**Group Members:**

Diya Valand (041101803)

Arshpreet Kaur (041100889)

Harsimranjit Singh (041100738)

Due date: 15th March

**Final Report**

**Professor: Anu Thomas**

**Integrated Data Analysis for Transportation Safety Enhancement**

[Introduction 4](#_Toc162648512)

[Business Understanding 4](#_Toc162648513)

[Project Plan: 5](#_Toc162648514)

[Data Understanding 5](#_Toc162648515)

[Collect Data 5](#_Toc162648516)

[Describe Data 6](#_Toc162648517)

[Explore Data 8](#_Toc162648518)

[Verify Data Quality 11](#_Toc162648519)

[Data Preparation 12](#_Toc162648520)

[Select Data 12](#_Toc162648521)

[Clean Data 13](#_Toc162648522)

[Construct Data 13](#_Toc162648523)

[Integrate Data 14](#_Toc162648524)

[Format Data 14](#_Toc162648525)

[Data Modeling 16](#_Toc162648526)

[Classification 16](#_Toc162648527)

[Decision tree 16](#_Toc162648528)

[Random Forest 16](#_Toc162648529)

[kNN 17](#_Toc162648530)

[Clustering 18](#_Toc162648531)

[Elbow Method 18](#_Toc162648532)

[K-Means 22](#_Toc162648533)

[Association (apriori) 23](#_Toc162648534)

[Outlier Detection 24](#_Toc162648535)

[LOF 24](#_Toc162648536)

[ISF 26](#_Toc162648537)

[Discussion of results 27](#_Toc162648538)

[Clustering 28](#_Toc162648539)

[Association 29](#_Toc162648540)

[Outlier Detection 29](#_Toc162648541)

[Conclusion 30](#_Toc162648542)

[References 31](#_Toc162648543)

# Introduction

In the modern era, urban transportation systems play a pivotal role in shaping the dynamics of cities, impacting not only mobility but also public safety. Ottawa, being a bustling urban center, is no exception to the complexities and challenges inherent in managing transportation networks. Understanding the intricacies of transportation collisions is crucial for policymakers, city planners, and law enforcement agencies to implement effective measures aimed at mitigating risks and enhancing road safety. Through this project, we endeavor to delve deep into the wealth of data encapsulated within the 2020 Tabular Transportation Collision Data for Ottawa, unraveling the underlying factors contributing to various types of transportation collisions. By employing advanced data science techniques, we aim to illuminate the intricate web of interactions between environmental conditions, road infrastructure, driver behavior, and collision outcomes, thus paving the way for evidence-based interventions to foster a safer and more sustainable transportation ecosystem in Ottawa.

(-Harsimranjit Singh)

# Business Understanding

Our project seeks to analyze the 2020 Tabular Transportation Collision Data from Ottawa with the overarching goal of enhancing transportation safety. By dissecting the underlying factors influencing collision occurrences and outcomes, we aim to provide stakeholders with actionable intelligence to reduce collision rates, mitigate injury severity, and improve overall road safety. This endeavor holds significant implications for city planners, transportation agencies, law enforcement, healthcare providers, and the broader community, as it enables targeted interventions, informed decision-making, and resource allocation. Through collaborative efforts grounded in methodological rigor and effective communication, we endeavor to deliver a comprehensive analysis that empowers stakeholders to foster a safer and more resilient transportation ecosystem in Ottawa and answer the question, “what are the contributing factors for collision to end up in fatality”.

(-Harsimranjit Singh)

### Project Plan:

|  |  |
| --- | --- |
| Name | Task |
| Diya Valand | Classification |
| Arshpreet Kaur | Outlier Detection |
| Harsimranjit Singh | Clustering |

# Data Understanding

## Collect Data

The 2020 Tabular Transportation Collision Data provides valuable insights into transportation collisions that occurred in Ottawa during the year 2020. This dataset contains detailed information about each collision, including the date, time, location description, classification of collision, road surface condition, environment, light conditions, initial impact type, traffic control, and various injury-related metrics.

To collect data from this dataset, researchers or analysts can access the dataset through the provided [2020 Tabular Transportation Collision Data | Open Ottawa](https://open.ottawa.ca/datasets/ottawa::2020-tabular-transportation-collision-data/about) and extract the relevant information based on their research or analysis objectives. They can use tools such as KML and CSV/XLS formats to access the data in latitude and longitude coordinates or MTM Zone 9, NAD83 (CSRS) projection format.

(-Arshpreet Kaur)

## Describe Data

The dataset contains 10047 instances with 28 attributes related to 2020 Tabular Transportation Collision Data.

