



# **Factors determine injury classification in Chicago car crashes**

-- Final Project Presentation



**Presented By : Pinqi Wang , Onur Önel**

**Date: 8/11/2023**



# Introduction

The datasets we are working with is sourced from the City of Chicago's Data Portal, specifically the "Traffic Crashes - Crashes" and "Traffic Crashes - People" dataset. These dataset provides detailed information about traffic crashes occurring within the city of Chicago. The data is collected by reporting officers who respond to crash scenes and is used by the Chicago Police Department and the Department of Transportation.

Each row in the dataset "Traffic Crashes - Crashes" represents a unique traffic crash and includes a variety of information about the crash, such as the date and time of the crash, the posted speed limit, the type of traffic control device present, the weather and lighting conditions, the first type of crash (e.g., rear end, angle, pedestrian), and the type of trafficway (e.g., one-way, two-way, divided).

Each row in the dataset "Traffic Crashes - People" represents a person involved in the crash and includes a variety of information about the person such as sex, age, people type(driver, passenger etc) , injury classification etc .




# Business Understanding

The main objective is to understand the **factors that contribute to the severity of injuries in traffic crashes** in the city of Chicago.

This could involve identifying which conditions or characteristics of crashes are associated with different types of injuries. The goal is to inform interventions that reduce the severity of injuries in crashes.

To achieve this, we will use classification techniques (k-Nearest Neighbors and Decision Trees) to predict the severity of injuries based on crash characteristics, and clustering (k-Means) to identify patterns in the data. We will also use an association rule model to see the underlying associations.



# Data Preparation

**Data Gathering and Cleaning:** Starts by connecting to the City of Chicago's Socrata API and downloading two datasets: a "Crash" dataset and a "People" dataset. Each dataset is imported as a Pandas DataFrame. It then applies a filtering step to discard columns where over 40% of the data is missing. I also filters the 'People' dataset to only consider rows where the 'person\_type' is either 'DRIVER' or 'PASSENGER'.

**Data Merging and Column Cleaning:** I then merges the two datasets (Crashes and People) based on the 'crash\_record\_id' field, which presumably is a common identifier in both datasets. After the datasets are merged, it performs a series of column removals, dropping columns that are irrelevant or repetitive, and also dropping categorical variables that have too many distinct values. All these steps help to make the data cleaner, more manageable, and more appropriate for further analysis.


We end up with **18632 records** with **37 attributes**.



# Data Preparation

**Crash Dataset:** crash\_record\_id, crash\_date\_x, posted\_speed\_limit, traffic\_control\_device, device\_condition, weather\_condition, lighting\_condition, first\_crash\_type, trafficway\_type, alignment, roadway\_surface\_cond, road\_defect, damage, prim\_contributory\_cause, num\_units, most\_severe\_injury, injuries\_total, injuries\_no\_indication, crash\_hour, crash\_day\_of\_week, crash\_month, latitude, longitude

**People Dataset:** person\_id, person\_type, sex, age, safety\_equipment, airbag\_deployed, ejection, injury\_classification, driver\_action, driver\_vision, physical\_condition, bac\_result, injuries\_bad\_fatal, injuries\_medium






# Data Preparation - Feature Engineering

Injury classification (5 to 3)  
Traffic Control Device (17 to 5)  
Device Condition ( 8 to 3 )  
Weather Condition( 12 to 5)  
First Crash Types (17 to 5)  
Trafficway Types (17 to 5)  
Road Defect ( 7 to 3)  
Primary Contributory Factor (38 to 5)  
Safety Equipment Usage (15 to 4)  
Driver Behavior (19 to 5)  
Driver Visibility (12 to 5)  
Physical Condition (12 to 5)

During the feature engineering phase of our predictive analysis, we undertook a systematic approach to refine and reorganize certain attributes in our dataset. By carefully recategorizing and relabeling the data, we aimed to extract valuable insights that would contribute to building a robust predictive model.



