CIND 820: Capstone Project

Classification Models for Road Safety Prediction

Student ID: 500866890

Supervisor: Tamer Abdou

Date: Dec 3, 2024

Ryerson University

Table of Contents

Github Project Link

https://github.com/mkleung/traffic-collision-analysis

Abstract

Every day, between one and two individuals unexpectedly lose their lives in traffic collisions on Ontario's road. According to the Ontario's Provincial Police, 411 people have passed away in road accidents during 2023, making it amongst the deadliest in recent years. (CBC News, 2024)

Existent studies have explored road accident predictions based on external factors such as road conditions, weather and light but very few have tried to come up with an accurate probability of road accident collision based on a given location or even a user input daily commute route.

In this paper, a number of machine learning models will be evaluated to predict the traffic collision hot spots and road accident collision probability based on the most recent Ottawa road collision dataset available for free on the city's open data portal. The collected data will be cleaned and an exploratory data analysis will be performed on it to determine the patterns and relationships between the combined attributes and formulating hypotheses for regression or classification techniques.

The classification part of this project will include the comparison of several predictive models such logistic regression, decision trees and K-nearest models algorithms will be employed in order to find the probability of road collisions.

The final part of the project,which is optional and for personal use, will be to show a live demo where users can view the high risk locations and enter a route where a collision probability score will be calculated based on several options such as light, road condition and weather is can be obtained from a weather and time API and employing the algorithm with the highest accuracy.

Introduction

Traffic collision analysis has long been a tool used by city engineers to improve road safety and reduce the number of accidents on the roads. The recent advent of data science and artificial intelligence has enabled them to apply machine learning techniques to predict the number of traffic collisions at specific areas of the city and offer important insights and advice for urban planners and city governments as well as members of the public.

This project hopes to contribute to safer experiences on our roads when drivers choose their routes by providing them with different routes and their corresponding risk of traffic collisions. After analyzing historical traffic collision data, the likelihood of road accidents can be measured and quantified using these machine learning algorithms and other independent data such as weather, road condition and light and offer drivers with more informed details about their upcoming routes or their daily commutes and enable them to avoid high risk routes.

Stakeholders

1. **Urban planners**: they can use the tool to plan safer roads by identifying high risk locations.
2. **Members of the public**: They can use the collision prediction tool to plan a safer route to their destination. This is especially useful for communities to work or home.
3. **Police departments**: The tool can help them identify historical trends in traffic collisions and allocate resources to the hotspots.
4. **Insurance Companies**: Car insurance companies can use the model to identify at risk routes and adjust their insurance rates.
5. **Cyclists**: People on two wheels suffer from additional stress when travelling because they do not have protection against collisions. The models can help them plan a safer route to their destination.

Research Questions

1. **What are the factors that contribute most significantly to the amount of road collisions at a specific road or location? Does the speed of the vehicle influence the collision rate?**

2. **Which locations or roads are the deadliest with the highest amount of fatal injuries? Are accidents more likely to occur where there are a lot of accidents that happened in close proximity?**

3. **How do road surface conditions like wet or snowy or icy affect the probability of road collisions at various locations? Does time of the day and days of the week affect the number of vehicular collisions?**

4. **Does the road speed limit affect traffic collisions? Like do collisions occur more in higher speed limits like 60km per hour roads vs 30 km/h roads?**

5. **Do traffic control measures such as red light cameras, speed cameras, speed display radar and speed bumps reduce collisions?**

Approach

Target Variable

Target variable will be the number of injuries or number of fatal injuries at a specific location. We will compare several classification models and evaluate the most accurate one. The approach to the research questions can be summarized in the flowchart below

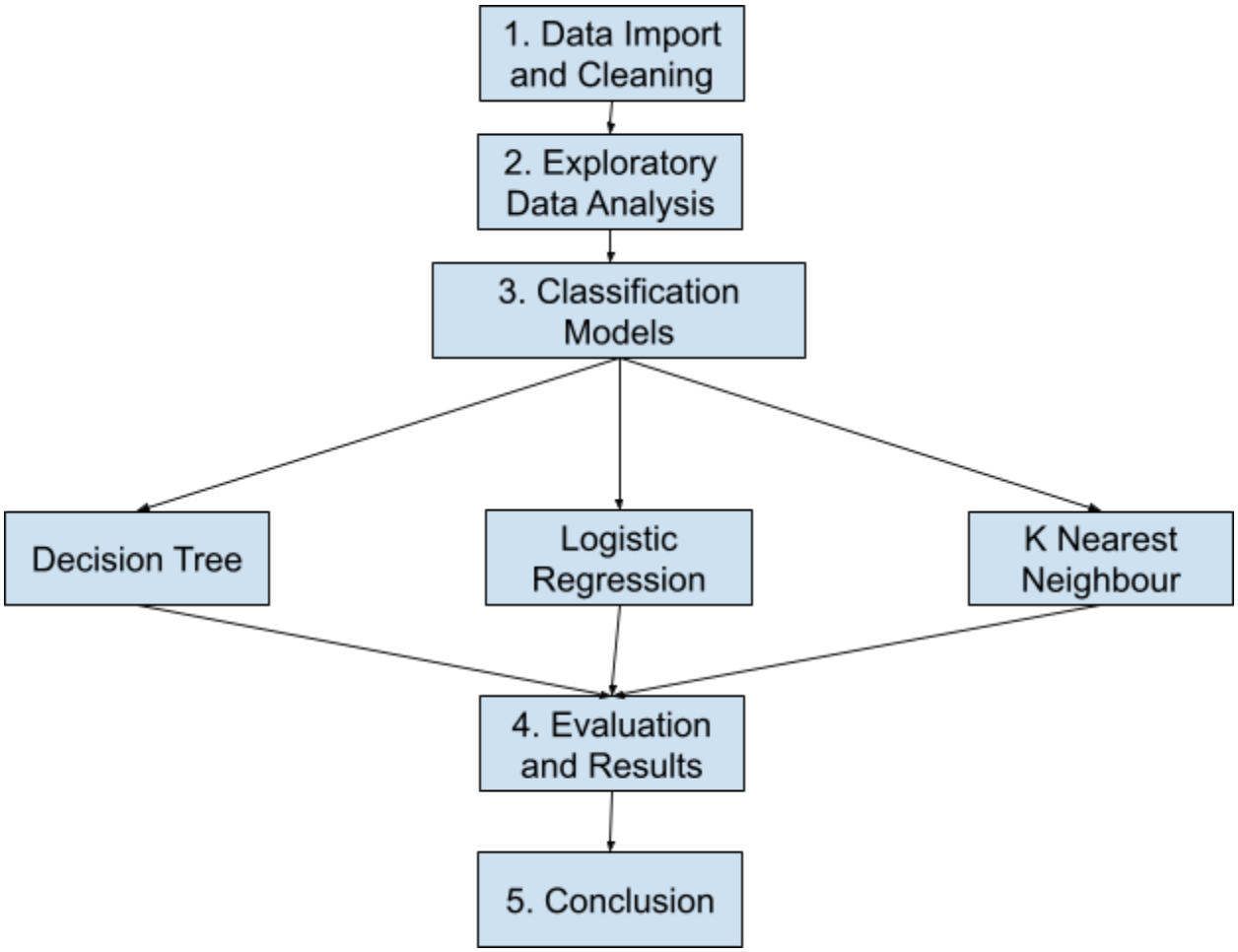


Figure 1: Flowchart of approach

Step 1: Import csv and Clean data

This part of the process is to import the data from the data source which is freely available from the data portal at the city of Ottawa in csv format and then removing null values, duplicates, repairing irregular data and performing one hot encoding for analysis.

Step 2: Exploratory Data Analysis

This part will focus on the analysis and will present graphs such as locations of the accidents, histograms of what time or day most accidents occur, what road condition has the most frequent collisions and other graphs about fatal injuries.

Step 3: Classification Model Comparison

The following algorithms will be applied to the cleaned data and each evaluated in order to find which one performs the best. This process is known as ensemble learning. These algorithms are all classification models like logistic regression, k nearest neighbor and decision trees. Random forests, although generally known to be more accurate, will not be used at this stage as it is computationally intensive and might not be appropriate for a demo web application.

