CST8390 Final Project

2013 Tabular Transportation Collision Data Analysis Report

Injury collision is more likely to happen in which location under what situation?

Due: Dec 10, 2021 11:59 PM

Professor: Dr. Anu Thomas

Lab section: 011

Student1: Yueming Zhou

SN: 040999611

College Email: zhou0194@algonquinlive.com

Student2: Zhicheng Zhu

SN: 040898620

College Email: [zhu00129@algoqnuinlive.com](mailto:zhu00129@algoqnuinlive.com)

Contents

[1. Introduction 3](#_Toc89954487)

[2. Data Collection 3](#_Toc89954488)

[1.1 Data resource 3](#_Toc89954489)

[1.2 Row dataset detail 3](#_Toc89954490)

[3. Data Preprocessing 4](#_Toc89954491)

[4. Analysis & Result 7](#_Toc89954492)

[4.1 Decision tree 7](#_Toc89954493)

[4.2 kNN 9](#_Toc89954494)

[4.2.1 Initial Conclusion of kNN prediction 10](#_Toc89954495)

[4.3 Clustering 11](#_Toc89954496)

[4.4 Association Rule 13](#_Toc89954497)

[5. Conclusion 14](#_Toc89954498)

# Introduction

Transportation collision occurs every day. It brings property damages, injuries, and even fatal injuries. It is a potential tragedy for every family. We want to explore fatal transportation collision is more likely to happen in which location in Ottawa, and with what superimposed factors. This analysis can help the municipality to plan measurements that can be taken to prevent these possible transportation collisions. It can also let drivers pay attention to these influence factors in order to reduce the accident rate.

# Data Collection

## 1.1 Data resource

To perform the analysis, we look for the datasets that record transportation collisions in Ottawa. Open Ottawa is a website where contains all kinds of datasets provided by the City of Ottawa. It is a reliable resource. On this website, we found a csv file that records transportation collision data for the year 2013. Here is the link to the 2013 Tabular Transportation Collision Data: <https://open.ottawa.ca/datasets/ottawa::2013-tabular-transportation-collision-data/about>.

## 1.2 Row dataset detail

The 2013 Tabular Transportation Collision Dataset records 15,156 collisions in Ottawa of the year 2013. It contains 16 attributes about the details of the collision id, the street names of the collision location, the longitude and latitude, the x and y coordinate projected in MTM Zone 9, the date and time when the collision happened, the environment, the light condition, the initial impact type, the traffic control type and if they functioned normally, number of pedestrians, and the collision classification. The collisions are classified as fatal injury, non-fatal injury, and property damage only. More details about each attribute are shown in the following table.

Table 1. Description of each attribute of the dataset

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Description** | **Datatype** |
| COLLISION\_ID | Unique file id of the collision | TEXT |
| LOCATION | Street names of the location where the collision happened. (In forms of RD1 @ RD2 or RD from RD 1 to RD 2) | TEXT |
| X | X coordinator projected in MTM Zone 9, NAD83(CSRS) | Number |
| Y | Y coordinator projected in MTM Zone 9, NAD83(CSRS) | Number |
| LONGITUDE | Longitude of the collision location | Number |
| LATITUDE | Latitude of the collision location | Number |
| DATE | Date of the collision | Date or Time |
| TIME | Time of the collision | Date or Time |
| ENVIRONMENT | The environment type when the collision happened. Includes: unknown, Clear, Rain, Snow, Freezing Rain, Drifting Snow, Strong wind, and Fog/mist/smoke/dust. | TEXT |
| LIGHT | The light condition when the collision happened. Includes: unknown, daylight, dawn, dusk, and dark. | TEXT |
| SURFACE\_CONDITION | The surface condition includes: unknown, dry, wet, loose snow, slush, packed snow, ice, mud, loose sand or gravel, spilled liquid, and other. | TEXT |
| TRAFFIC\_CONTROL | Type of traffic control includes: traffic signal, stop sign, yield sign, school bus, traffic gate, traffic controller, no control, and roundabout. | TEXT |
| TRAFFIC\_CONTROL\_CONDITION | The traffic control condition includes: unknown, functioning, not functioning, and obscured. | TEXT |
| COLLISION\_CLASSIFICATION | Collisions are classified as: fatal injury, non-fatal injury, and P.D only (property damage only). | TEXT |
| IMPACT\_TYPE | Impact type includes: approaching, angle, rear end, sideswipe, turning movement, SMV unattended vehicle, SMV other, and other type. | TEXT |
| NO\_OF\_PEDESTRIANS | Number of pedestrians involved in the collision. 0,1,2,4 | NUMBER |

# Data Preprocessing

In this dataset, COLLISION\_ID is just the file id that is independent and identically distributed for a collision case, it is irrelevant to our analysis. There are 56% of the TRAFFIC\_CONTROL\_CONDITION values are unknown or missing, which may be a problem on the analysis accuracy. As for LONGITUDE and LATITUDE, they actually provide similar information as X and Y, it is not necessary to keep them. These attributes are removed at this point for performance and accuracy.