# Data Preparation - Missing Values

	missing_percentage
latitude	0.68
longitude	0.68
sex	3.12
airbag_deployed	0.01
ejection	0.01
injury_classification	0.04
driver_vision	32.02
physical_condition	32.02
bac_result	32.02
age	21.28


Interestingly, we noted that attributes like 'driver\_vision', 'physical\_condition', and 'bac\_result' were collected only for drivers and not passengers. This observation led us to a strategy for imputing the missing values. We decided to fill the missing values in these columns based on the corresponding values from records with the same 'crash\_record\_id' where 'person\_type' was identified as 'DRIVER'.





# Classification Techniques

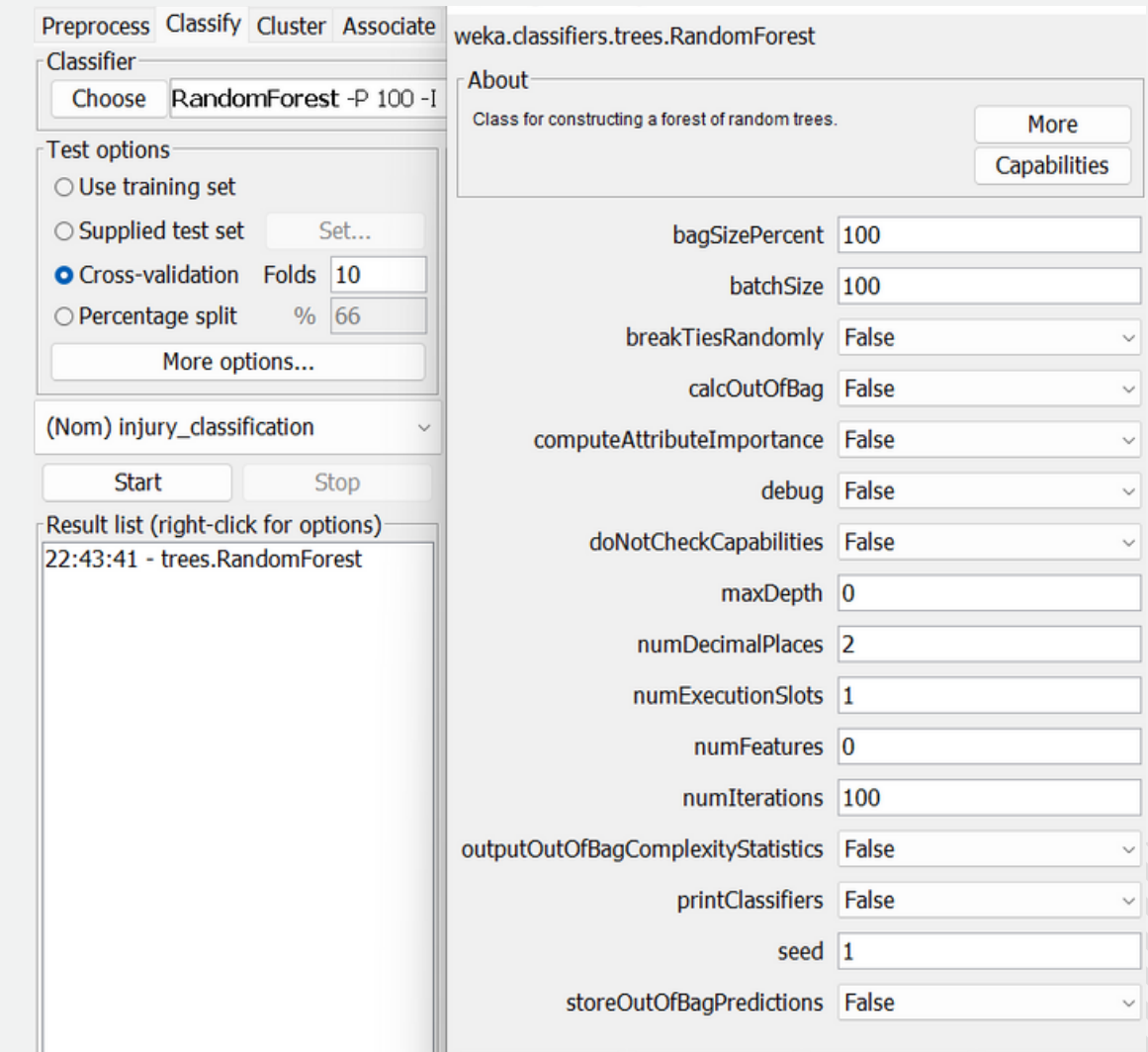
We employed classification techniques such as **k-Nearest Neighbors** and **Decision Trees** for predicting the severity of injuries. Further, we leveraged the **k-Means clustering algorithm** to identify inherent patterns in the data.