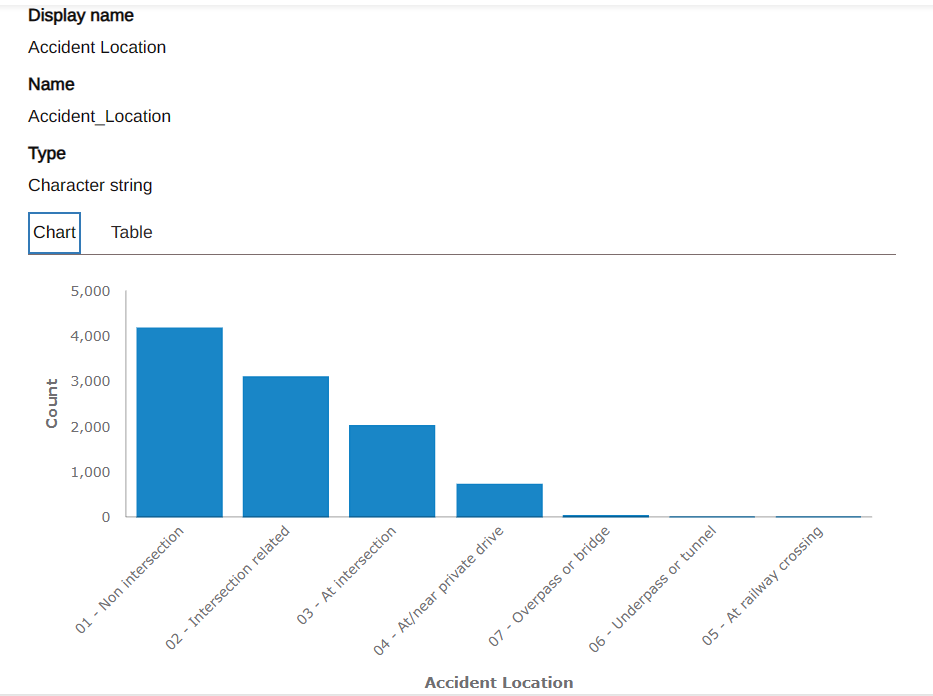
|  |  |  |
| --- | --- | --- |
| Attributes | Data type | Description |
| ObjectID | Integer | Id of Object |
| Accident\_Date | Date-Time | Date |
| Accident\_Time | Nominal | Time |
| Location | Nominal | Location description (RD1 @ RD2 or RD from RD 1 to RD 2) |
| Accident\_Location | Nominal | Collision location (Intersection, non-intersection, at/near private driveway) |
| Classification\_of\_Accident | Nominal | Classification of collision (non-fatal, fatal, property damage only) |
| Environment\_Condition | Nominal | Environment (Clear, rain, snow…) |
| Light | Nominal | Light (daylight, dawn, dusk…) |
| Road\_Surface\_Condition | Nominal | Road surface condition (Ice, wet, dry snow...) |
| Traffic\_control | Nominal | Traffic control (stop, traffic signal, no control…) |
| Max\_injury | Nominal | Max Injury (Highest injury level in the collisions) |
| No\_of\_Injuries | Integer | No. of Injuries |
| Anom ID | Nominal |  |
| Geo\_ID | Nominal |  |
| Initial Impact Type | Nominal | Initial impact type (Angle, turning movement, rear-end…) |
| Traffic Control Condition | Nominal |  |
| No. of Vehicles | Integer | No. of Vehicles |
| No. of Motorcycles | Integer | No. of Motorcycles |
| No. of Bicycles | Integer | No. of Bicycles |
| No. of Pedestrians | Integer | No. of Pedestrians |
| No. of Minimal | Integer | No. of Minimal (Person did not go to hospital when leaving the scene of the collision) |
| No. of Minor | Integer | No. of Minor (Person went to hospital and was treated in the emergency room, but not admitted) |
| No. of Major | Integer | No. of Major (Person admitted to hospital. Includes person admitted for observation. This could be either life threatening or non-life threatening) |
| No. of Fatal | Integer | No. of Fatal (Person killed immediately or within 30 days of the motor vehicle collision) |
| X | Real | X and Y coordinate format is projected in MTM Zone 9, NAD83 (CSRS) |
| Y | Real | X and Y coordinate format is projected in MTM Zone 9, NAD83 (CSRS) |
| Latitude | Real | Latitude |
| Longitude | Real | Longitude |

(-Arshpreet Kaur)

## Explore Data

1. This bar chart shows the number of transportation collisions by location type, with most occurring at non-intersection locations, followed by intersection-related and private drive or alley incidents, and the fewest at overpasses, bridges, underpasses, tunnels, and railway crossings.

(-Diya Valand)



1. The bar chart represents transportation collision data categorized by environmental conditions, indicating that most collisions occurred in clear conditions, with significantly fewer in snow and rain, and the least in other conditions like freezing rain, fog, or unknown situations.

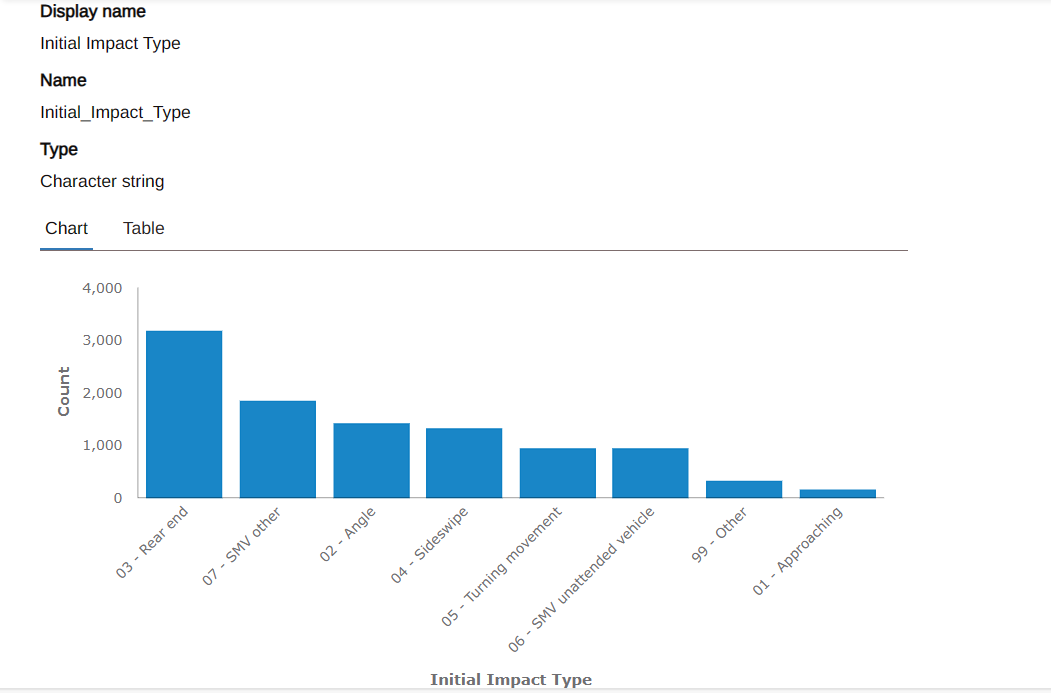
(-Arshpreet Kaur)

A graph with blue squares

Description automatically generated with medium confidence

1. The bar chart details transportation collisions by initial impact type, with the highest number of collisions involving rear-end impacts, followed by other specified impacts, angle collisions, and sideswipe impacts, with approaching impacts and collisions with unattended vehicles being the least frequent.

