CST8390

BUSINESS INTELLIGENCE AND DATA ANALYTICS

FINAL REPORT

PROFESSOR: Subair Abayomi Oloko

STUDENT DETAILS:

Arpandeep Singh : 040950261

Param Savalia: 040963842

Question: What is the Type of accident in Toronto and East York District with respect to the time of the day, age and road class.

Table of Contents

INTRODUCTION

The dataset that we chosen has been obtained from Open Toronto and is provided by peel police. It contains information which includes traffic collision events where a person was either Killed or Seriously Injured between the years 2006 and 2020.
That dataset has a total of 54 attributes with over 16K records.
Attributes of dataset are:

_id	Unique row identifier for Open Data database
ACCNUM	Accident Number
YEAR	Year Collision Occurred
DATE	Date Collision Occurred
TIME	Time Collision Occurred
HOUR	Hour Collision Occurred
STREET1	Street Collision Occurred
STREET2	Street Collision Occurred
OFFSET	Distance and direction of the Collision
ROAD_CLASS	Road Classification
DISTRICT	City District
WARDNUM	City Ward Identifier, will show multiple if collision occurred along a border
DIVISION	Police Division(s), will show multiple if collision occurred along a border
LOCCOORD	Location Coordinate
ACCLOC	Collision Location
TRAFFCTL	Traffic Control Type
VISIBILITY	Environment Condition
LIGHT	Light Condition

RDSFCOND	Road Surface Condition
ACCLASS	Classification of Accident
IMPACTYPE	Initial Impact Type
INVTYPE	Involvement Type
INVAGE	Age of Involved Party
INJURY	Severity of Injury
FATAL_NO	Sequential Number
INITDIR	Initial Direction of Travel
VEHTYPE	Type of Vehicle
MANOEUVER	Vehicle Manouever
DRIVACT	Apparent Driver Action
DRIVCOND	Driver Condition
PEDTYPE	Pedestrian Crash Type - detail
PEDACT	Pedestrian Action
PEDCOND	Condition of Pedestrian
CYCLISTYPE	Cyclist Crash Type - detail
CYCACT	Cyclist Action
CYCCOND	Cyclist Condition
PEDESTRIAN	Pedestrian Involved In Collision
CYCLIST	Cyclists Involved in Collision
AUTOMOBILE	Driver Involved in Collision
MOTORCYCLE	Motorcyclist Involved in Collision
TRUCK	Truck Driver Involved in Collision
TRSN_CITY_VEH	Transit or City Vehicle Involved in Collision
EMERG_VEH	Emergency Vehicle Involved in Collision

PASSENGER	Passenger Involved in Collision
SPEEDING	Speeding Related Collision
AG_DRIV	Aggressive and Distracted Driving Collision
REDLIGHT	Red Light Related Collision
ALCOHOL	Alcohol Related Collision
DISABILITY	Medical or Physical Disability Related Collision
POLICE_DIVISION	Toronto Police Service Division
HOOD_ID	City of Toronto Neighbourhood Identifier
NEIGHBOURHOOD	City of Toronto Neighbourhood Name
ObjectId	Object ID (Unique Identifier)
geometry	

Initial Guess

The data from the surveys and the data from Toronto Police Service (TPS) statistics indicate that the roads are getting safer and Toronto has seen much fewer fatal-accidents than non-fatal accidents. Study also suggests that youth involvement in the collisions has increased over the years.

Also, as of the data collected, during the spring and summer seasons non fatal accidents peaked between 12pm - 3pm. However fatal-crashes are more likely to occur between 8pm and midnight.

Data Collection

The data set has been obtained from the official website of Open Toronto and the data is provided by Toronto Police Services. Toronto peel police does not guarantee the accuracy and completeness of the dataset.

Link:
https://open.toronto.ca/dataset/motor-vehicle-collisions-involving-killed-or-seriously-injured-persons/

Data Pre - Processing

Finding Duplicates: We applied RemoveDuplicates Filter, but none were removed implying that there were no duplicates.

Data Cleaning: Out of total attributes , some of the attributes have been removed which do not meet the needs of our study. The removed attributes are listed below.