Step 4: Evaluation and Results

In this part, we will use metrics such as accuracy, precision, recall and f1-score to evaluate each model and come up with the best model for a real live demo application which will ask the user to input a location or route and the application will display the probability of collision to the nearest accuracy.

Step 5: Conclusion

Literature Review

Paper 1 - Effects on inclement weather events on road surface conditions and traffic safety: An event based empirical analysis framework
Abohassan et Al. 2022

Abohassan et Al. conducted a study on the relationship between pavement friction coefficient and collision counts during inclement weather in Edmonton, Alberta during the years 2017-3029 and it shows how snowy road surface conditions in -15 degrees celsius for extended periods of time in late January increase the risk of collisions by 1,091% in injury crash rates and 2,113% increase in non-injury crash rates. The researchers used hourly weather datasets along with negative binomial safety performance functions to create models and found a statistically significant relationship between road friction and traffic collisions.

The study puts forward the idea that the environmental variables such as snow and ice decreases road friction to below 0.35, leading to a rise in traffic collisions. It also mentions larger roads such as main roads, motorways and highways with higher traffic volumes and speed limits contribute to higher traffic collisions compared to lower traffic smaller roads and neighbourhood streets. The paper concludes that a strong statistically significant relationship exists between pavement friction and collision counts and that if a severe snowstorm is forecasted, preventive measures have to be broadcasted to the general public.

Even though the paper used snow as its main focus, other extreme conditions such as low visibility situations such as early morning fog, lightning and heavy rainfall directly affect the risk of traffic collisions. The study is directly related to the research question which asks if weather has any effect on traffic collisions.

Paper 2 - Examining the relationship between wind speed, climatic conditions, and road accidents in Iran

Mahshid Eltemasi et al. 2024

In Iran, a study by Eltemasi et al. (2024) looked into the relationships between the different causes of road accidents ranging from distracted driving, wind speed, excess speed, human fatigue, environmental conditions and focusing particularly on wind speed using road accident data and wind speed data. The goal of the study was to identify which attributes or variables are the primary causes of car accidents and to provide recommendations for policy makers.

The study looked into road accident data from 2017 to 2022 consisting of 15 thousand cases and also wind speed data and used logistic regression to assess the relationship between wind speed and the type of accidents. The logistic regression results concluded that wind speed has no effect on the category of accidents like fatal or non-fatal but the only wind speed influence has is in the way the vehicles collide. They also used data mining like the J48 decision tree to visualize the relationships between numerous weather related attributes. The decision tree was able to predict collision probabilities in road accidents and found that rainy conditions increase the probability of non-fatal accidents and higher wind speed increases the probability of fatal accidents in front-to-back collisions.

A number of accident experts have been interviewed as well as past texts and articles have been reviewed which reveal human factors, not just environmental factors, contribute significantly more to traffic collisions. Fatigue and distractions such as mobile phones are the primary contributors to collisions while environmental factors come at a close second. The paper is directed related to the first research question and directly addresses whether the environmental factors contribute to traffic accidents.

Paper 3 - Reducing Road Traffic injuries: Effectiveness of speed cameras in an urban setting

Perez et al. 2007

A study investigated the effectiveness of speed cameras installed in Barcelona (Perez et al. 2007) at reducing traffic collisions between two groups using a local police database. The first group was conducted on the beltway of Barcelona which is a major highway without any speed cameras and the second group was conducted on the same beltway with new traffic cameras installed. The goal of the study was to find out if the installation of speed cameras led to a reduction in collisions and injuries.

The experiment conducted was based on a time series design and they used Poisson regression models to deal with historical trends and seasons to find that there is a 27% reduction in road collisions since the traffic camera installation. There was a reduction in injuries by 507 and 789 fewer vehicles were involved in these collisions. The study also found that the cameras prevented collisions at all times of the day, including nighttime and also on weekends.

This paper is directly connected to the research question number 3 which asks whether traffic cameras have an effect on collisions. It focuses on the effectiveness of speed cameras and provides empirical evidence on the impact of reducing road collisions, especially using two groups, one without and one with speed cameras installed after. Although the study's main focus is on speed cameras, other traffic control measures such as red light cameras, speed bumps, speed display radars and even the presence of police cars will have an effect on reducing traffic collisions.

Paper 4 - Safer Roads Owing to Higher Gasoline Prices: How long it takes
Guangqing Chi et al. 2015

This paper by Guangqing Chi et al. studies the correlation between gasoline prices and traffic collisions with a special consideration on the time between a change in gasoline prices and its effect on traffic collisions using the 2004 to 2023 traffic crash data and several variables such as economic difficulty, seat belts, alcohol consumption and fuel costs data from the US Department of Energy Information Administration.

The goal of the study consisted of three goals which were to find the positive relationship between higher gasoline prices and traffic collisions. The second goal was to check if variables like age, gender and race correlates with traffic accidents and the final aim is to identify how many months passed before gasoline prices have any effect on collisions which will be our main focus because it is related to research questions number 4.

The study used negative binomial regression and Pearson correlations to investigate how fuel costs affect traffic crashes. The researchers found out that the effect of gasoline changes on traffic collisions did not occur until 9-10 months after using a negative binomial regression model. The researchers also found that gasoline prices have different effects on different age groups. For teenage drivers, gas prices have an immediate effect on traffic collision reduction and this age range is the only group that has been affected by gas prices. There are no effects on adult drivers.

The findings concluded that higher gasoline prices lead to a smaller number of traffic collisions and suggested that government officials and policy makers should consider increasing gasoline prices in order to reduce traffic collisions.

Paper 5 - Factors influencing accident severity: An analysis by road accident type
Laura Eboli et al. 2007

This paper by Laura Eboli et al. conducted research on the different types of traffic collisions such as front/side and rear end collision and the relationships between the type of collision and the number and seriousness of injuries. They analyzed a dataset of 40, 172 road accidents that took place in Italy in 2016 and focused on three characteristics that might contribute to traffic collisions such as the type of road, weather conditions and the driver details such as his or her age, gender and license type.

Using a binary logistic regression, they modeled the relationship between accident type and various categorical attributes and tried to determine the importance of each attribute's effect on traffic collisions. The findings showed that there are significant differences in accident severity and the type of collisions. The location of the collision is a major factor which include intersection type. Other factors include driver related relationships such as license type, experience directly influence the severity of accidents. Gender has no impact on accident severity.

This study is interesting because it directly answers the last two research questions of this capstone project.

Data Description and Preprocessing

This project's dataset is available through the open Ottawa's government website which is a web portal that offers datasets related to the city of Ottawa like details about traffic, environmental and housing data amongst others. It is called "Traffic Collision data" and is one big csv file and contains data about road surface, weather, type of collisions, number of injuries, etc for the year 2017-2022 and the dataset can be found in the references.

The dataset contains 74,612 total number of records and 30 Columns or attributes. Here is a list of attributes at first glance and their corresponding descriptions.

- Date
- Time
- Location (RD1 @ RD2 or RD from RD 1 to RD 2)
- Location Type (Intersection, non-intersection, at/near private driveway)
- Classification of collision (non-fatal, fatal, property damage only)
- Initial impact type (Angle, turning movement, rear-end…)
- Road surface condition (Ice, wet, dry snow...)
- Environment (Clear, rain, snow…)
- Light (daylight, dawn, dusk…)
- Traffic control (stop, traffic signal, no control…)
- Number of Vehicles
- Number of Pedestrians
- Number of Bicycles
- Number of Motorcycles
- Max Injury (Highest injury level in the collisions)
- Number of Injuries
- Number of Minimal Injuries (Person did not go to hospital when leaving the scene of the collision)

- Number of Minor Injuries (Person went to hospital and was treated in the emergency room, but not admitted)
- Number of Major Injuries (Person admitted to hospital. Includes persons admitted for observation. This could be either life threatening or non-life threatening)
- Number of Fatal Injuries (Person killed immediately or within 30 days of the motor vehicle collision)
- X and Y Coordinate (MTM Zone 9, NAD83)
- Latitude and longitude (WGS1984)

Here are some statistics about the data. There are 17 numeric and 13 categorical columns. Running a null checker scan using python reveals the following details. Accident_year consists 100% of missing values. A few missing null values (less than 5) are in Accident_Year, Initial_Impact_type, Road_Surface_Condition, Traffic_Control and Environment_Condition.