(-Harsimranjit Singh)



## Verify Data Quality

To verify data quality, we must ensure the accuracy, completeness, and reliability of the 2020 Tabular Transportation Collision Data from Ottawa. This includes cross-checking the correctness of information, such as accident dates, times, and locations, assessing consistency in data entry for attributes like road conditions and environmental factors, and confirming there are no duplications or missing values that could skew the analysis. In Max\_Injuries we have 8280 Missing values. There are some outliers in the selected attributes.

(-Arshpreet Kaur)

# Data Preparation

## Select Data

We will use all 10047 instances in the dataset, with no exclusions, but choose only the following 12 attributes for analysis.

Select Attributes: This step selects a subset of attributes from the retrieved data, including attributes like Accident\_Date, Accident\_Location, Accident\_Time, etc.

Set Role: This operator sets the role of the attribute "ObjectId" as the identifier.

|  |  |  |
| --- | --- | --- |
| Attributes | Data type | Description |
| ObjectID | Integer | Id of Object |
| Accident\_Date | Date-Time | Date |
| Accident\_Time | Nominal | Time |
| Location | Nominal | Location description (RD1 @ RD2 or RD from RD 1 to RD 2) |
| Accident\_Location | Nominal | Collision location (Intersection, non-intersection, at/near private driveway) |
| Classification\_of\_Accident | Nominal | Classification of collision (non-fatal, fatal, property damage only) |
| Environment\_Condition | Nominal | Environment (Clear, rain, snow…) |
| Light | Nominal | Light (daylight, dawn, dusk…) |
| Road\_Surface\_Condition | Nominal | Road surface condition (Ice, wet, dry snow...) |
| Traffic\_control | Nominal | Traffic control (stop, traffic signal, no control…) |
| Max\_injury | Nominal | Max Injury (Highest injury level in the collisions) |
| No\_of\_Injuries | Integer | No. of Injuries |

(-Diya Valand)

## Clean Data

Cleaning the data from the 2020 Tabular Transportation Collision Dataset is imperative to ensure its reliability and accuracy for meaningful analysis. This process involves identifying and rectifying errors, inconsistencies, and missing values within the dataset. By accessing data quality, handling missing values (here we are replacing missing values in max\_injury by unknown), normalizing data (to ensure that all attributes are on a similar scale), although we don’t have any duplicates to remove. By this analysts can obtain a dataset that is consistent, complete, and ready for analysis.

(-Diya Valand)

## Construct Data

We will construct new attributes for this dataset. Firstly, we will converts the attribute "Accident\_Date" from a date format to a nominal format, possibly for easier handling or processing. Then we will split the attribute "Accident\_Time" into multiple attributes, possibly to extract specific time-related information. Lastly we will renames the split attributes to denote parts of the accident date, like day, month, and year.

(-Diya Valand)

## Integrate Data

We will not integrate this dataset with any other datasets.

(-Diya Valand)

## Format Data

Nominal to Numerical: Converts several nominal attributes into numerical format, possibly for compatibility with certain algorithms.

(-Diya Valand)

For clustering and distance-based calculations, we typically want to choose attributes that provide meaningful information about the characteristics of the data points and are suitable for measuring similarity or dissimilarity between data points. In the context of transportation collision data and focusing on clustering or distance-based analysis, we are considering attributes such as:

Location: Geographic coordinates or categorical descriptions of the collision location could be crucial for identifying clusters of accidents in specific areas.

Accident\_Date and Accident\_Time: Temporal information can help identify patterns based on when accidents occur, such as clustering accidents during certain times of the day, days of the week, or seasons.

Classification\_of\_Accident: The type or severity of the accident could be informative for clustering similar types of accidents together.

Environment\_Condition, Light, and Road\_Surface\_Condition: These attributes provide insights into the environmental factors affecting accidents and can help identify clusters based on similar environmental conditions.

Traffic\_Control: Clustering accidents based on the presence or absence of traffic control measures can reveal patterns related to intersections, traffic signals, or other regulatory factors.

Max\_Injury and No\_of\_Injuries: While these attributes are more directly related to the severity of accidents, they can also be informative for clustering accidents based on their impact.

ObjectId: Depending on its nature, this attribute might serve as an identifier for individual collisions and may not be directly relevant for clustering or distance-based calculations.

When choosing attributes for clustering or distance-based analysis, we find it essential to consider the characteristics of the data and the objectives of our analysis.

For **Outlier Detection** and **Clustering**, we use normalized data. For **Classification**, we will use the same normalized data by applying the numeric to polynomial operator to convert it into categorical data for the decision tree.

(-Harsimranjit Singh)