Some attributes are not making sense to our further analysis, we need to modify them to meet our requirements. For DATE, we divided all dates into two groups: weekdays or weekends. For TIME, we divided the time into 7 groups: Early Morning (6 a.m. to 9 a.m.), Morning (9 a.m. to 11 a.m.), Noon (11 a.m. to 14 p.m.), Afternoon (14 p.m. – 16 p.m.), Evening (16 p.m. to 20 p.m.), Night (20 p.m. to 24 p.m.), and Midnight (0 a.m. to 6 a.m.). For LOCATION, we create a new attribute ROAD\_TYPE to represent the type of road, if it is a highway or a normal road. Last, for COLLISION\_CLASSIFICATION, we modify it to represent if a case is Injury or P.D. only.

Then we load the file to WEKA, convert all the attributes to the right data type and apply removeDuplicates filter to remove the duplicates instances. After the filter, there are still 15,165 instances remaining in the dataset.

To analysis which location of Ottawa has most injury collisions, we need to create an attributes LOCATION to represent different districts of Ottawa. We create a new csv file that records X, Y, and COLLISION\_CLASSIFICATION and then load it to WEKA. According to the statistics from the SimpleKmeans performance, we can see in the following chart, the line flattens from the point of 5.

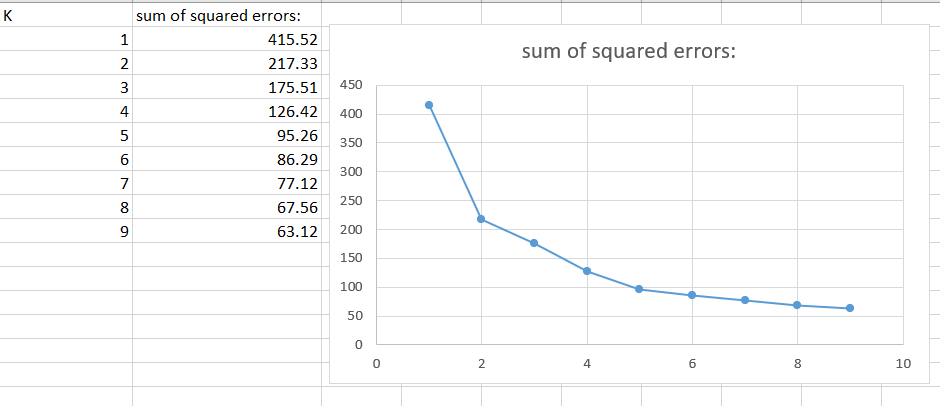


Figure 1. Find best k for location clustering

Visualize the clustering on WEKA, we notice that we can distinguish 5 districts of Ottawa. Cluster0 is Ottawa\_south, cluster1 is downtown, cluster2 is Ottawa\_west, cluster4 is mid-town, and cluster5 is Ottawa\_east.