Data Cleaning

The figure 1 below shows the columns and their corresponding reasons for removal

Column Names	**Reason for removal**
Accident_Year	Contains only null values
ObjectID, Geo_ID, ID	These are leftover database keys
X_Coordinate, Y_Coordinate, X , Y	These are duplicates of the lat and long
Location	Categorical location values containing street names will not be used as they are a less accurate descriptions

Num_of_Minimal_Injuries, Num_of_Minor_Injuries, Num_of_Major_Injuries	We will use num_of_injuries and fatal_injuries to simplify our data analysis
Location_Type, Classification_Of_Accident	Both contain only 2 unique categorical values as therefore do not contribute to our model
Initial_Impact_Type	Irrelevant
Accident_date, Accident_time	Combined into a column called Accident_dateTime which will be a timestamp for more accurate values

Figure 1. Summary of attributes removal

Null Values

The dataset contains tons of null values and other data irregularities and the table below shows the fix for these values

Column Names	**Null Values**	**Description and Fix**
Accident_Time	2757	Fix 2757 “unknown” values inside the column and replace them with a median value
Road_Surface_Condition	1	These contain numbers in front of them that need to be removed in order to be used as categorical data
Environment_Condition	2	
Traffic_Control	1	

Max_Injury	61143	These contains a lot of null values that were supposed to be zero but assumed to be incorrectly input as null during the data entry process
Num_of_Bicycles,	73265	
Num_of_Motorcycles	73975	
Num_of_Injuries	61195	
Num_of_Fatal_Injuries	74471	

Figure 2. Irregular data fix

Attribute Summary

Attribute	**Attribute Type**	**Quantitative**	**Min**	**Max**	**Mean**	**Std**	**Unique Values**
Road_Surface_Condition	Categorical / Nominal	Discrete	-	-	-	-	11
Environment_Condition	Categorical / Nominal	Discrete	-	-	-	-	9
Light	Categorical / Nominal	Discrete	-	-	-	-	6
Traffic_Control	Categorical / Nominal	Discrete	-	-	-	-	12

Num_of_Vehicles	Quantitative	Discrete	1.000000	25.000000	1.841648	0.586337	10
Num_of_Pedestrians	Quantitative	Discrete	0.000000	3.000000	0.022305	0.153909	4
Num_of_Bicycles	Quantitative	Discrete	0.000000	3.000000	0.018223	0.135354	4
Num_of_Motorcycles	Quantitative	Discrete	0.000000	3.000000	0.008675	0.094315	4
Injury_Type	Categorical / Ordinal	Discrete	-	-	-	-	5
Num_of_Injuries	Quantitative / Continuous	Discrete	0.000000	38.000000	0.233150	0.571718	10
Num_of_Fatal_Injuries	Quantitative / Continuous	Discrete	0.000000	2.000000	0.001987	0.046887	3
Lat	Quantitative	continuous	0.000000	45.5249	45.292519	1.836272	42382
Long	Quantitative	continuous	-79.237290	-75.261583	-75.710279	0.167509	42165
Accident_Timestamp	Ordinal	continuous	-	-	-	-	70045

Figure 3. Attribute Summary After Cleaning

Outliers

For figure 3, there are a number of outliers that stand out. For example, the latitude has the minimum value of zero which is impossible as it is out of bounds for the city of Ottawa which has a longitude of -75 and latitude within 45.

Another anomaly is the number of injuries which is 38. Upon further examination, there is a single outlier in that column and it can be easily removed. The case in question is a famous bus crash that happened in 2019 where 38 passengers were injured and 3 people passed away.

No further outliers have been detected as the injury data and other numerical values are very singular, ranging from zero to 5 in the case of injuries and fatal injuries.

Duplicates

There are only 8 duplicates found in the dataset of 74,000 rows and these duplicate rows have easily been removed.

Feature Engineering

These are the two columns that have been added to our dataset.

Accident_Timestamp : The Accident_Timestamp is a new feature column that is in datetime format which indicates the exact time that crash happened and it is the combination of the existing Accident_Time and Accident_Date attributes. This feature will help is visualize the hour and days where the most collisions occur.

Collision_Happen : Another new feature column that can be created is the Collision_Happen which is a binary feature taken from the “number_of_injuries”. It will be 1 if the number is greater than zero. This feature will be the target variable for our machine learning.

One-hot Encoding

All the categorical attributes such as road_surface_condition, environment_condition, light and traffic_control will be one hot encoded and be converted into numerical values. These are the columns that have been one hot encoded

- Road_Surface_Condition
- Environment_Condition
- Light
- Traffic_Control
- Injury_Type

Standardization

These columns have been standardized to bring them to a common scale and to improve the performance of machine learning algorithms.

- Num_of_Vehicles
- Num_of_Pedestrians
- Num_of_Bicycles
- Num_of_Motorcycles
- Num_of_Injuries
- Num_of_Fatal_Injuries
- Latitude and Longitude

Exploratory Data Analysis

Correlation Heatmap

The histograms below show that none of the graphs are normally distributed and many of the quantitative variables such as num_of_vehicles, num_of_Pedestrians, Num_of_Bicycles, Num_of_fatal are skewed towards zero. They follow a poisson distribution.

The longitude and latitude follow a binomial distribution and are normal for geometry data as the latter are not normally distributed. Latitude is skewed to the left because Ottawa's population density seems to be higher to be on the larger latitude (near the parliament) which might correlate to more collisions.

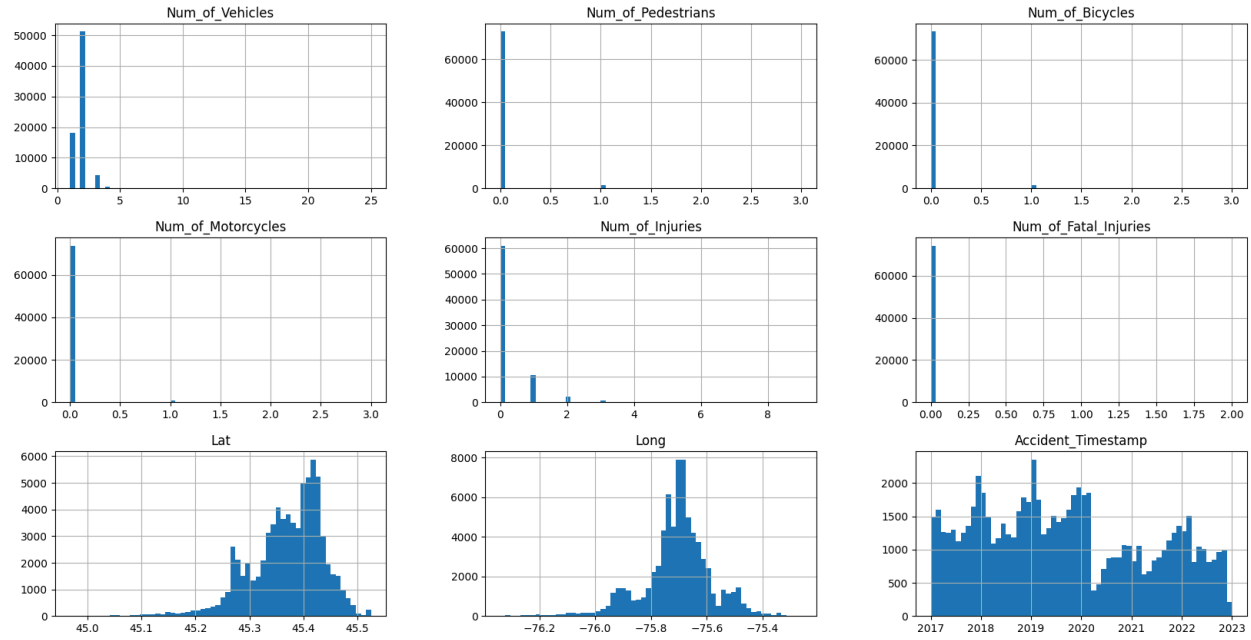


Figure 4: histogram of distributions

Correlation Matrix and Heatmap

A correlation matrix was created and a heatmap was generated to determine any correlations between the attributes.

