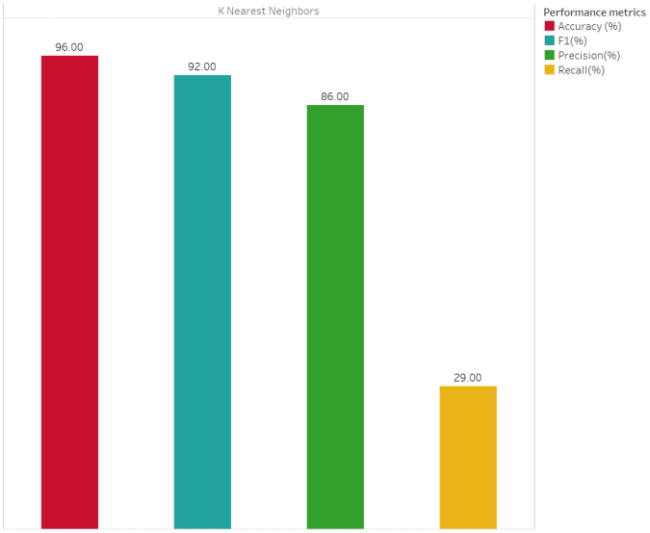

Figure 18. Performance metrics of the KNN model

5.3 Classification tree

The classification tree had the best results as shown in **Fig 19.** We can see that the accuracy was 98% within the validation data and the test data. There is a visual representation of the tree algorithm in **Fig 20.**

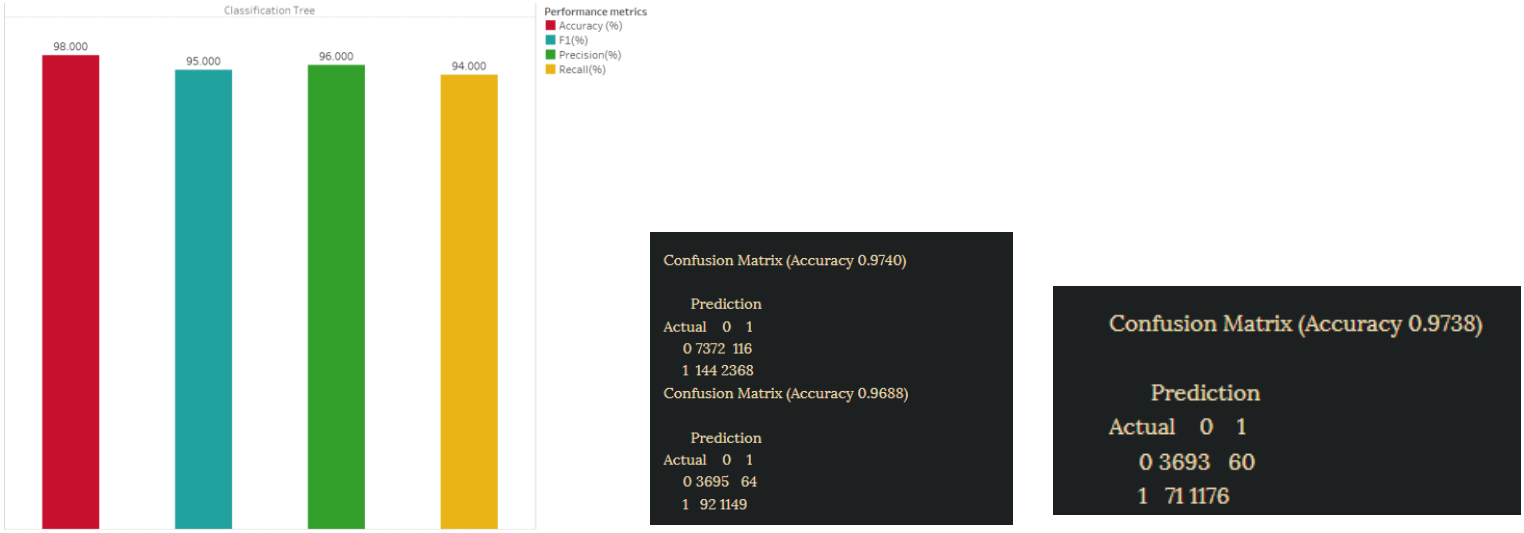

Figure 19. A. Performance metrics of the decision tree model B. Confusion matrix of the validation dataset. C. Confusion matrix of the testing dataset.