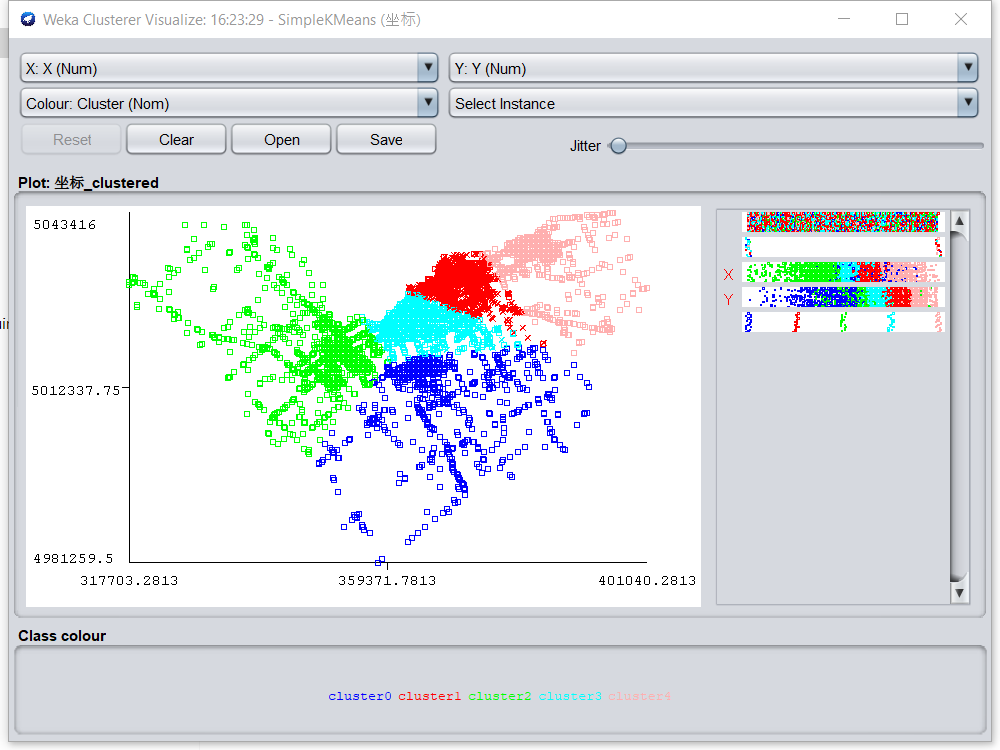


Figure 2. location clustering

A new attribute LOCATION is added to the dataset, and the X and Y attributes are removed. After the data processing, the dataset still remains 15,156 instances and 11 attributes.

Table 2. Statistic of processed dataset

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Label & count** | **Datatype** |
| **DATE** | Weekdays :11974  Weekends: 3182 | nominal |
| **TIME** | Early\_Morning: 375  Morning:201  Noon: 2066  Afternoon:1537  Evening: 3704  night:4864  Mid\_Night: 2409 | nominal |
| **Road\_type** | Road: 13022  highway: 2134 | nominal |
| **ENVIRONMENT** | 01-Clear: 11588  02-Rain:1420  03-Snow: 1798  04-Freezing Rain: 145  05-Drifting Snow: 86  06-Strong wind: 26  07-Fog, mist, smoke, dust: 50  00-Unknown: 43 | nominal |
| **SURFACE\_CONDITION** | 01-Dry: 9500  02-Wet: 2850  03-Loose snow: 1168  04-Slush: 458  05-Packed snow: 416  06-Ice: 692  07-Mud: 3  08-Loose sand or gravel: 18  09-Spilled liquid: 2  00-Unknown: 35  99-Other: 14 | nominal |
| **LIGHT** | 01-Daylight: 10622  03-Dawn: 353  05-Dusk: 682  07-Dark: 3368  00-Unknown: 131 | nominal |
| **TRAFFIC\_CONTROL** | 01-Traffic signal: 5579  02-Stop sign: 1477  03-Yield sign: 160  07-School bus:2  08-Traffic gate:13  09-Traffic controller: 9  10-No control: 7904  11-Roundabout: 6  Unknown: 6 | nominal |
| **IMPACT\_TYPE** | 01-Approaching: 215  02-Angle: 2058  03-Rear end: 5101  04-Slideswipe: 1821  05-Turning movement: 1693  06-SMV unattended vehicle: 1141  07-SMV other: 2761  99-Other: | nominal |
| **LOCATION** | downtown:1973  mid-town:1282  Ottawa\_south: 6287  Ottawa\_west: 1457  Ottawa\_east: 4157 | nominal |
| **COLLISION\_CLASSIFICATION** | Injury: 2795  P.D. only: 12361 | nominal |
| **NO\_OF\_PEDESTRIANS** | 0:14790  1:353  2:12  4:1 | nominal |

Before any analysis of this dataset, we assume that injury collision is more likely to happen in downtown at rush hour under bad weather.

# Analysis & Result

## Decision tree

Decision tree classifiers are one of the most popular and used classification techniques because the tree is constructed from the given data based on simple equations and uses the attribute selection measures such as a gain ratio measure, which ranks the attributes and determines the most useful attribute, and accordingly we can realize the most efficient attributes on the predicted purpose.