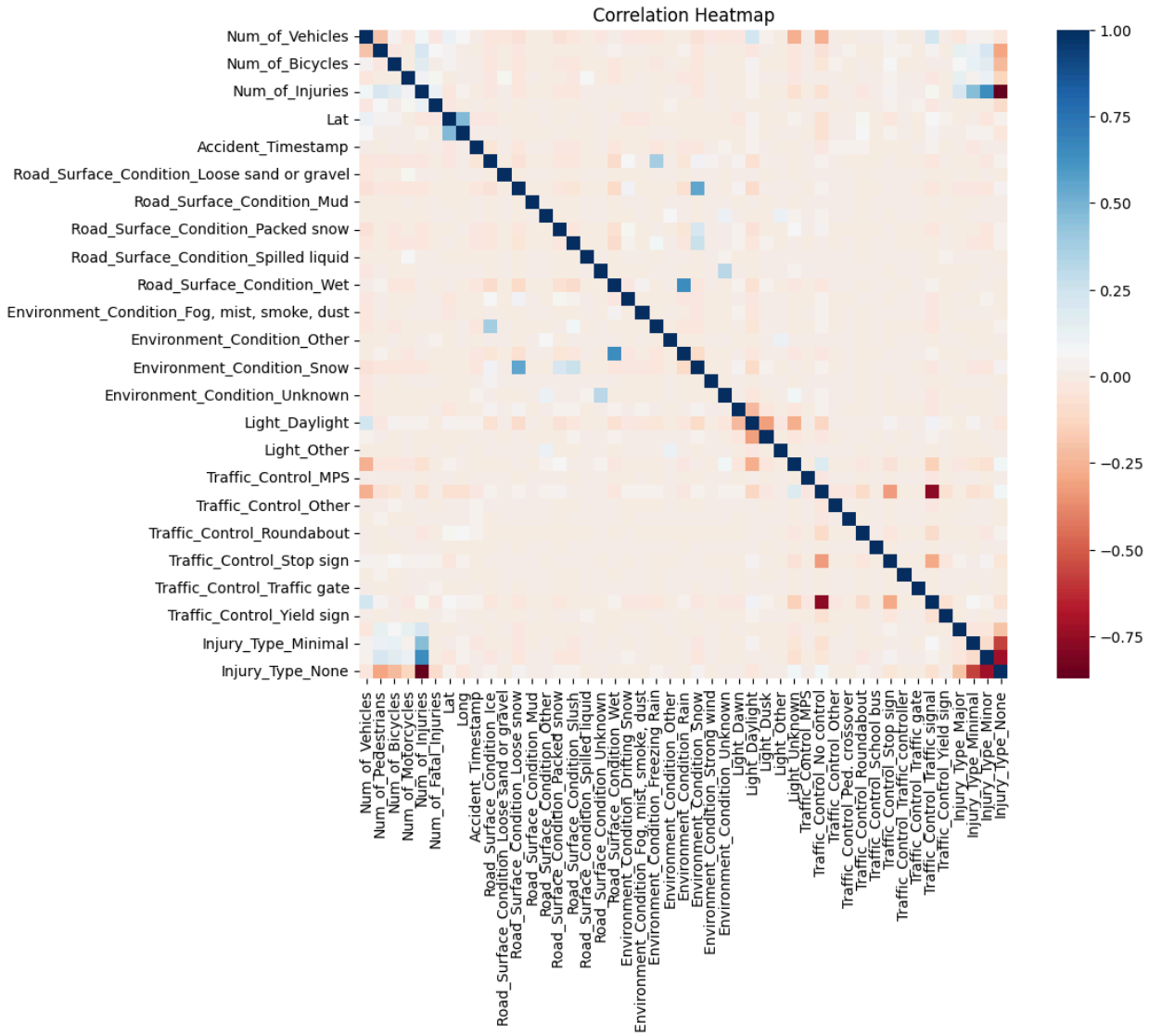


Figure 5: Traffic Collision Heatmap displaying correlation values

These are the top 5 strongest correlations. A strong positive correlation between a rainy weather and wet road surfaces indicates that roads tend to get wet when it rains. This is important for traffic safety and it will affect the road collision probability because wet roads increase stopping distances.

	Attribute 1	**Attribute 2**	**Correlation Value**
1	Environment_Condition_Rain	Road_Surface_Condition_Wet	0.655352
2	Road_Surface_Condition_Wet	Environment_Condition_Rain	0.655352
3	Injury_Type_Minor	Num_of_Injuries	0.644281
4	Num_of_Injuries	Injury_Type_Minor	0.644281
5	Environment_Condition_Snow	road_surface_condition_loose_snow	0.554040

Figure 5: Top 5 Highest Correlation relationships

Statistical Analysis and Dimensionality Reduction

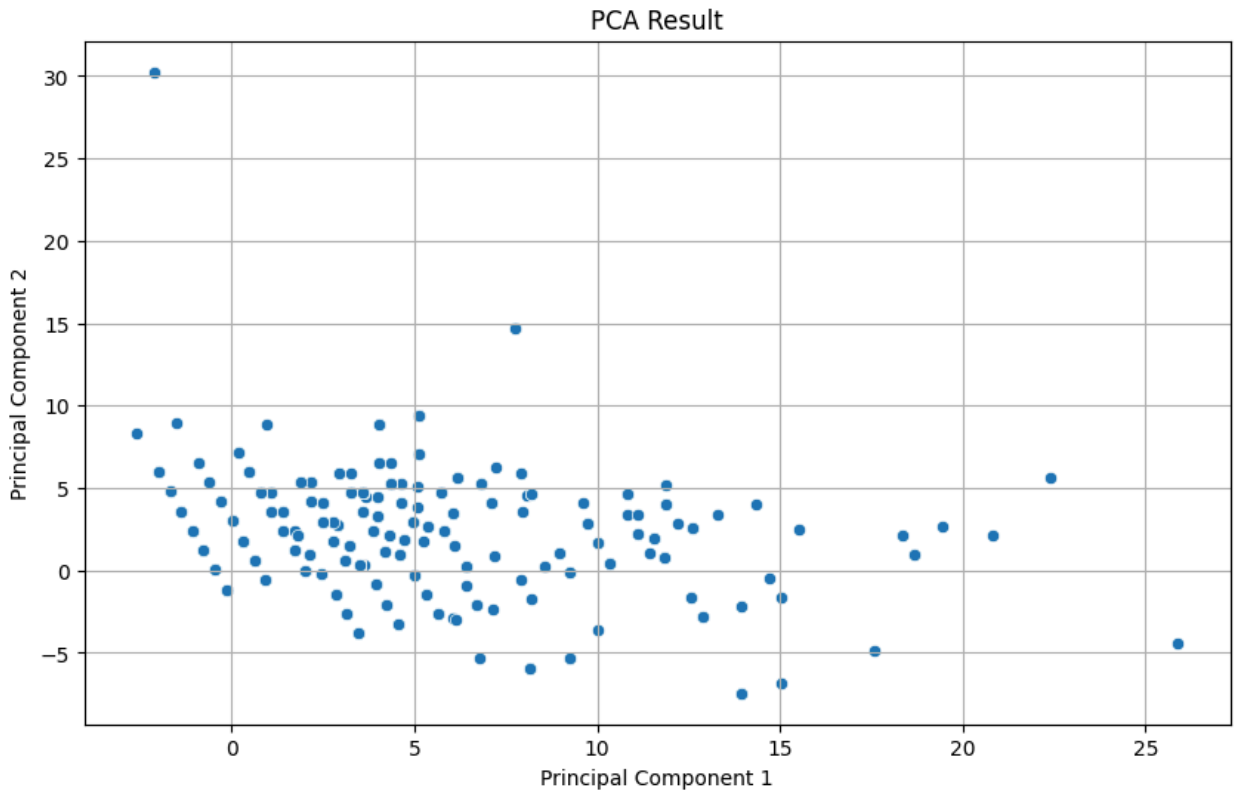
Anova Test

Target variable is Num_of_Injuries

Attribute	**F-Statistic**	**P-Value / PR(>F)**	**Correlation Probability**
Environment_Condition_Rain	0.238268	0.625461	No