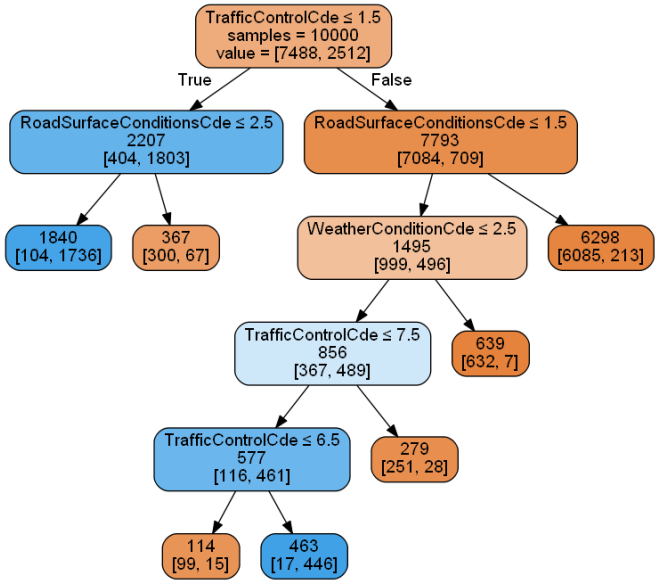

Figure 20. Decision tree algorithm model

5.4

Random Forest

In a similar way, the random forest outperformed KNN and Logistic Regression as shown in **Fig 21.** We can also observe the most important feature for this model in **Fig. 22**