ACCLOC, ACCNUm, AG_DRIV, ALCOHOL, AUTOMOBILE CYCACT, CYCCOND, CYCLIST, CYCLIST TYPE, DATE, DISABILITY, DIVISION, DRIVACT, DRIVCOND, EMERG_VEH, FATAL_NO, HOOD_ID, IMPACTYPE, INITDIR, INVTYPE, LOCCOORD, MANOEUVER, MOTORCYCLE, NEIGHBOURHOOD, OFFSET, PASSENGER, PEDACT, PEDCOND, PEDESTRIAN, PEDTYPE, POLICE_DIVISION, RDSFCOND, REDLIGHT, SPEEDING, STREET1, STREET2, TRAFFCTL, TRSN_CITY_VEH , VEH TYPE, WARDNUM, YEAR, GEOMETRY.

Final Attributes:

No.	Name
1	DISTRICT
2	INVAGE
3	ROAD_CLASS
4	VISIBILITY
5	ACCLASS
6	TIME_OF_DAY
7	VEHTYPE

Data Modifications:

1. District Names have been shortened for better readability

	Label
1	Tor & E York
2	Scar
3	Eto York
4	N York
5	NA
6	Tor E York

2. Age ranges have been grouped together into age groups of Child, Adolescent, Youth, Adult and Senior.

0 to 4, 5 to 9	Child
10 to 14, 15 to 19	Adolescent
20 to 24, 25 to 29	Youth
30 to 34 … 60 to 64	Adult
>65	Senior
(Blanks)	NA

3. Hour column is converted to time of the day (Morning /Afternoon / Evening/ Night) using the formula :

   =IF(AND(C2>=5,C2<12),"Morning",IF(AND(C2>=12,C2<17),"Afternoon",IF(AND(C2>=17,C2<22),"Evening","Night")))

4. For the Road_Class, similar road types have been put together into a single road class. for example:

- Major Arterial, Major Arterial Ramp and Minor Arterial —> Arterial
- Expressway and Expressway Ramp → Expressway

5. Empty records are replaced with NA implying Not Available.

Data Analysis

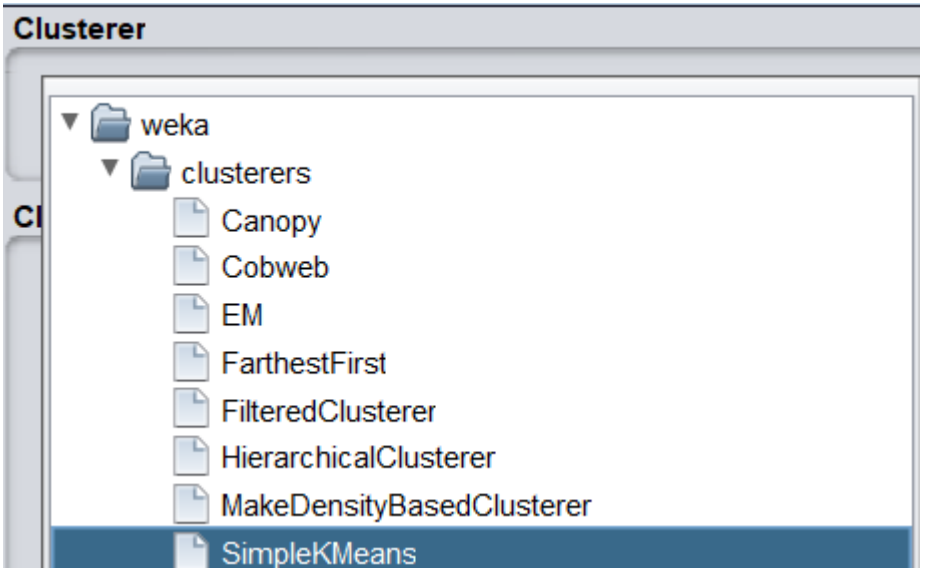
Information regarding the attributes:

Attribute	Distinct Count
District	6
Invage	6
Road_Class	7
Visibility	9
AccClass	3
Time_of_day	4
Veh_type	22

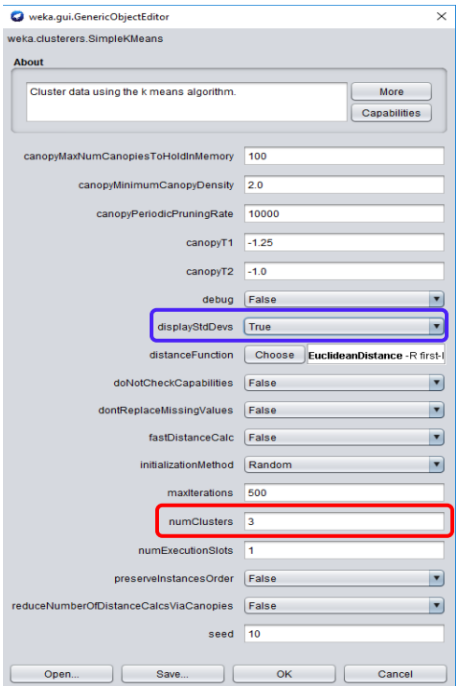
Now it's our time to perform some algorithms to get to our desired results.
We used the below algorithms to analyse the data.