Road_Surface_Condition_Ice	35.34936	2.767821e-09	No
Environment_Condition_Snow	148.987339	3.112163e-34	No

Principal Component Analysis



Chi Square Test

Categorical Attribute 1	Categorical Attribute 2	Chi-Square Statistic	p-value	Association

Road_Surface_Condition	Environment_Condition	92373.3573 4425038	0	Strong
Road_Surface_Condition	Traffic_Control	496.745677 60756645	1.62632 1097870 968e-50	Very Strong
Road_Surface_Condition	Light	3879.60431 92033585	0	strong

Visualizations

1. **Accident Count per hour**

The figure 6 below shows that accidents are the highest in the afternoon from 2pm to 6pm and the lowest at night after 11pm. It is very interesting as we assume that most accidents happen at night because of the low visibility but this is not the case. One assumption is that during the day, there are more cars on the road and this will lead to a bigger chance of collisions.

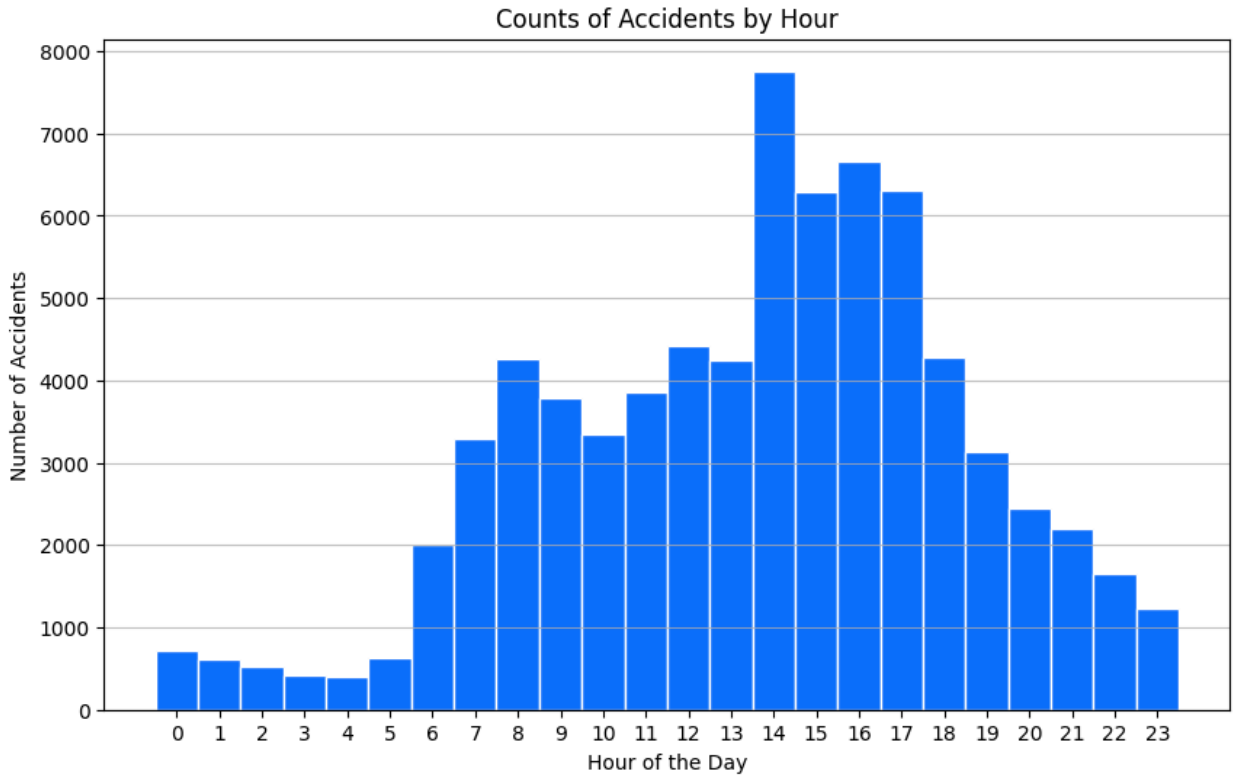


figure 6: Histogram showing daily collisions

2. Accident Count per day of the week

The figure 7 below shows that collisions are not very high during weekends and most accidents happen on a friday due to the higher amounts of traffic on Fridays. Alcohol or other leisurely activities which generally happen on Fridays likely contribute to the higher accident count.

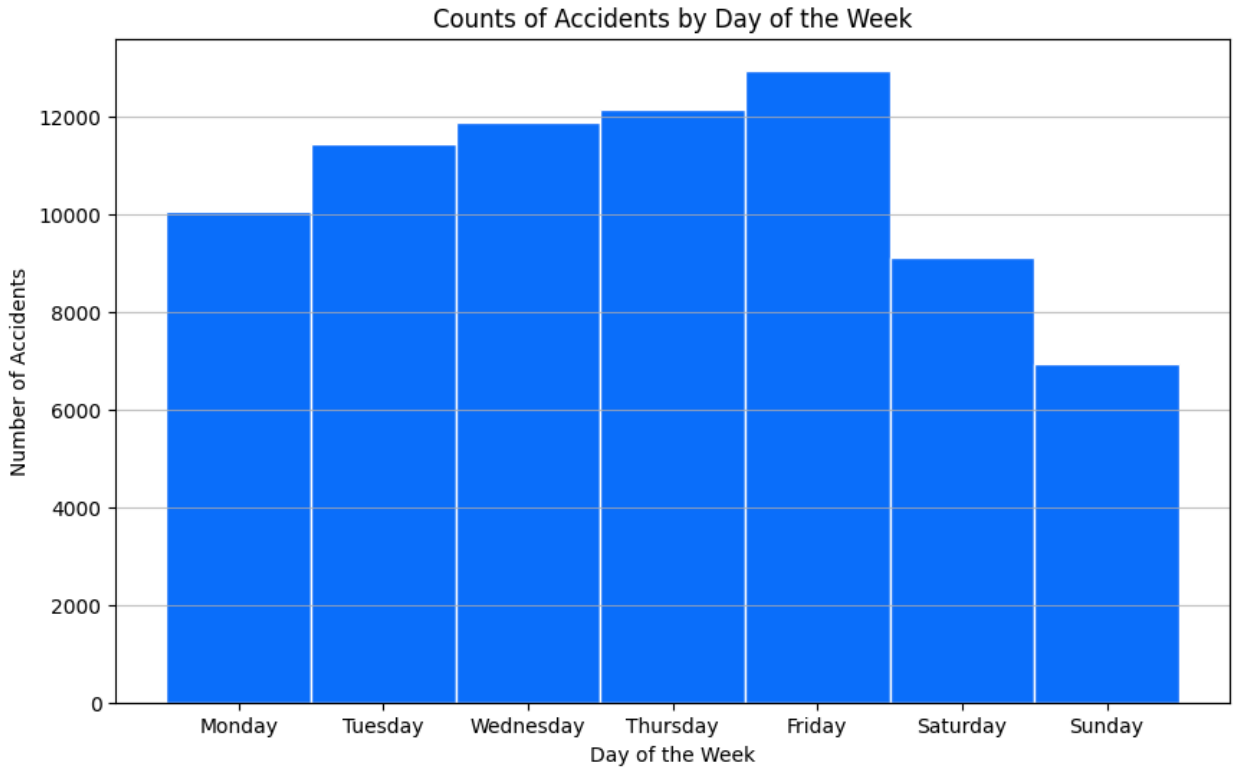


Figure 7: Number of Accidents for each day of the week

3. Accident count by traffic control measure

Figure 8 shows the accident count by traffic control. Notice that traffic and no signal have the highest collisions.