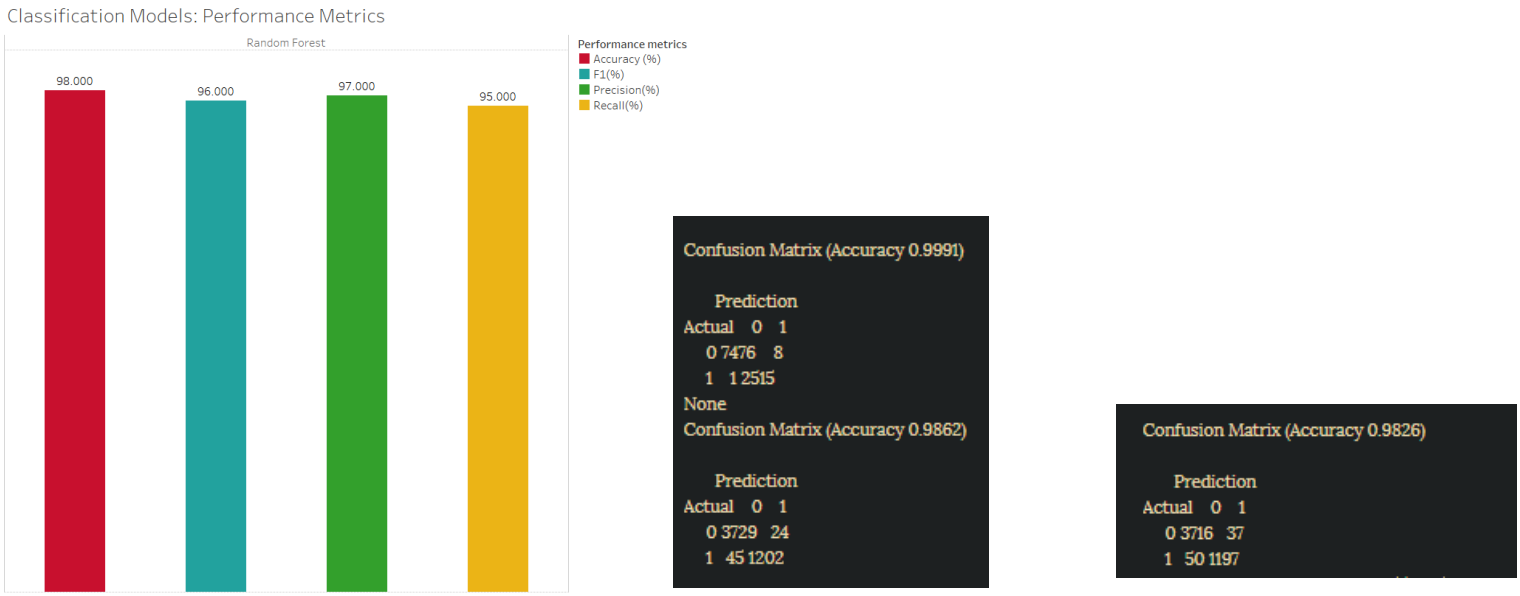

Figure 21. A. Performance metrics of the random forest model B. Confusion matrix of the validation dataset. C. Confusion matrix of the testing dataset.

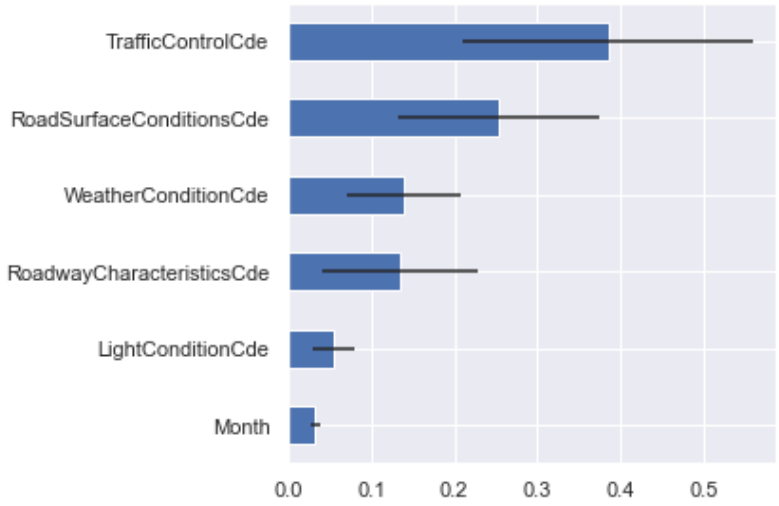

Figure 22. Feature importance graph

Gradient Boosting

In a similar way, the gradient boosting yielded similar results to the random forest with an accuracy of 98%, as seen in **Fig 23.**