Clustering

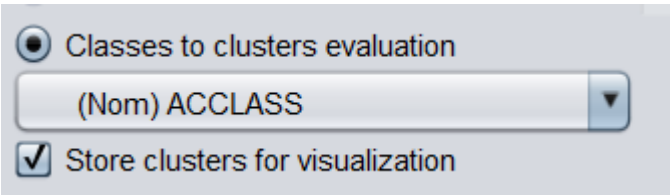
The motive of performing clustering is to find similar groups of data and to analyse it. We used SimpleKMeans to perform the clustering. From the cluster tab we chose weka→ clusters → SimpleKMeans



For the parameters we started with choosing K=3 since we had three different classes forACClass (Fatal/Non-Fatal/Property Damage Only) and displayStdDevs was set to True , the rest other parameters remained default.



For “Cluster mode”, we selected “Classes to clusters evaluation” and selected ACClass



After running the algorithm, Weka generated the result and we plotted the graph with different values for K (3,4,5,6).

K	Within cluster sum of squared errors
3	37800
4	34670
5	32420
6	31808
7	30455

From these points, we constructed the line chart and found the best K using the Elbow Method.

X Axis denoted the number of clusters and Y axis represents the Sum of Squared Errors.
From the graph we can see that ,from K = 5 onwards and above , the y axis is not fluctuating significantly.

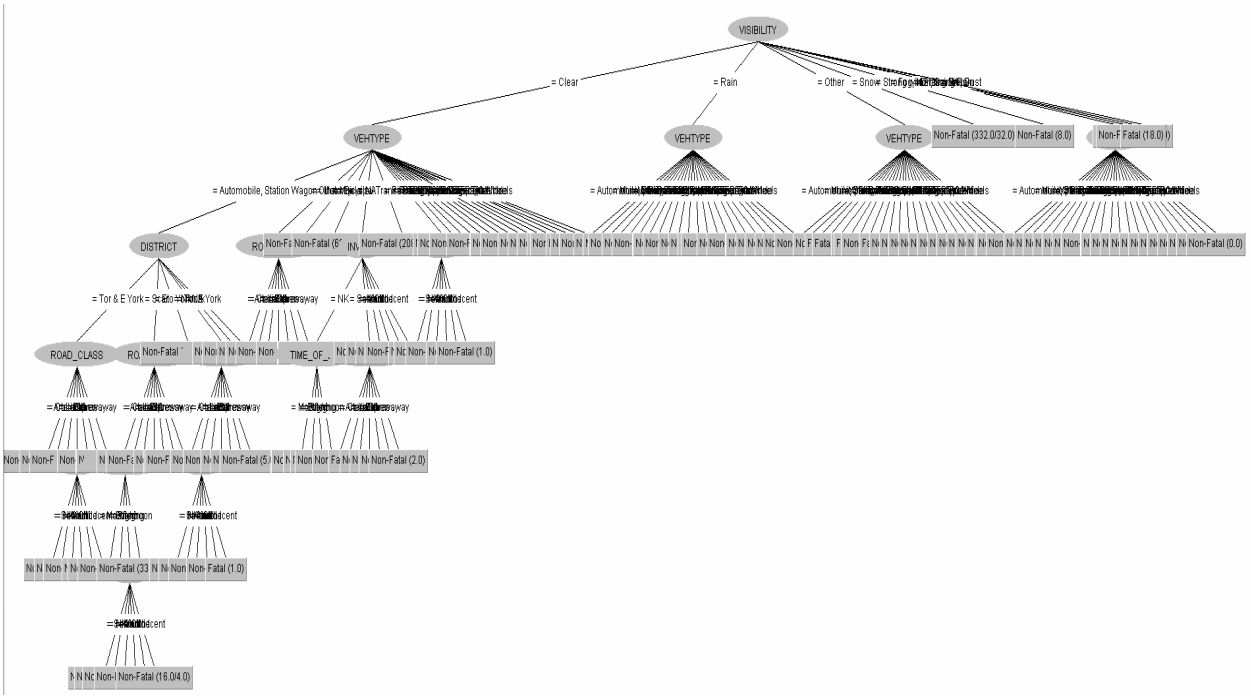
The results below show the clusters made from the provided data.

```
Number of iterations: 5
Within cluster sum of squared errors: 32420.0

Initial starting points (random):

Cluster 0: 'Tor & E York',Youth,Arterial,Rain,Night,'Automobile, Station Wagon'
Cluster 1: 'Eto York',Youth,NA,Clear,Morning,'Automobile, Station Wagon'
Cluster 2: 'Tor & E York',Adult,Arterial,Clear,Night,'Automobile, Station Wagon'
Cluster 3: 'Eto York',Adult,Arterial,Clear,Morning,'Passenger Van'
Cluster 4: 'N York',Senior,Arterial,Clear,Night,Other
```