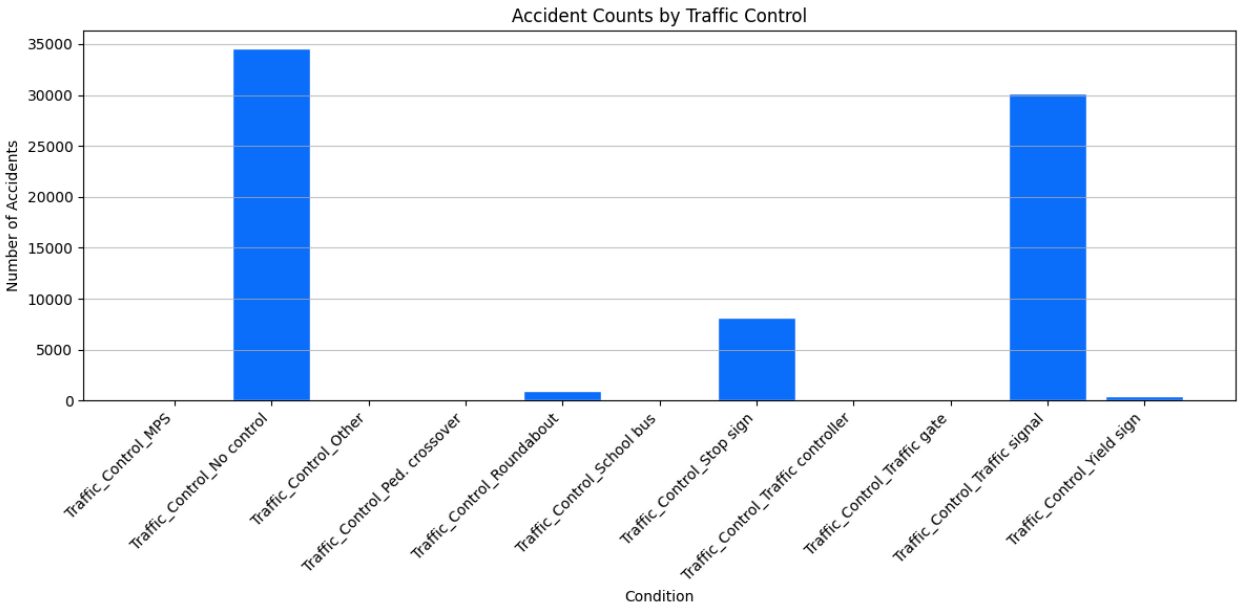


Figure 8: Histogram of Traffic Control against Accident Count

4. Accident Counts by Road Surface Condition

Rainy weather tends to lead to higher collisions because a slippery surface will decrease the brakes functionally and effectiveness and this may lead to more road accidents.

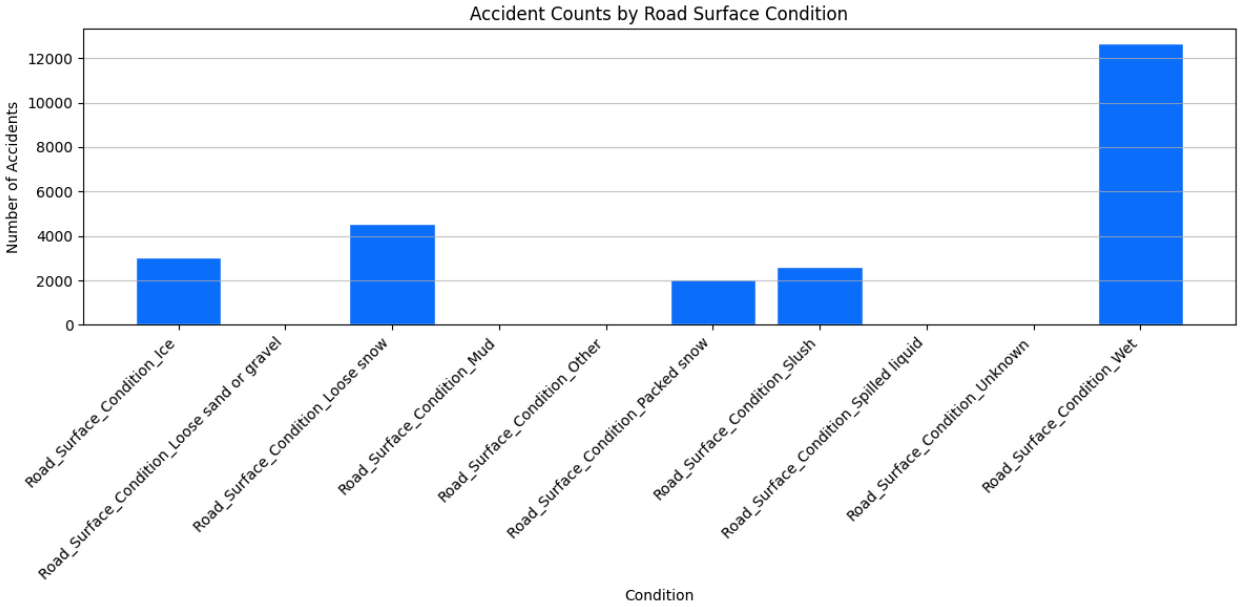


Figure 9: Relationship between Road Surface Condition and Accident Count

5. Fatal Injuries map

Figure 8 below shows the fatal injuries that have ocurred at the different locations around the city. Even if we are not going to use the fatal_injuries as our target variable, it is an important metric that brings attention to road collisions.

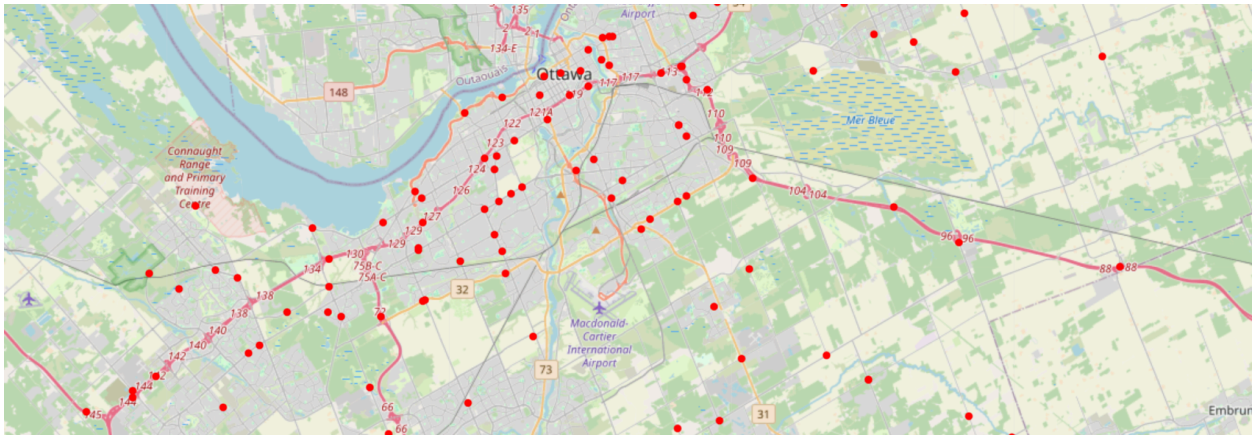


Figure 8: Geospatial visualization of fatal accident locations

Classification Models

We are going to test three classification models for this dataset for this project. These are the logistic regression, decision tree, random forest and K-Nearest Neighbour models. These algorithms are simple to implement and are suitable for analysing traffic collisions and gaining insights into patterns that contribute to collisions and eventually go on the develop predictive estimations of traffic collisions in a web project.