# Data Modeling

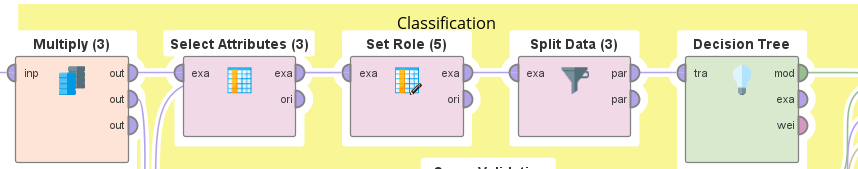
## Classification

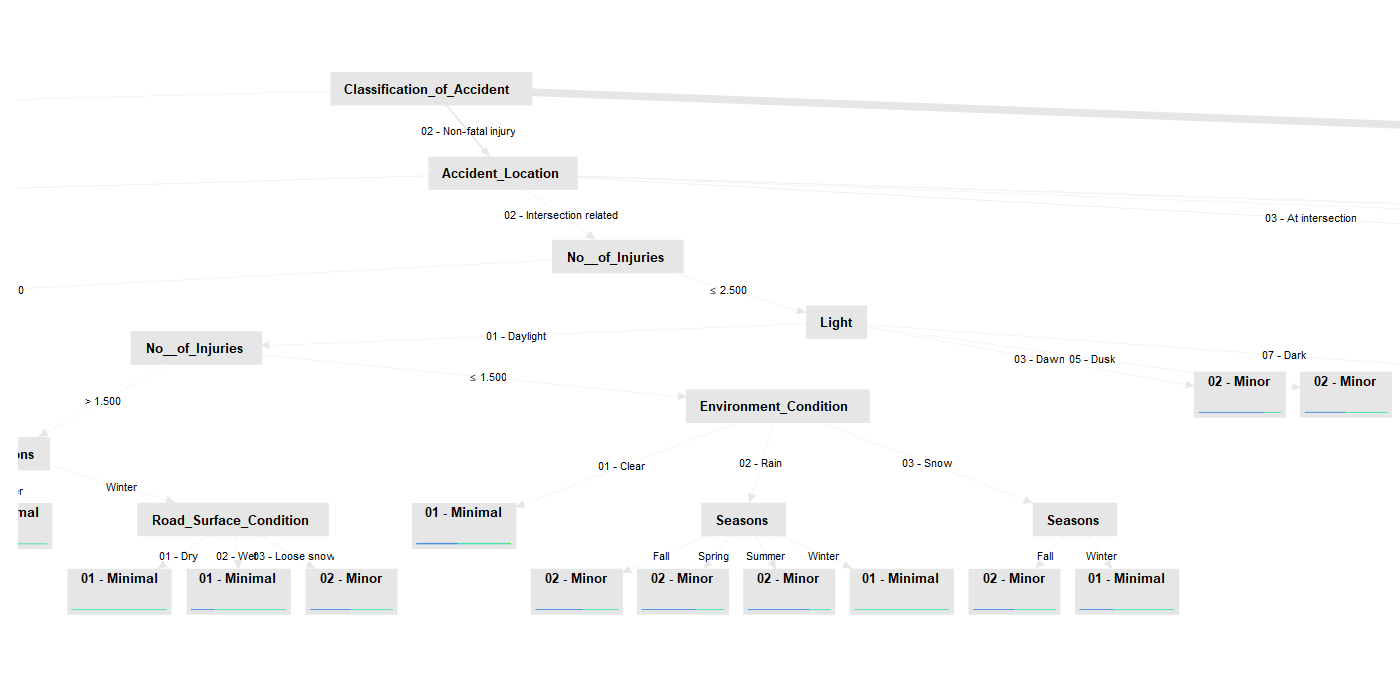
Classification means predicting the classified data, it involves three main methods: kNN, Decision tree, and Random Forest.

### Decision tree

In the Decision tree method, we split data into 70% for training purposes and 30% for testing purposes.

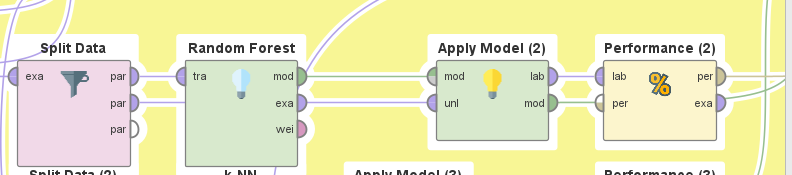
Here, we have provided one image which illustrates how Seasons, Environment\_Condition, Light, Road\_Surface\_Condition, and No\_Of\_Injuries are important factors, that are main reasons to reduce Number of Accidents.

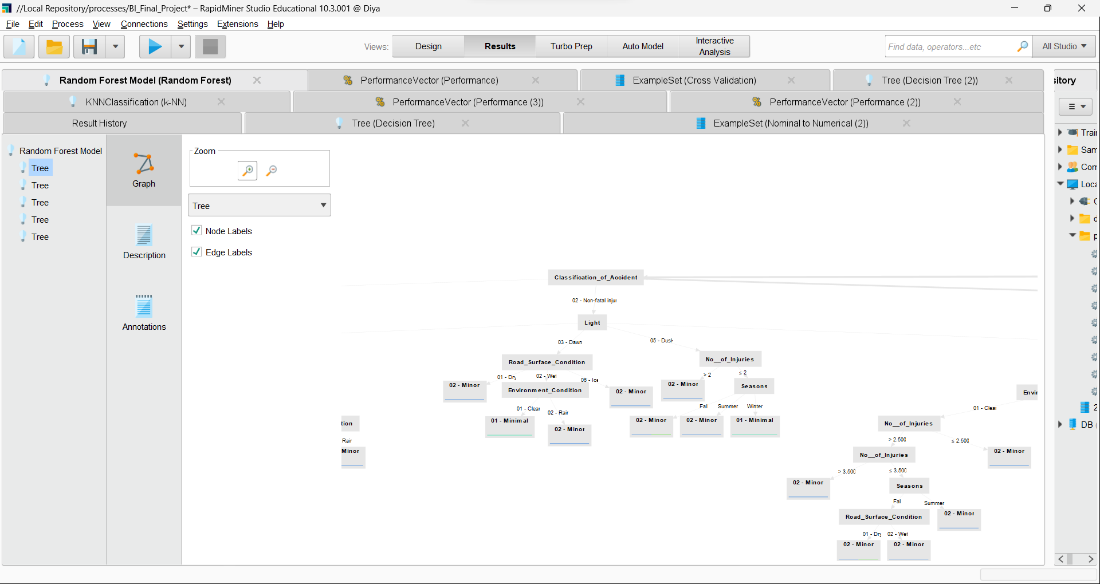




### Random Forest

This is Random Forest tree, which is being used to classify accidents based on conditions like light, road surface, Seasons and the number of injuries. This Random Forest contains 5 trees. The tree segments accidents into categories such as "Non\_Fatal\_Injury" with further subdivisions based on environmental and temporal conditions to refine the classification.