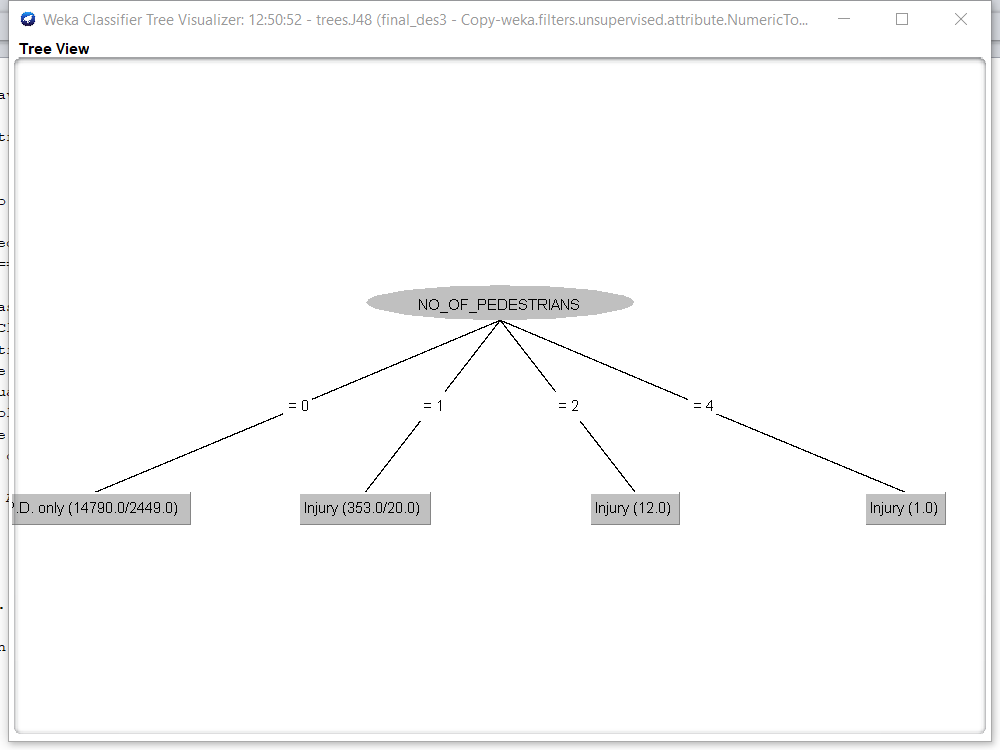


Figure 3. J48 Decision Tree 1

Based on the visualization graph of J48 analysis has pointed out that the most significant related to Collision Classification is the NO\_OF\_PEDESRIANS. However, since NO\_OF\_PEDESRIAN has unbalanced data, the result might be affected and inaccurate. We will temporarily remove this attribute to get another result of the J48 analysis.