Logistic Regression

Logistic regression is a statistical method that predicts if a boolean event will happen or not like collision or no collision. In this case, the boolean variable will be Num_of_Injuries which will be the dependent variable and a new column has been generated which stores true if the number of injuries is greater than zero.

```
Accuracy: 0.8583993554451457
Confusion Matrix:
[[12147    88]
 [ 2021   638]]
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.99      0.92     12235
           1       0.88      0.24      0.38      2659

    accuracy                           0.86     14894
   macro avg       0.87      0.62      0.65     14894
weighted avg       0.86      0.86      0.82     14894
```

Decision Tree Classifier

A decision tree is a type of pyramidal flowchart representing a tree with branches and each branch is a decision based on questions asked. The decision tree used will have a max depth of 5. The final outcome is a yes or no on whether a collision will happen.

```
Accuracy: 0.858533637706459
Confusion Matrix:
[[12152    83]
 [ 2024   635]]
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.99      0.92     12235
           1       0.88      0.24      0.38      2659

    accuracy                           0.86     14894
   macro avg       0.87      0.62      0.65     14894
weighted avg       0.86      0.86      0.82     14894
```

K Nearest Neighbor

The k-nearest neighbor (KNN) is a good candidate choice for examining traffic collisions because this dataset contains a lot of spatial data like latitude and longitude where nearby collisions may share similar collision patterns. KNN is particularly effective for location based datasets.

```
Accuracy: 0.8400698267758829
Confusion Matrix:
[[11788   447]
 [ 1935   724]]
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.96      0.91     12235
           1       0.62      0.27      0.38      2659

    accuracy                           0.84     14894
   macro avg       0.74      0.62      0.64     14894
weighted avg       0.82      0.84      0.81     14894
```

Results

This table below shows the comparison of the performance of four different classifications algorithms using key metrics like accuray, precision, recall and f score.

Algorithm	**Accuracy**	**Precision**	**Recall**	**f1-score**
Logistic Regression	0.8583993554451457	0.86	0.99	0.92
Decision Tree	0.8585336377064459	0.86	0.99	0.92
Random Forest	0.8292601047401639	0.86	0.95	0.90
K-Nearest Neighbor	0.8585336377064459	0.86	0.99	0.92

Logistic Regression:

- High accuracy (85.84%), high recall (99%), and good precision (86%) mean that it is effective in identifying positive cases.

Decision Tree:

- Similar performance to logistic regression in terms of accuracy, precision, and recall, indicating it is also a strong model.

Random Forest:

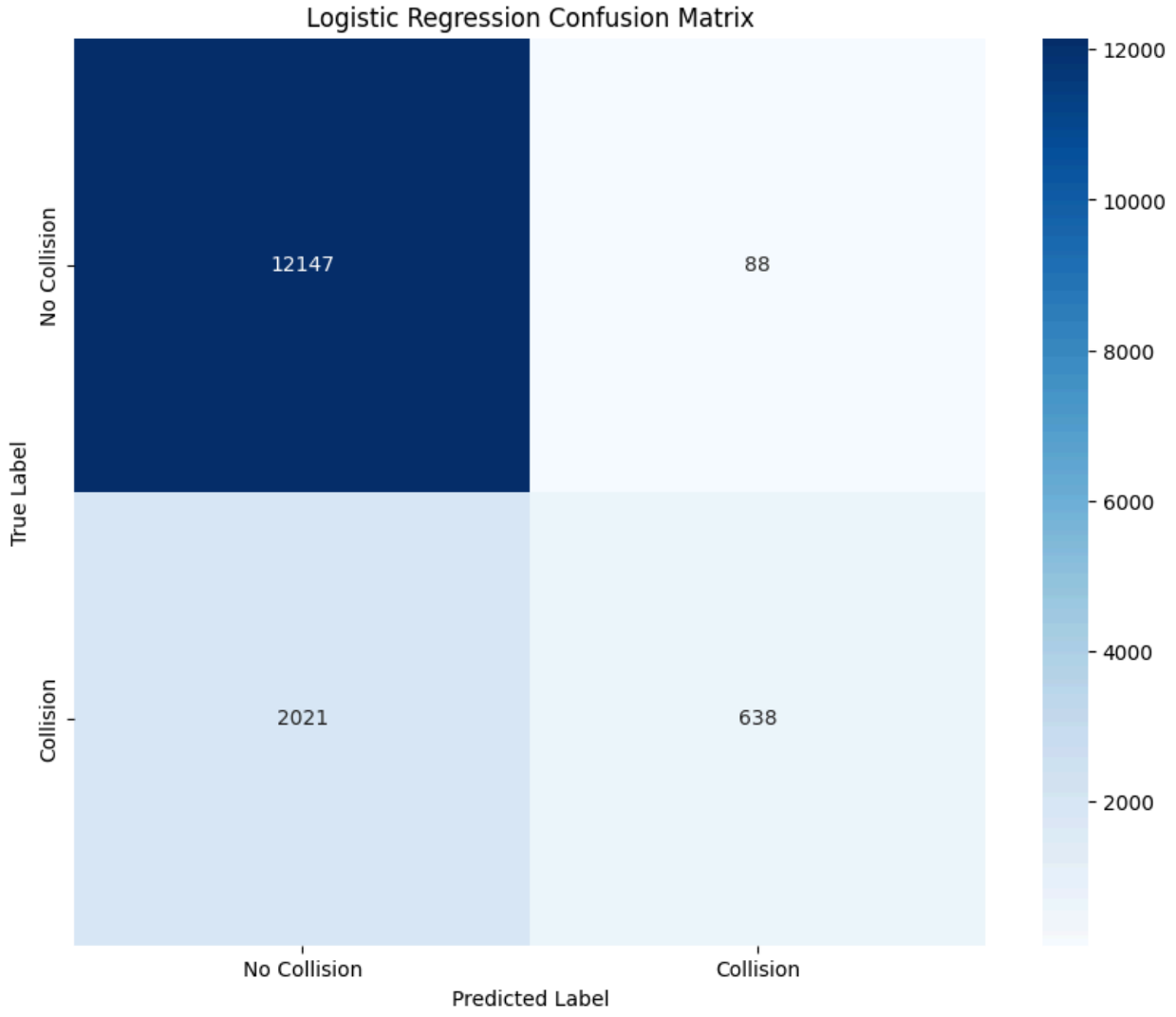
- Lower accuracy (82.92%) compared to the other models, with good precision (95%) but lower recall (90%) suggests it misses some positive cases.