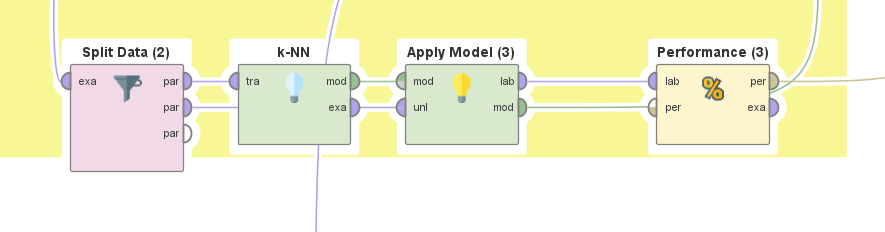




### kNN

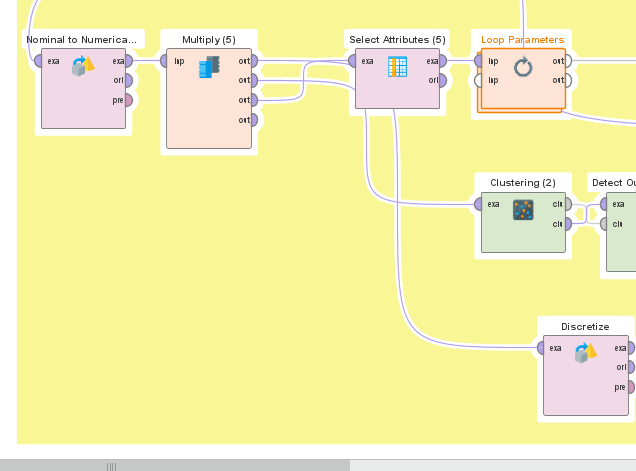
In kNN method we have used 70% data for training set and 30% data for testing set which results into model like this:

**Weighted 15-Nearest Neighbour model for classification.**  
**The model contains 7033 examples with 42 dimensions of the following classes:**  
 **02 - Minor**  
 **01 - Minimal**  
 **03 - Major**  
 **04 - Fatal**  
 **Unknown**



## Clustering

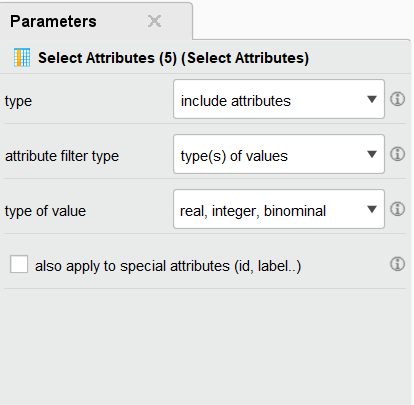
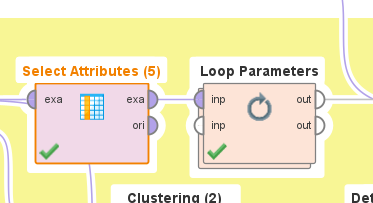
In the modeling phase for clustering, I begin by using the "Numeric to Nominal" operator to format the data appropriately. This preprocessing step is pivotal for clustering algorithms such as K-Means, which necessitate numerical data to compute the distances between points effectively. By converting categorical data into a numerical format, this operator ensures that all data types are suitably prepared for the algorithm, enhancing the accuracy and efficiency of the clustering process. Later I connect the operator to the multiply operator so that I can use it for K-mean iteration to find out elbow point for best k value, for k-mean Clustering outlier Detection and for Association rule.



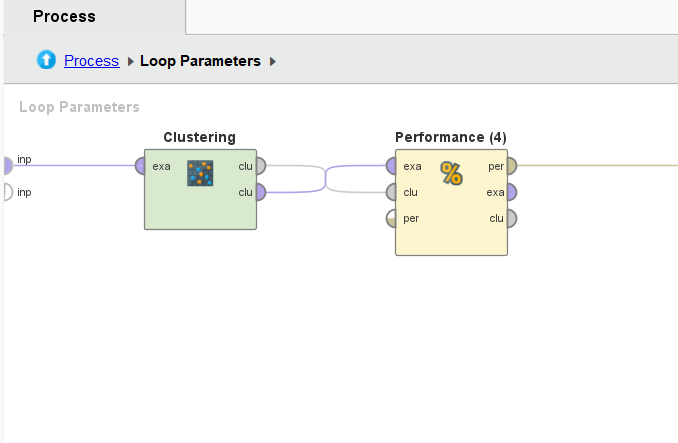
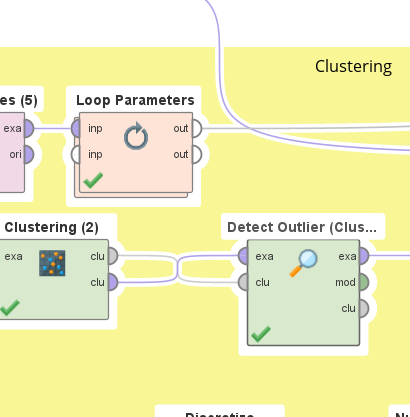
@Harsimranjit Singh