For instance in Cluster 0 we can see that for Toronto and East York district majority of the collisions occurred in Youth age group on the **Arterial** roads at the **Night** time
And similarly results show the trends for other clusters as well.

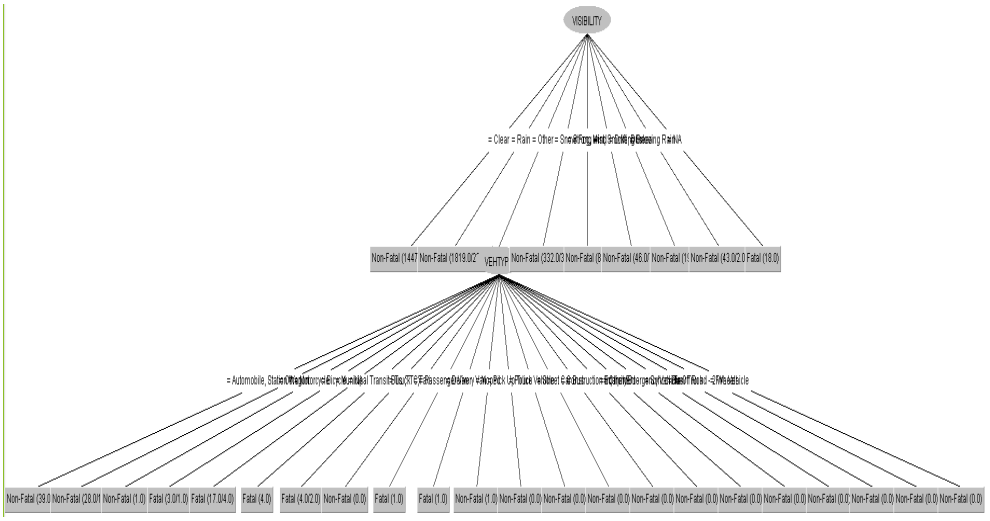
Decision Trees

The second algorithm that we used is the Decision Tree Classifier. This algorithm provides an advantage of easy interpretation by humans and from tracing the tree , it makes the task of making the decision easier.

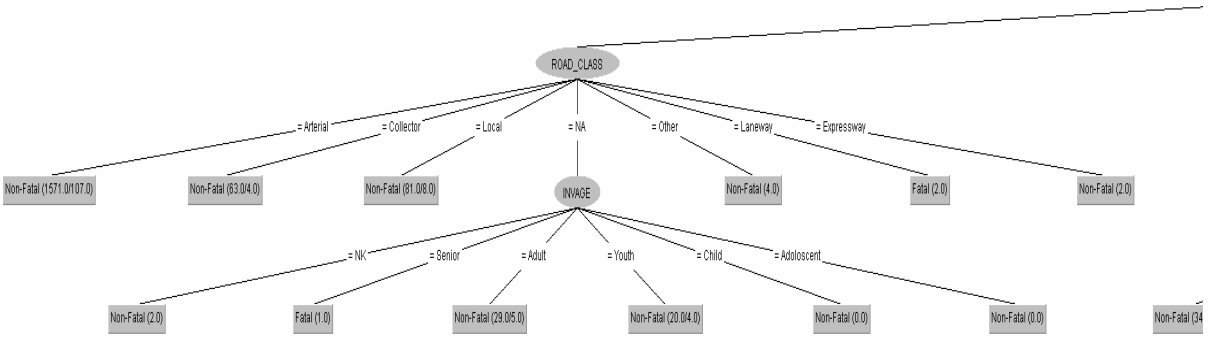


Since the above tree has many branches due to the complexity in the dataset, the picture is not very clear but the tree can be zoomed when the model file is opened in Weka.
The tree can be pruned by tweaking the default parameters of Weka.
With the following parameters, we constructed another tree.