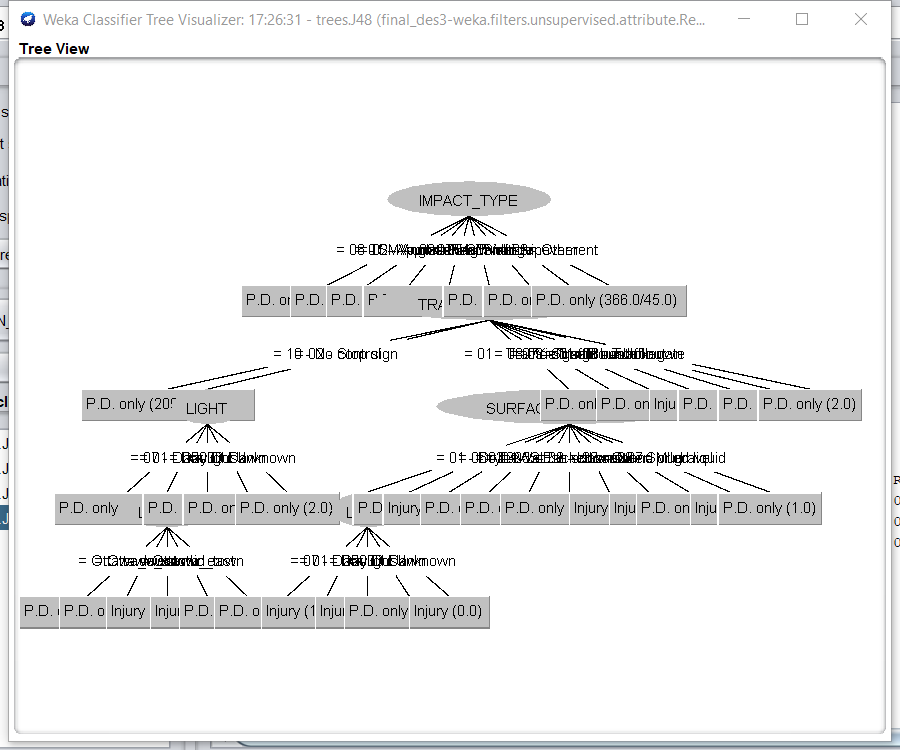


Figure 4. J48 Decision Tree 2\_1

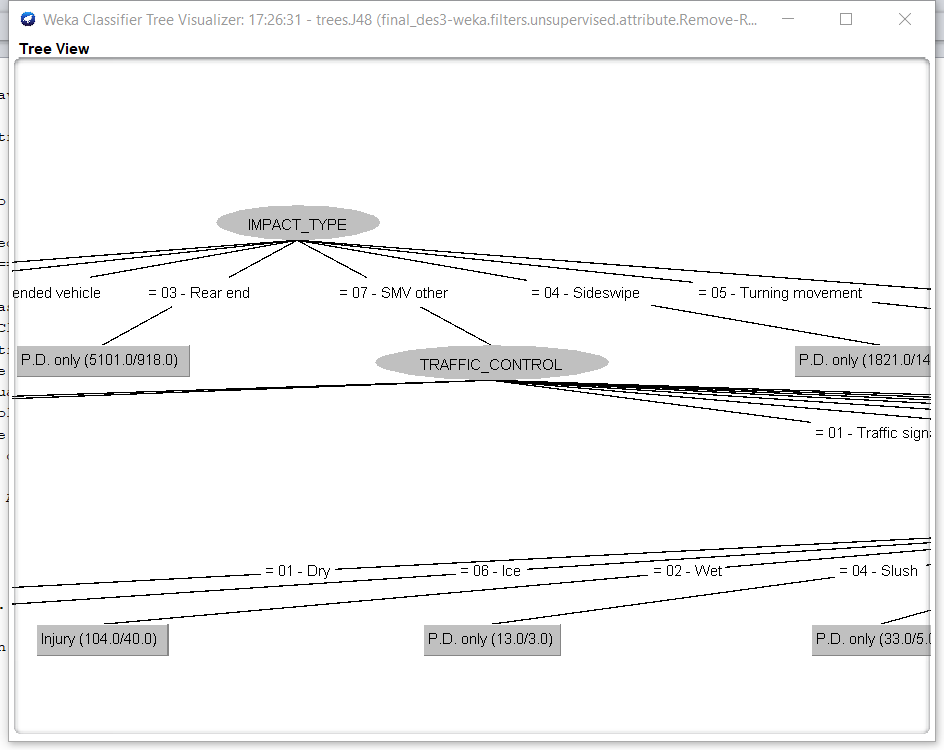


Figure 5. J48 Decision Tree 2\_2

Here is another J48 Decision Tree that we got from the datasets. From the graph, we can see the most relative attribute with Collision Classification is Impact type and the secondly related to it is Traffic control. Then, we will review the cluster analysis result in the further steps.

## kNN

KNN is one of the easiest classification algorithms. It is a method for classifying objects based on the closest exemplary instances in the attribute vector space. It helps us with the prediction.

Figure 6. TP rate of injury’s graph

Table

Description automatically generated*Figure 7. TP rate of injury’s chart*

A picture containing text

Description automatically generatedWe set the cross-validation to 10 folds to find the kNN classification. from the graph we could see the highest TP value is when k = 3. So, we decide to use the predictable result when k equals 3.

*Figure 8. Confusion matrix when kNN = 3*

Here is the confusion matrix when kNN = 3. We can see that there are 403 of injury instances has been correctly predicted and 12040 of P.D. only instances have been predicted.

## Initial Conclusion of kNN prediction

Graphical user interface, application

Description automatically generated

*Figure 9. Prediction result of kNN = 3*

Graphical user interface, text, application

Description automatically generated

*Figure 10. actual percentage of pedestrians, traffic signal and SVM. other*

Graphical user interface, text, application, Word

Description automatically generatedHere is some of the result found that about kNN= 3. We can predict the result of the top 3 attributes mostly related to collision type that we found in the decision tree. The overall number of instances is predicted as injury case is 724.