### Elbow Method

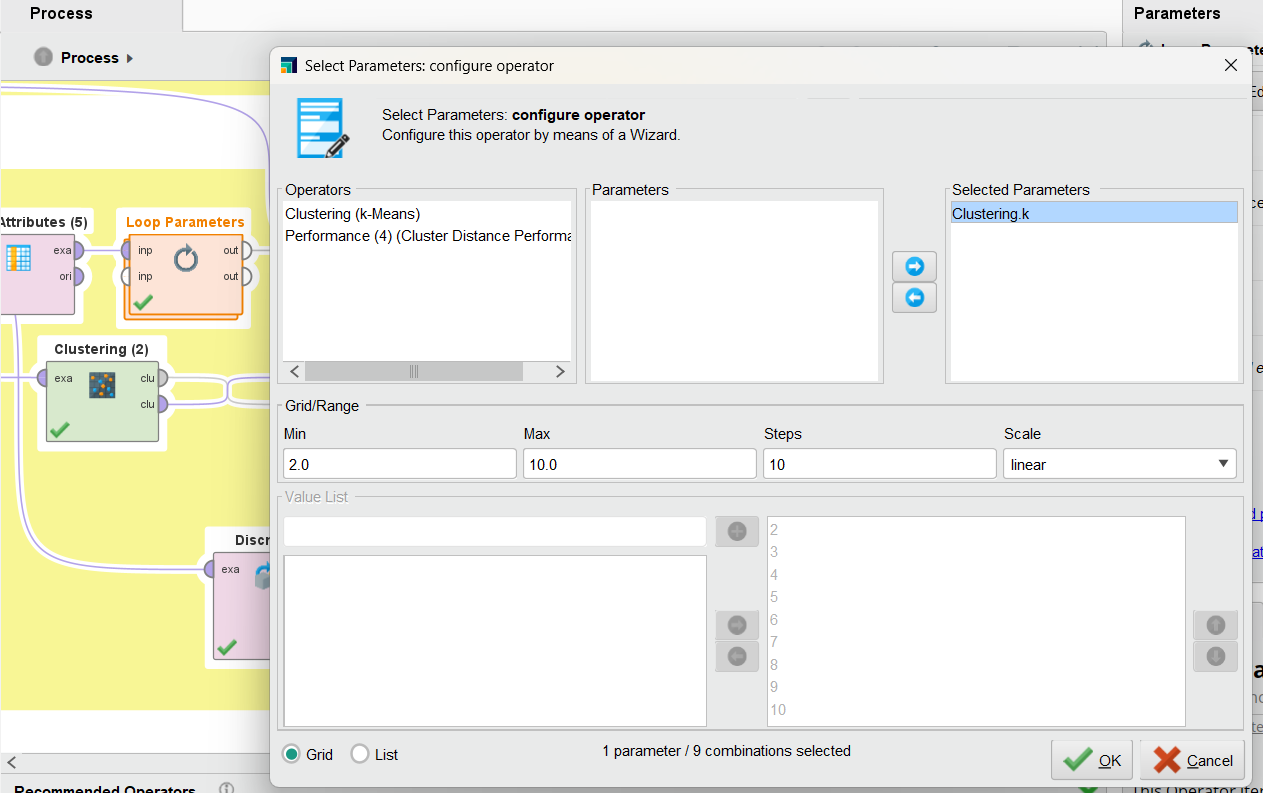
Finding the Elbow Point is a crucial step in cluster analysis, especially when employing the k-means clustering algorithm. This method involves determining the optimal number of clusters (k) that best represents the underlying patterns in the data. To achieve this, I utilized a technique involving the **Loop parameter operator**, focusing on **k-means Clustering** and evaluating the performance through **Cluster Distance Performance** metrics. Here's an expanded and refined explanation of the process:



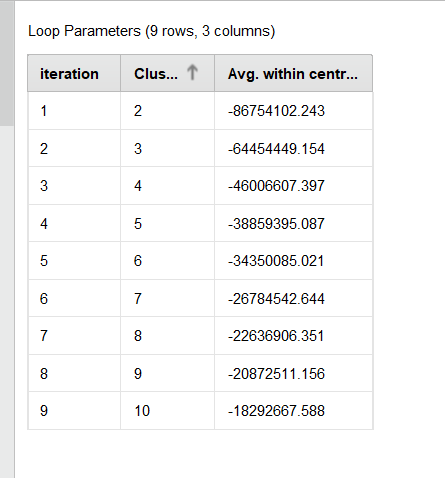
So first I used the select attribute and select the Value of type ,which are Real , Interger and bionominal.



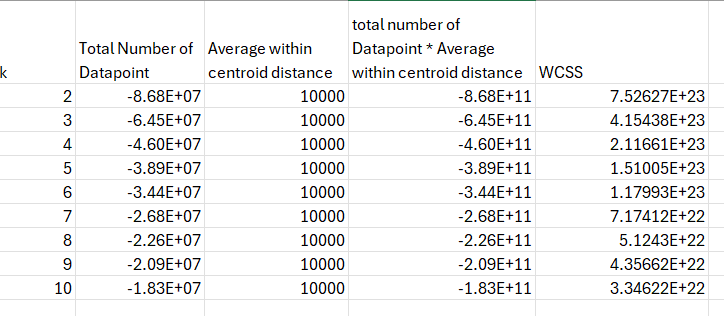
then I began by setting the parameters for the k-means algorithm, specifically the range of potential clusters. The minimum number of clusters was set to 2 (Minimum = 2), acknowledging that at least two groups exist within the data. The maximum boundary was defined as 10 (Maximum Range = 10), limiting the exploration to a manageable number of clusters and avoiding excessive computation.



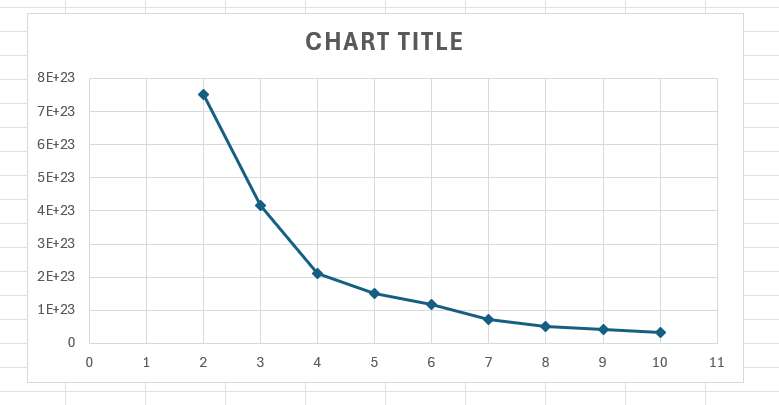
After running the Loop Parameter, we got the result for the K and it Avg of within Centroid Distance



We utilized this data in an Excel sheet to compute the **Within-Cluster Sum of Squares** (WCSS) and generated the Elbow Diagram