Parameters : “Unpruned: False, confidence factor: 0.15 (15%), minNumObj: 25”



Confidence factor is used for better pruning the tree. Lower is the confidence factor , more is the pruning. minNumObj refers to the minimum number of instances per leaf. By tweaking the parameters the accuracy remained almost the same (<0.05% difference), but the path taken by the decision tree to classify instances is now clear.



Result from Binary Tree

```
Correctly Classified Instances       14557               86.3405 %
Incorrectly Classified Instances      2303               13.6595 %
Kappa statistic                          0.0454
Mean absolute error                      0.15
Root mean squared error                  0.2758
Relative absolute error                 95.4755 %
Root relative squared error             98.4333 %
Total Number of Instances            16860

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.033     0.005     0.487       0.033   0.062      0.099    0.616     0.226     Fatal
               0.995     0.967     0.867       0.995   0.926      0.099    0.616     0.903     Non-Fatal
               0.000     0.000     ?           0.000   ?          ?        0.989     0.008     Property Damage Only
Weighted Avg.  0.863     0.836     ?           0.863   ?          ?        0.616     0.810
```

From the above results, our decision tree is good enough to identify Non-Fatal Accidents, but not very good in identifying Fatal Accidents. This biassed nature of

prediction is due to the provided dataset where , about 86% of the total instances were Non-Fatal collisions .
This means that for example if out of 100 instances , if 95 instances are True and our tree classified all of them as True, it will be correct with 95% accuracy, It dosent mean that our accuracy is good, it is just that it is not able to predict false instances from the datset

```
=== Confusion Matrix ===

     a     b     c   <-- classified as
    76  2221     0 |     a = Fatal
    80 14481     0 |     b = Non-Fatal
     0     2     0 |     c = Property Damage Only
```

From the above confusion matrix, the decision tree is not able to predict Fatal instances with good enough accuracy and did not predict anything for Property Damage Only class and it classified almost all Fatal Instances as Non-Fatal for the reasons already mentioned above.

K-Nearest Neighbour

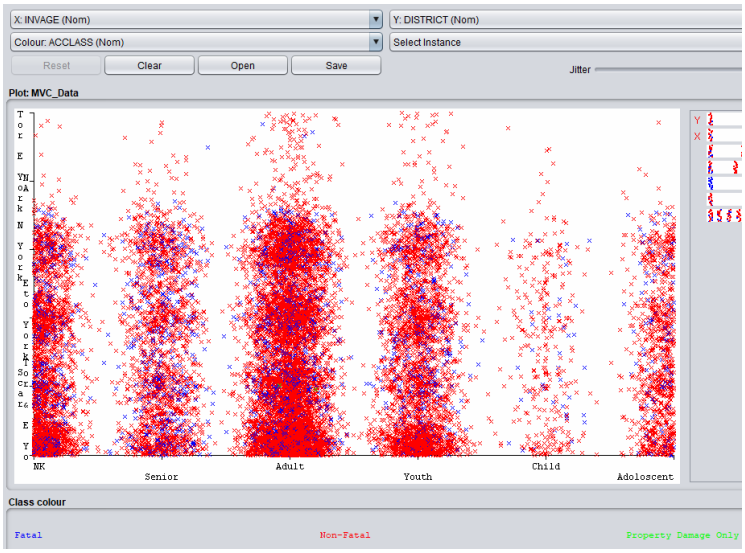
We performed Knn classification from classify tab. With cross fold validations set to 10 , we got the following results with different values of K