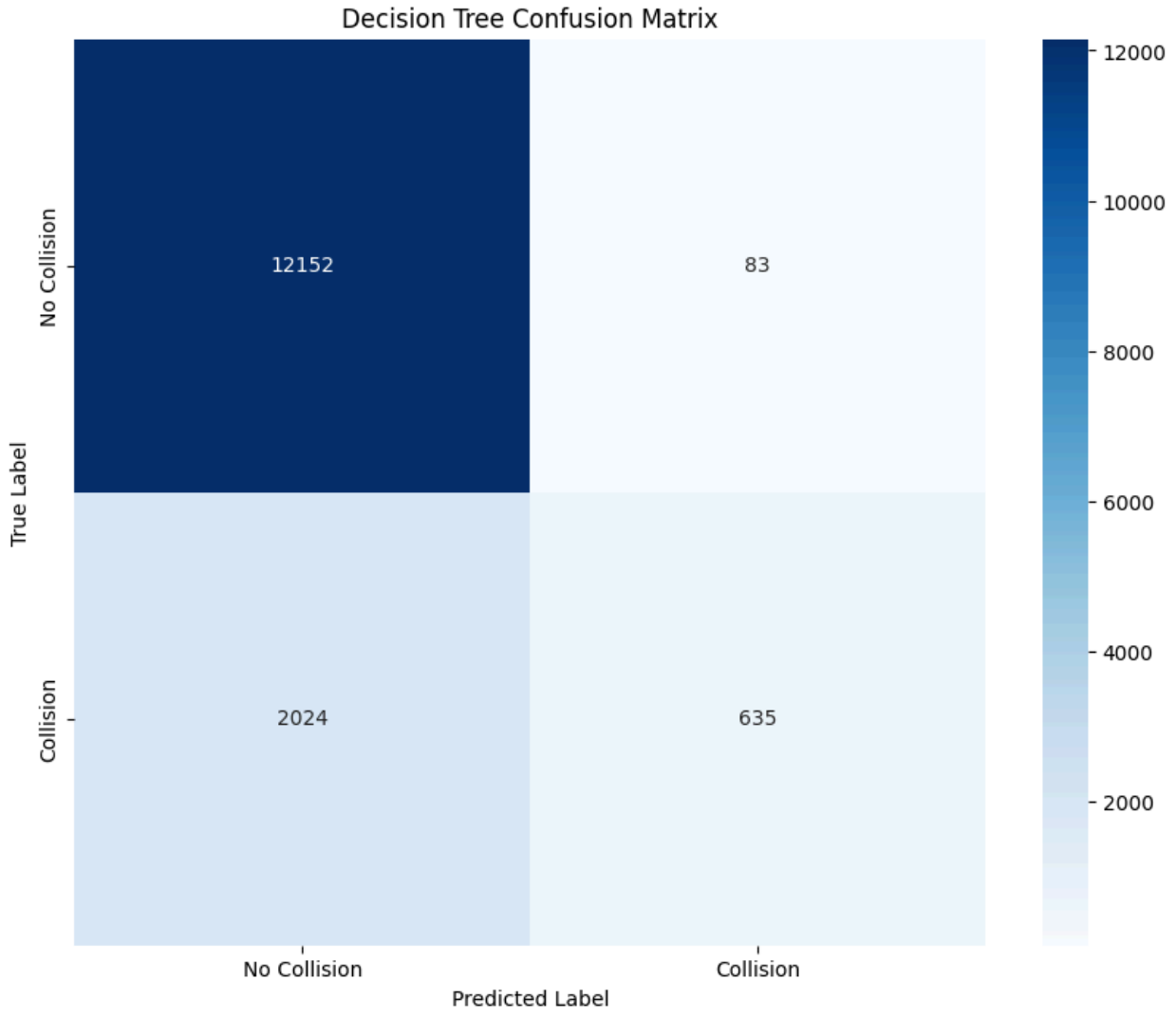
K-Nearest Neighbor:

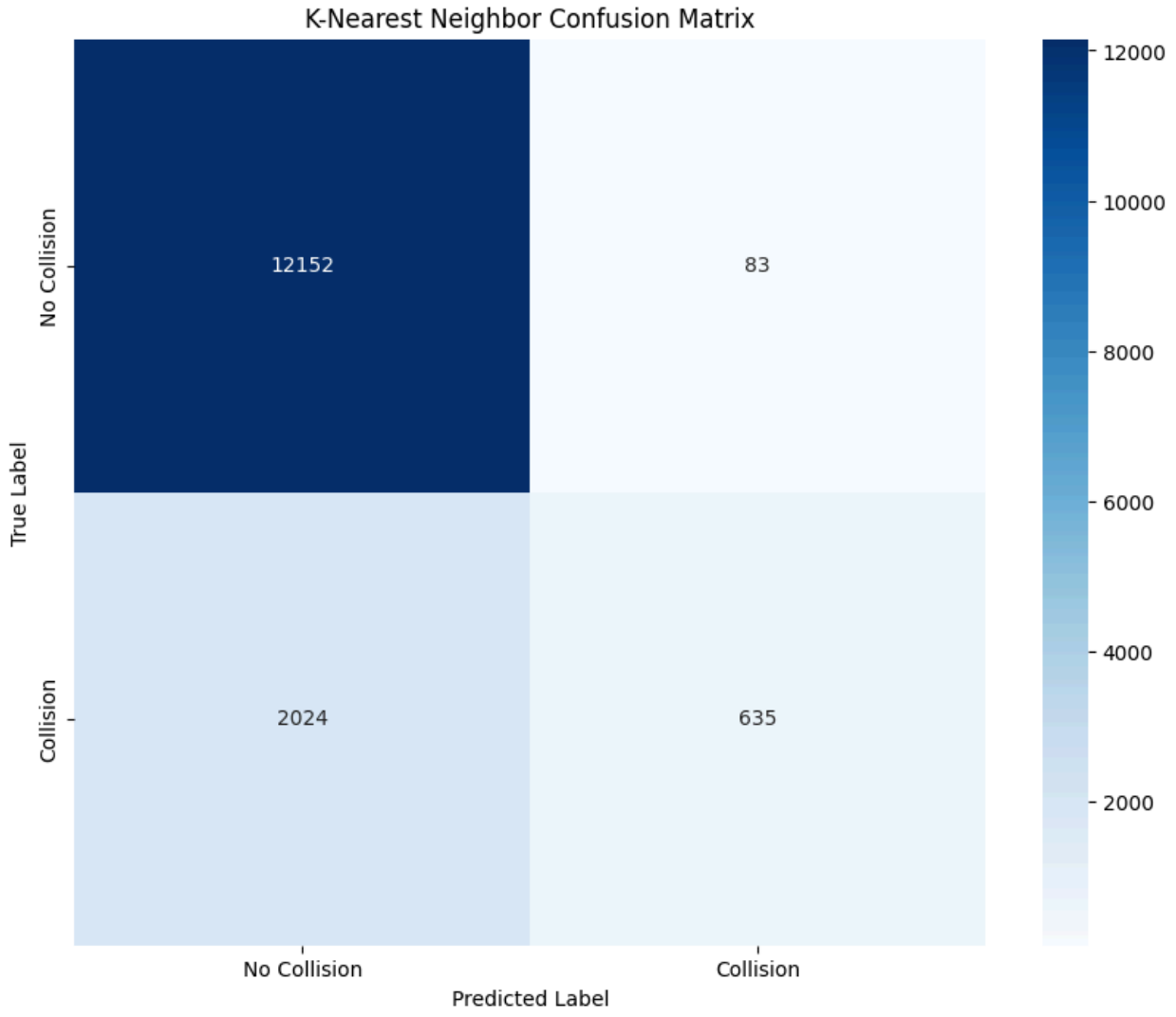
- Performance is similar to logistic regression in terms of accuracy, precision, and recall, which indicates it also works well for this classification task.

Due to time constraints, I was not able to perform hypertuning on the algorithms.

Confusion Matrix Comparison




Decision Tree Confusion Matrix
12152
83
2024
635
No Collision
Collision
True Label
No Collision
Collision
Predicted Label
12000
10000
8000
6000
4000
2000


K-Nearest Neighbor Confusion Matrix
12152
83
2024
635
No Collision
Collision
True Label
No Collision
Collision
Predicted Label
12000
10000
8000
6000
4000
2000

Recommendations

The data analysis on the project is still incomplete due to time constraints and more needs to be done like hyperparameter tuning.

It would be wise to use other data sources from other Ontario cities like Kingston and Toronto to improve the models. The next step is to build an application where the general public can enter a route from their home to a destination and the model will split out a collision probability.

Conclusion

In this project, several machine learning classification models have been employed to predict collision probability based on historical road collision data. Decision tree and logistic regression exhibit the highest performance at 85% in accuracy score and the decision tree outscore the other by a little bit.

Out of all models examined, decision trees have emerged as the best performing model based on accuracy for this classification and will be used in the online application. Decision trees are perfect for recommending collision probabilities.

Additional work will need to be done to explore additional feature engineering, hyperparameter tuning and the use of ensemble solutions like random forest and gradient booster. Additionally associative rule mining algorithms like apriori can be used to detect frequent itemsets in the collision data to improve feature engineering.

References

Dataset Source https://open.ottawa.ca/datasets/ottawa::traffic-collision-data/about

[1] Ahmed Abohassan et. al. retrieved from
https://journals.sagepub.com/doi/full/10.1177/03611981221088588

[2] Eltemasi et al. (2024) retrieved from
https://www.cell.com/heliyon/fulltext/S2405-8440(24)09259-4

[3] Perez et al. 2007 retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC1963295/

[4] Guangqing Chi et al. retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC4504271/

[5] Laura Eboli et al. retrieved from
https://www.sciencedirect.com/science/article/pii/S2352146520303197