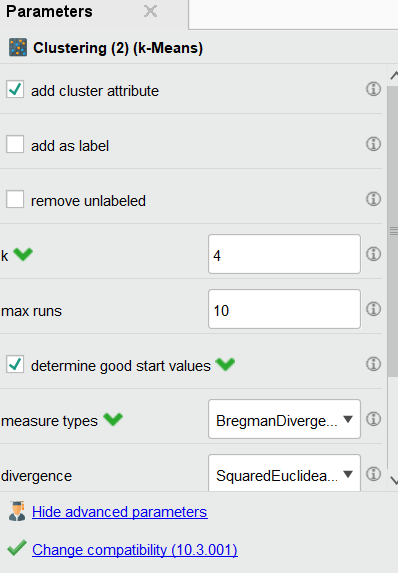
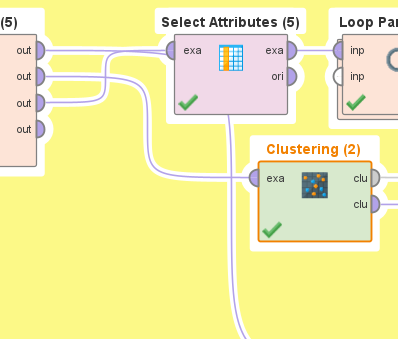
We created a line graph from the Within-Cluster Sum of Squares (WCSS) and the corresponding K points to observe the optimal elbow point.



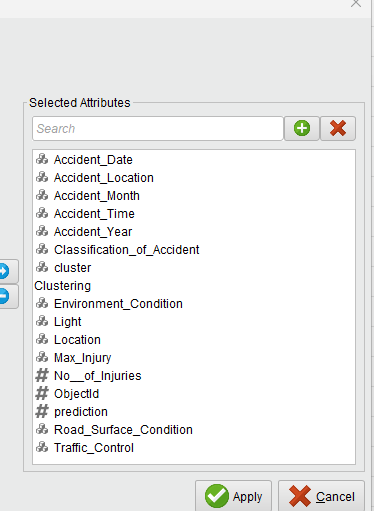
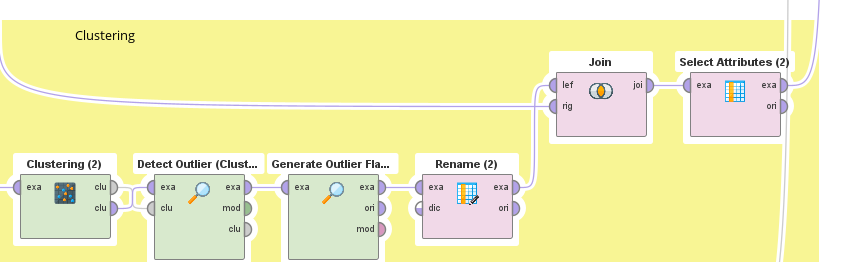
Here, we discover that K=4 is the optimal value because beyond 4, the dataset exhibits greater stability. This value will be employed in both clustering and outlier detection flag improvement.

### K-Means

For K-means clustering, I initiated the process by utilizing the data previously formatted with the Nominal to Numeric Operator. Then, I employed the K-means Clustering operator and set the value of K = 4, as determined from the elbow method.



After that, I employed the "Detect Outlier (Clustering)" operator, keeping the default parameters. Then, I connected it to the "Generate Outlier Flag" to extract outliers from the clustered dataset. Next, I utilized the "Rename" operator to rename the outlier flag to "Clustering Outlier" and proceeded to use the "Join" operator to combine the clustered data with the prepared data. Finally, I employed the "Select Attribute" operator to obtain the selected result focusing on the condition for accidents.

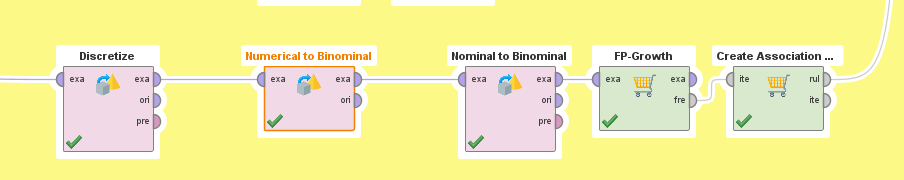


### Association (apriori)

So, to Use Apriori algorithm, I get to know that we can use FP-Growth instead of which work fine for Association rule [1].

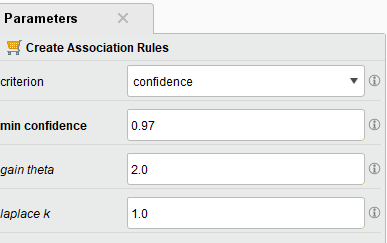
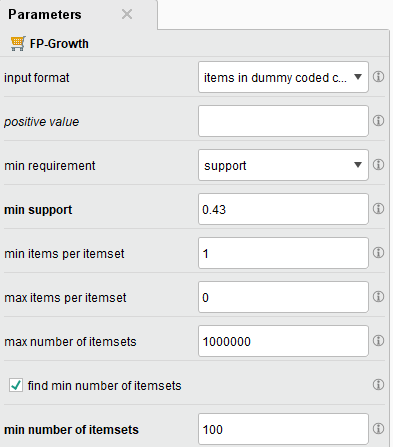
For the process,

I use **Discretize** (by frequency) and set Number of bins =2 then I formatted the data by using **Numeric to Binominal** and **Nominal to Binominal** so that data can be use for **FP-Growth** operator



In **FP-Growth** operator I set the **Minimum Support** to 0.43

And for the **Create Association Rule** Operator I set the **Minimum confidence** to 0.97



## Outlier Detection

A diagram of a diagram

Description automatically generated

### LOF

The **Local Outlier Factor (LOF)** method is an algorithm used for identifying outlier data points by measuring the local deviation of a given data point with respect to its neighbors. It is particularly useful in datasets where the density varies across observations. LOF is a type of unsupervised learning algorithm, meaning it does not require labeled data to function. Instead, it relies on the concept of local density, with outliers being identified as points that have a substantially lower density than their neighbors.