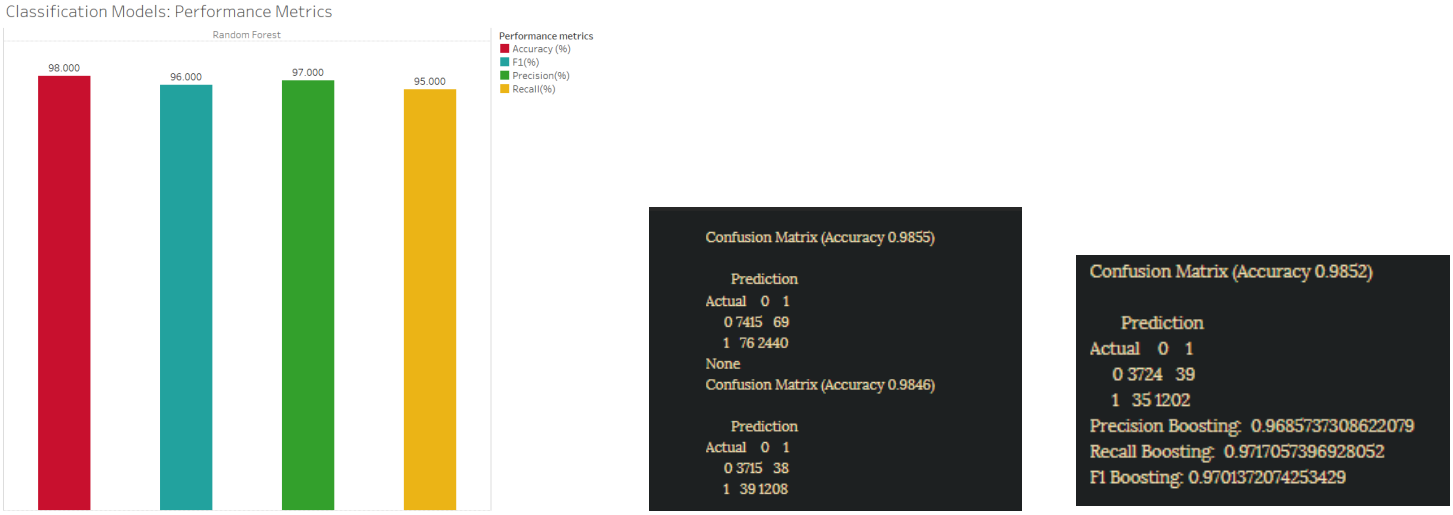

Figure 23. A. Performance metrics of the Grading boosting model B. Confusion matrix of the validation dataset. C. Confusion matrix of the testing dataset.

6. Conclusion

All the models performed very well because they obtained accuracies of 90% or higher. Since the data was purposely imbalanced due to the nature of traffic accidents, we expected some of the models to perform better. For this reason, we evaluated precision, recall, and F1 score. We can see in **Fig 18**. that KNN has a recall of 29%, which means that the model is classifying many accidents as non-accidents (Japkowicz, 2006). Decision trees and random forests performed better than their counterparts. The reason could be because of the nature of that data, and decision trees are very good at capturing non-linear relationships. In conclusion, the Decision Trees and Random Forest performed better. As a recommendation, it would be better to obtain the negative samples with a different method or use an unsupervised learning model to keep the data intact.

7. References

Bissonette, J. A., Kassar, C. A., & Cook, L. J. (2008). Assessment of costs associated with deer–vehicle collisions: human death and injury, vehicle damage, and deer loss. *Human-Wildlife Conflicts*, *2*(1), 17-27.

Bissonette, J. A., & Rosa, S. (2012). An evaluation of a mitigation strategy for deer-vehicle collisions. *Wildlife Biology*, *18*(4), 414-423.

Gonser, R. A., Jensen, R. R., & Wolf, S. E. (2009). The spatial ecology of deer–vehicle collisions. *Applied Geography*, *29*(4), 527-532.

Japkowicz, N. (2006, July). Why question machine learning evaluation methods. In AAAI workshop on evaluation methods for machine learning (pp. 6-11).

Mastro, L. L., Conover, M. R., & Frey, S. N. (2008). Deer–vehicle collision prevention techniques. *Human-Wildlife Conflicts*, *2*(1), 80-92.

Parra, C. A., Duarte, A., Luna, R. S., Wolcott, D. M., & Weckerly, F. W. (2014). Body mass, age, and reproductive influences on liver mass of white-tailed deer (Odocoileus virginianus). Canadian Journal of Zoology, 92(4), 273-278.

Reeve, Archie F., and Stanley H. Anderson. "Ineffectiveness of Swareflex reflectors at reducing deer-vehicle collisions." *Wildlife Society Bulletin (1973-2006)* 21.2 (1993): 127-132.

Ujvari, M., Baagøe, H. J., & Madsen, A. B. (1998). Effectiveness of wildlife warning reflectors in reducing deer-vehicle collisions: a behavioral study. *The Journal of wildlife management*, 1094-1099.

Yuan, Z., Zhou, X., Yang, T., Tamerius, J., & Mantilla, R. (2017, August). Predicting traffic accidents through heterogeneous urban data: A case study. In Proceedings of the 6th international workshop on urban computing (UrbComp 2017), Halifax, NS, Canada (Vol. 14, p. 10).