*Fig 11. kNN prediction – NO\_OF\_PEDESTRAINS vs. predicted COLLSION\_CLASSIFICATION*

The graph shows that the highest possibility of NO\_OF\_PEDESTRAINS having an injury traffic collision is the second column which is when there is one pedestrian. The prediction percentage is 318/724which is 43.92%.

Chart, bar chart

Description automatically generated

*Fig 12. kNN prediction – IMPACT\_TYPE vs. predicted COLLSION\_CLASSIFICATION*

The graph shows that the highest possibility of impact type to have injury traffic collision is the fifth column which is SVM - other. the prediction percentage is 391/724 which is 54.01%.

Chart

Description automatically generated

*Fig 13. kNN prediction – TRAFFIC\_CONTROL vs. predicted COLLSION\_CLASSIFICATION*

The graph shows that the highest possibility of impact type to have injury traffic collision is the fifth column which is SVM - other. the prediction percentage is 344/724 which is 52.49%.

## Clustering

Clustering is an unsupervised machine learning task. It involves automatically discovering natural grouping in data. Unlike supervised learning, clustering algorithms only interpret the input data and find natural groups or clusters in feature space. We use it to find a clustering that has the highest injury percentage and explore the clustering to find some underlying patterns.

|  |  |
| --- | --- |
| K | sum of squared errors: |
| 1 | 55908 |
| 2 | 50609 |
| 3 | 48473 |
| 4 | 47309 |
| 5 | 44171 |
| 6 | 42599 |
| 7 | 40687 |
| 8 | 39978 |
| 9 | 39213 |
| 10 | 38185 |

*Fig 14. sum of squared errors graph*

We are going to use k\_means clustering during the clustering process. The first step is choosing an appropriate k value We test the k value from 1-10 and here is the result of the sum of squared the errors. From the graph, we could see that when k = 6 the graph line goes stable. So we choose k = 6 to do the next step.

Graphical user interface, application

Description automatically generated

*Fig 15. K-means result*

Table

Description automatically generated

*Fig 16. Injury instances percentage in each cluster*

Here is the result of the data after clustering. after the clustering process, we calculate the injury case concluded in each cluster. We found that cluster 5 has the highest injury percentage. Then we choose cluster 5 as a sample to find out the instances that weighted highest in each of attribute. The result shows as follows:

Graphical user interface, table, Excel

Description automatically generated

*Fig 17. highest Injury instances in cluster 5*

From Fig 15 we could see that the number of injuries is denser when there is 1 pedestrian than the P.D. only, and the P.D. only is denser in pedestrian 0. Seems like the number of pedestrians indeed effect the injury instances. And injury cases are more likely to happen when there is a pedestrian.

Graphical user interface, application

Description automatically generated with medium confidence

*Fig 18. Simple k-means. COLLSION\_CLASSIFICATION vs NO\_OF\_PEDESTRAINS*

## Association Rule

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently an item set occurs in a transaction. We are using it to figure out what are sets of conditions can lead to an injury collision.

We are using the Apriori algorithm to run the Association process. We set the class index to be 9 which indicate to collision classification attribute and min metric as 0.1. The result shows as follows.

A picture containing text

Description automatically generated

*Fig 19. Aprior Result of Injury*

Then, we got the top 10 results that associate with injury, and the highest conf. value is 0.2. The result is out of our expectations, the injury type always appeared with the mild weather and environmental condition.

# Conclusion

Table

Description automatically generated

The prediction result is getting a bit higher than the actual result. However, the accuracy of the overall result is high. We can find that most of the correct classified instances are P.D. property instead of injury. Our conclusion is, even though the KNN provides high accuracy, but the disadvantage is it does not always make sense for each attribute.

From the result of clustering, we could see that injury collision is more likely to happen on weekdays, in downtown. And from the result of association rules, it is out of our expectations, the injury type always appeared with the mild weather and environmental condition.

Under all the analysis results, our prediction is that injury collision is more likely to happen downtown on weekdays under mild weather and environmental conditions, and also, the most like to affect the injury collision is impact type of SMV -other and the traffic control of traffic signal，and also the number of pedestrians will also affect the result especially affect the injury result when there is upper than 1 pedestrian. However, due to the NO\_OF\_PEDESTRAINS having imbalanced data, the result might be inaccurate.

For the further effect of society, it could be a suggestion for government to make sure traffic signals work well all the time and a suggestion for the driver to be more careful to drive under this kind of situation. And the driver better more carefully check if there are pedestrians passing the intersection.