Knn	Correctly Classified Instance	Incorrectly Classified Instance	Confusion Matrix
1	85.1542 %	14.8458 %	a b c <-- classified as 143 2154 0 | a = Fatal 347 14214 0 | b = Non-Fatal 0 2 0 | c = Property Damage Only
3	85.7651 %	14.2349 %	a b c <-- classified as 67 2230 0 | a = Fatal 168 14393 0 | b = Non-Fatal 0 2 0 | c = Property Damage Only
5	86.0261 %	13.9739 %	a b c <-- classified as 34 2263 0 | a = Fatal 91 14470 0 | b = Non-Fatal 0 2 0 | c = Property Damage Only
7	86.1269 %	13.8731 %	a b c <-- classified as 20 2277 0 | a = Fatal 60 14501 0 | b = Non-Fatal 0 2 0 | c = Property Damage Only

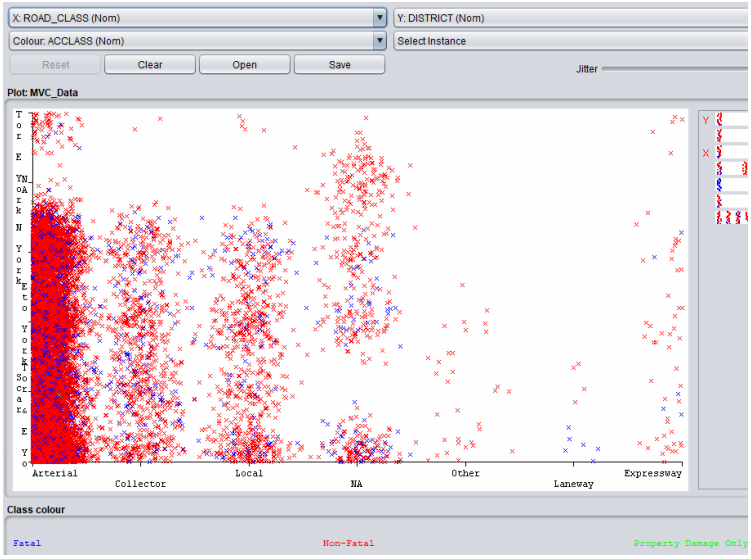
With confusion matrix , we can easily calculate the
Accuracy : a+b+c+d
Precision: a/(a+c)
Recall: a/(a+b)
Specificity: d/(b+d)

Interesting Graphs

During our analysis , we found some of the interesting graphs which are as below.



From the above graph we can see that, in all the collisions, adult age groups are the ones that are mostly impacted. This trend is common in almost all the cities. Child age group is least impacted than other age groups.

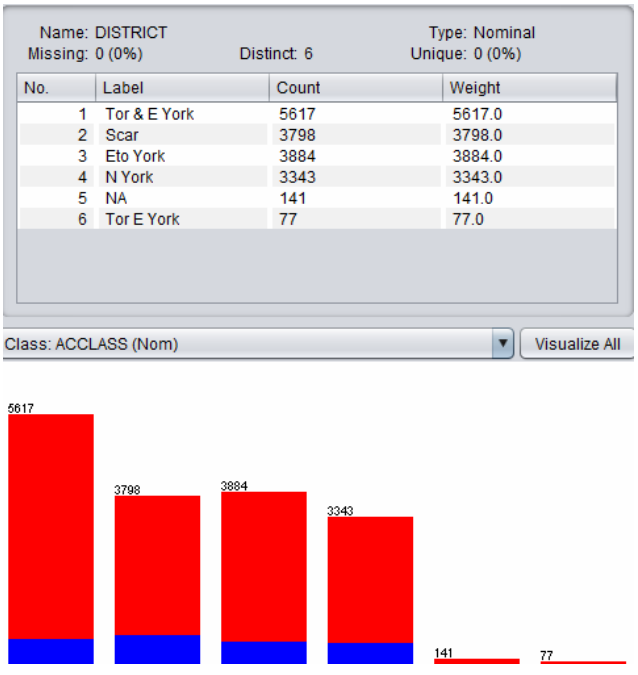


In the above graph of Road_Class vs District , we can see that almost all the collisions occurred on the Arterial road. It is also to be noticed that the density of blue regions (Fatal collisions) is almost the same for all road classes.