# Random Forest Our Principal Classification Model

Our main classification model is the **Random Forest algorithm**, the model was validated using 10-fold cross-validation to ensure reliability and stability of our predictions.



The screenshot displays the Weka software interface for configuring the Random Forest classifier. The 'Classify' tab is active, and the 'RandomForest -P 100 -I' classifier is selected. The 'Test options' section shows 'Cross-validation' with 'Folds' set to 10. The 'Result list' shows a single entry: '22:43:41 - trees.RandomForest'. The right-hand pane, titled 'weka.classifiers.trees.RandomForest', contains an 'About' section and a list of parameters with their current values:

Parameter	Value
bagSizePercent	100
batchSize	100
breakTiesRandomly	False
calcOutOfBag	False
computeAttributeImportance	False
debug	False
doNotCheckCapabilities	False
maxDepth	0
numDecimalPlaces	2
numExecutionSlots	1
numFeatures	0
numIterations	100
outputOutOfBagComplexityStatistics	False
printClassifiers	False
seed	1
storeOutOfBagPredictions	False

# Random Forest Model Training & Testing

The data was divided into a **70:30** ratio for training and testing purposes  
The model demonstrated exceptional performance,  
with an **accuracy of approximately 96%** in the training phase.

```
Size of the tree : 690

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      3815          95.9748 %
Incorrectly Classified Instances    160           4.0252 %
Kappa statistic                    0.7879
Mean absolute error                 0.0268
Root mean squared error             0.162
Relative absolute error             20.8252 %
Root relative squared error         63.9484 %
Total Number of Instances          3975

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.797	0.021	0.824	0.797	0.810	0.788	0.894	0.690	Bad Injury/Fatal
	0.979	0.203	0.976	0.979	0.977	0.788	0.894	0.975	Medium Injury
	?	0.000	?	?	?	?	?	?	No Injury
Weighted Avg.	0.960	0.183	0.959	0.960	0.959	0.788	0.894	0.944	

```
=== Confusion Matrix ===

  a    b    c  <-- classified as
342  87    0 |   a = Bad Injury/Fatal
 73 3473    0 |   b = Medium Injury
  0    0    0 |   c = No Injury
```

# Random Forest Model

## Training & Testing

On the test set, the model maintained a high accuracy, classifying **around 95%** of the instances correctly. However, a minor reduction in the **Kappa statistic** indicates a **slight decrease** in the agreement between the predicted and observed categorizations.