This process is performed by setting the parameters minimal points lower bound to 10 , minimal points upper bound to 20 and taking the distance function as Euclidian distance.

A screenshot of a computer

Description automatically generated

**Generating an outlier flag** after using the Local Outlier Factor (LOF) method is a crucial step in data preprocessing and analysis, enabling the identification and segregation of anomalies from normal observations within a dataset. By applying a threshold to the LOF scores—where scores indicate the degree of deviation of an observation from its neighbors—data points are classified as either outliers or inliers, with a binary flag assigned accordingly.

In this we took the method as contamination method , score column as outlier and contamination threshold 0.05.

### ISF

A screenshot of a graph

Description automatically generated

The **Isolation Forest (ISF)** algorithm is an effective method for detecting outliers and anomalies in datasets. It is based on the principle that anomalies are data points that are few and different, and hence, they are susceptible to being isolated from the rest of the data. The algorithm isolates anomalies instead of modeling the normal points, which is a distinctive approach compared to many other anomaly detection techniques.

Here we have set the parameters number of trees 1000, max leaf size to 2 and bootstrap ratio is 0.9 and score calculation average\_path.

**Then Filter Example** is used to display the only those observations flagged as outliers by both methods.

A screenshot of a computer

Description automatically generated

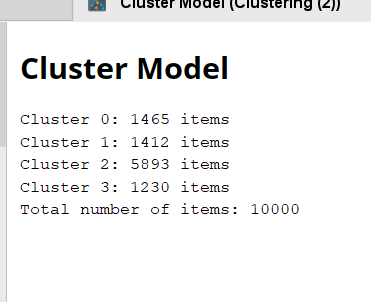
# Discussion of results

Classification:

The classification models, including decision trees, random forests, and k-nearest neighbors (kNN), likely provided valuable insights into the relationships between different attributes and the likelihood of collisions resulting in various levels of injury severity or property damage. For instance, certain combinations of environmental conditions, light, and road surface conditions might have emerged as significant predictors of collision outcomes.

The insights from classification models and clustering analysis can guide targeted interventions and infrastructure improvements to reduce collision rates and enhance road safety in Ottawa.

## Clustering



"Cluster\_0" reflects traffic incidents primarily resulting in property damage without major injuries. Most incidents occurred at non-intersections during clear weather, although snow and slushy conditions are also noted. Accidents were evenly spread throughout the day with a considerable number occurring in the evening. Many locations lacked formal traffic control, which may have contributed to the incidents. The data is from January 2020, and injuries reported were minimal to minor. This cluster may have been formed based on similarities in accident conditions and outcomes.

"Cluster\_1" shows traffic incidents with a range of injury severities, including minor and major injuries, mostly occurring in clear weather during various times of the day. Both intersection and non-intersection incidents are present, with many areas lacking formal traffic control. The data spans January 2020.

"Cluster\_2" features traffic incidents mainly with unknown injuries, occurring across various times, predominantly in clear weather and daylight conditions. Accidents are spread between intersections with traffic control and non-intersections without control, suggesting a mix of potentially risky locations. Instances of minor injuries and minimal damage are noted, with a few outliers identified, which could indicate exceptional cases within this cluster. Data is again from January 2020.

"Cluster\_3" involves traffic incidents largely with unknown injury severity, occurring under various conditions, with a significant number in clear weather and a mix of daylight, dusk, and dark. Accidents are split between controlled intersections and areas with no control, including many near private drives, suggesting different risk points. Snow and wet conditions are frequent, pointing to potential weather-related driving hazards. Data is from January 2020, with incidents spanning across all times of the day.

### Association

Rule 1 states that if an accident is classified as property damage only (P.D. only) and the road surface condition is dry, then the environmental condition is clear with a confidence of 99.6%.

Rule 2 indicates that if the maximum injury from an accident is unknown and the road surface condition is dry, then the environmental condition is clear with a confidence of 99.6%.

Rule 3 specifies that if an accident is classified as property damage only (P.D. only), the road surface condition is dry, and the light condition is daylight, then the environmental condition is clear with a confidence of 99.9%.

Rule 4 establishes that if an accident occurs at a non-intersection location and is classified as property damage only (P.D. only), then there is no traffic control at the accident site with a confidence of 100%.

Rule 5 states that if an accident occurs at a non-intersection location, the maximum injury is unknown, and the accident occurs in the year 2020, then there is no traffic control at the accident site with a confidence of 100%.

### Outlier Detection

The combination of LOF and ISF methods yield insights into data patterns, identifying outliers and anomalous behaviors within the transportation collision dataset. By leveraging these techniques, analysts can improve data quality, enhance predictive modeling accuracy, and identify potential areas for intervention to enhance road safety measures.

The results of this analysis could be interpreted to suggest that while the majority of the data points fit within expected parameters, there are specific observations that deviate in a manner significant enough to warrant further investigation. These outliers could represent errors, unusual but valid scenarios, or indicate underlying patterns or trends that could be of interest in the context of the analysis, such as potential injury severities in a healthcare dataset or anomalies in equipment performance data.

A pie chart with different colored circles

Description automatically generated

This chart shows that most of the outliers are detected with the collisions occurring in the **Winter** season followed by Spring .

# Conclusion

From the analysis it is clear that most collisions in Ottawa occur –

* At Intersection
* At Wet Road Surface
* In Winter Season

Though there are some outliers in

* Winter has the most outliers: This suggests a possible increase in anomalies during colder months.

* Fall is second in outlier count: Indicating a lesser but still significant presence of anomalies.
* Summer and Spring have fewer outliers: With Spring having the least, implying that warmer months have fewer anomalies.

# References

|  |  |
| --- | --- |
| [1] | el\_chief, "RapidMiner Community," RapidMiner, November 2010. [Online]. Available: https://community.rapidminer.com/discussion/11263/apriori-algorithm-in-rapidminer. |