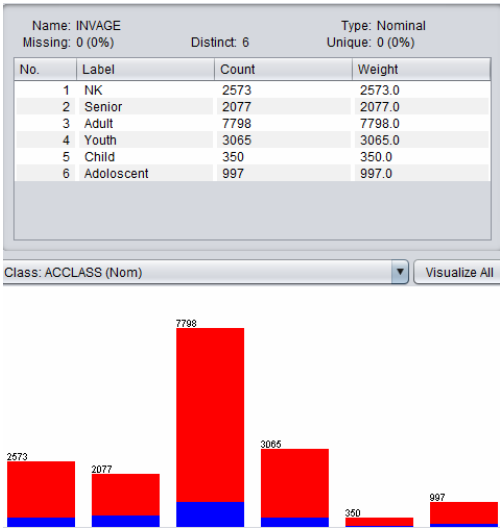
Results / Conclusion

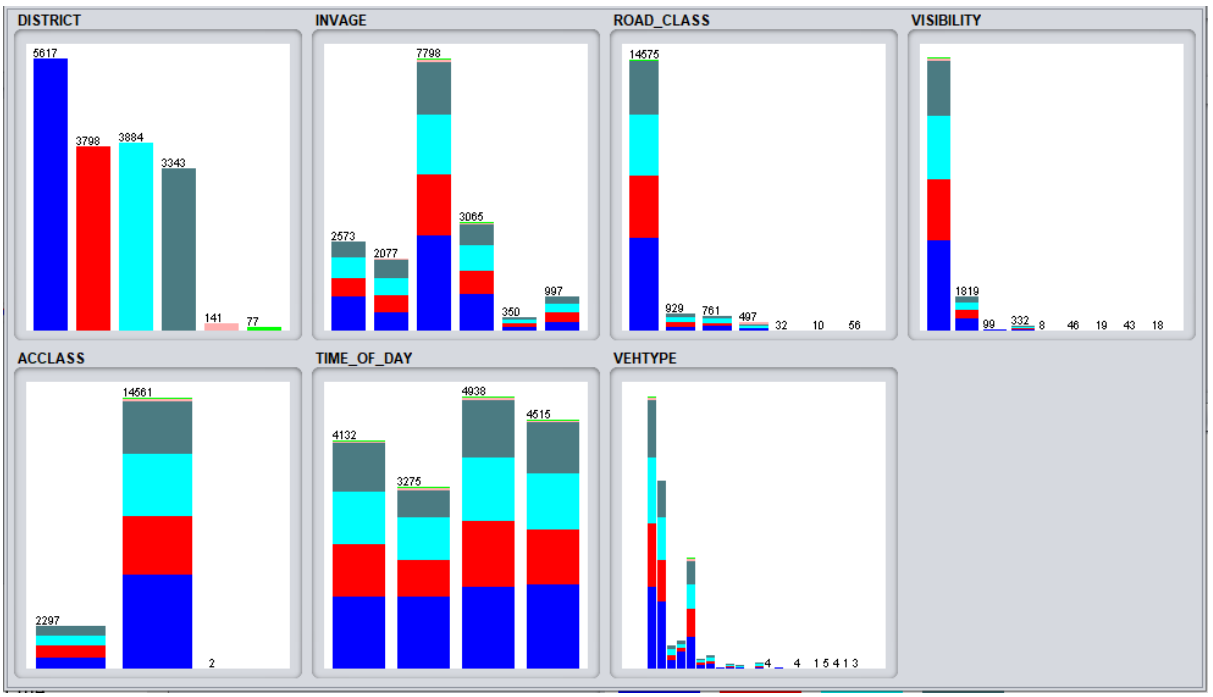
1. Visual Representation of Time of day shows that most of the collisions occurred during Evening. However the numbers for Morning and Afternoon are significantly similar while least collisions occurred at Night time.

2. Visual Representation of District shows that most of accidents occured in Toronto and East York. In fact Fatal accidents are significantly lower that Non-Fatal in all the cities



3. Visual Representation of INVAGE (Age of Involved Party) depicts that in most of the collisions , Adults (Age 30 -64) were the one that were more involved followed by Youth and then Seniors.





After performing the algorithms and along with the help of visualised results, we can determine the most common type of collision in Toronto and East York district with respect to type of road, visibility, time of the day and age group that is mostly impacted.

In the end we conclude that, in Toronto and East York District, trends show that non-Fatal accidents are more common in the city during the evening time on the Arterial Roads and youth generation is mostly the impacted age group which correctly aligns with our initial assumption.

References:

https://www.moseslawsc.com/blog/2021/july/when-and-where-most-car-accidents-occur/

https://open.toronto.ca/dataset/motor-vehicle-collisions-involving-killed-or-seriously-injured-persons/