```
Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.09 seconds

=== Summary ===

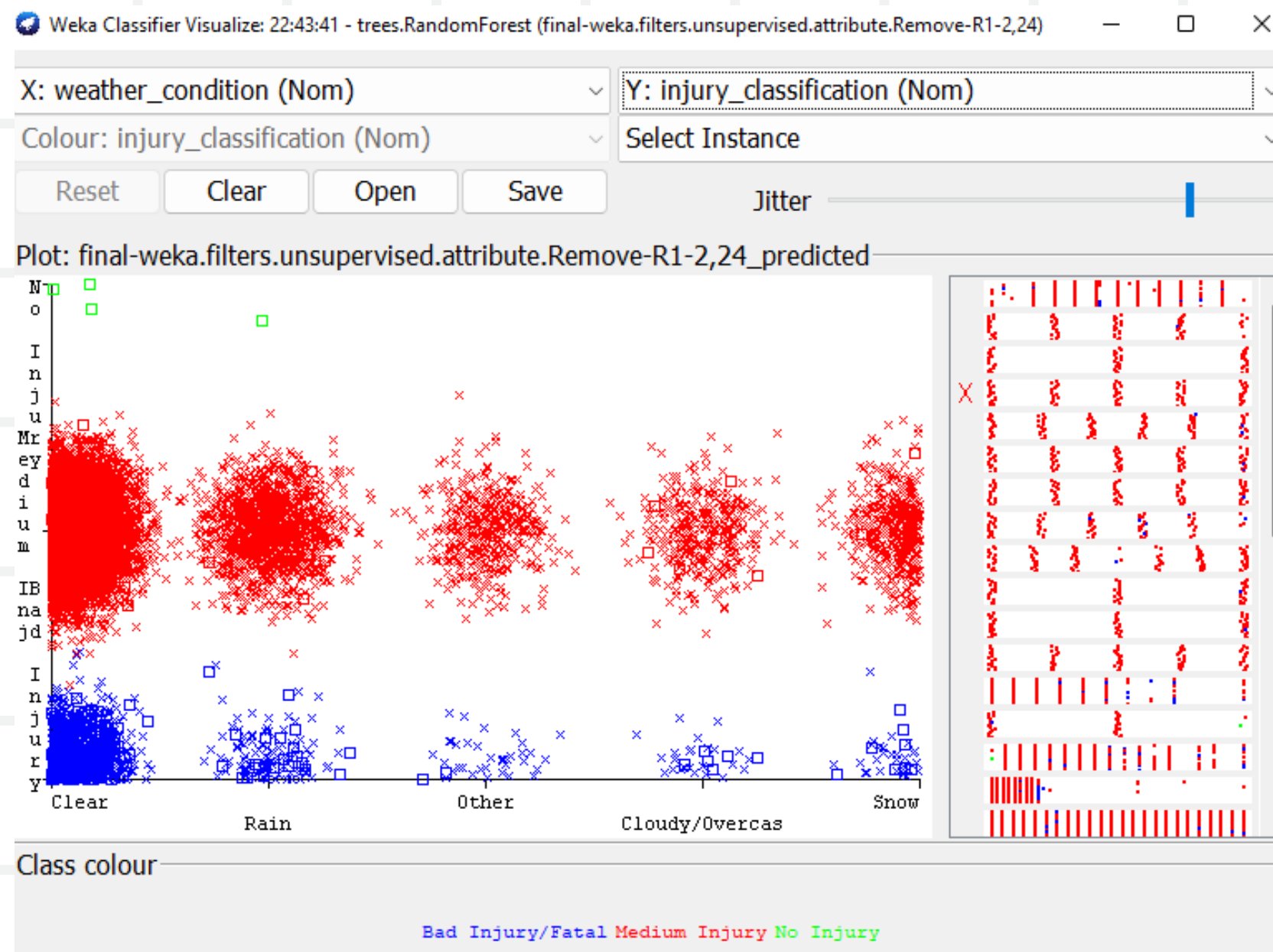
Correctly Classified Instances      8791      94.7817 %
Incorrectly Classified Instances    484      5.2183 %
Kappa statistic                    0.7048
Mean absolute error                0.0354
Root mean squared error            0.1836
Relative absolute error             28.7973 %
Root relative squared error         75.8156 %
Total Number of Instances          9275

=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.742	0.030	0.728	0.742	0.735	0.706	0.867	0.585	Bad Injury/Fatal
	0.970	0.261	0.972	0.970	0.971	0.705	0.865	0.972	Medium Injury
	0.000	0.000	?	0.000	?	?	0.500	0.000	No Injury
Weighted Avg.	0.948	0.238	?	0.948	?	?	0.865	0.934	

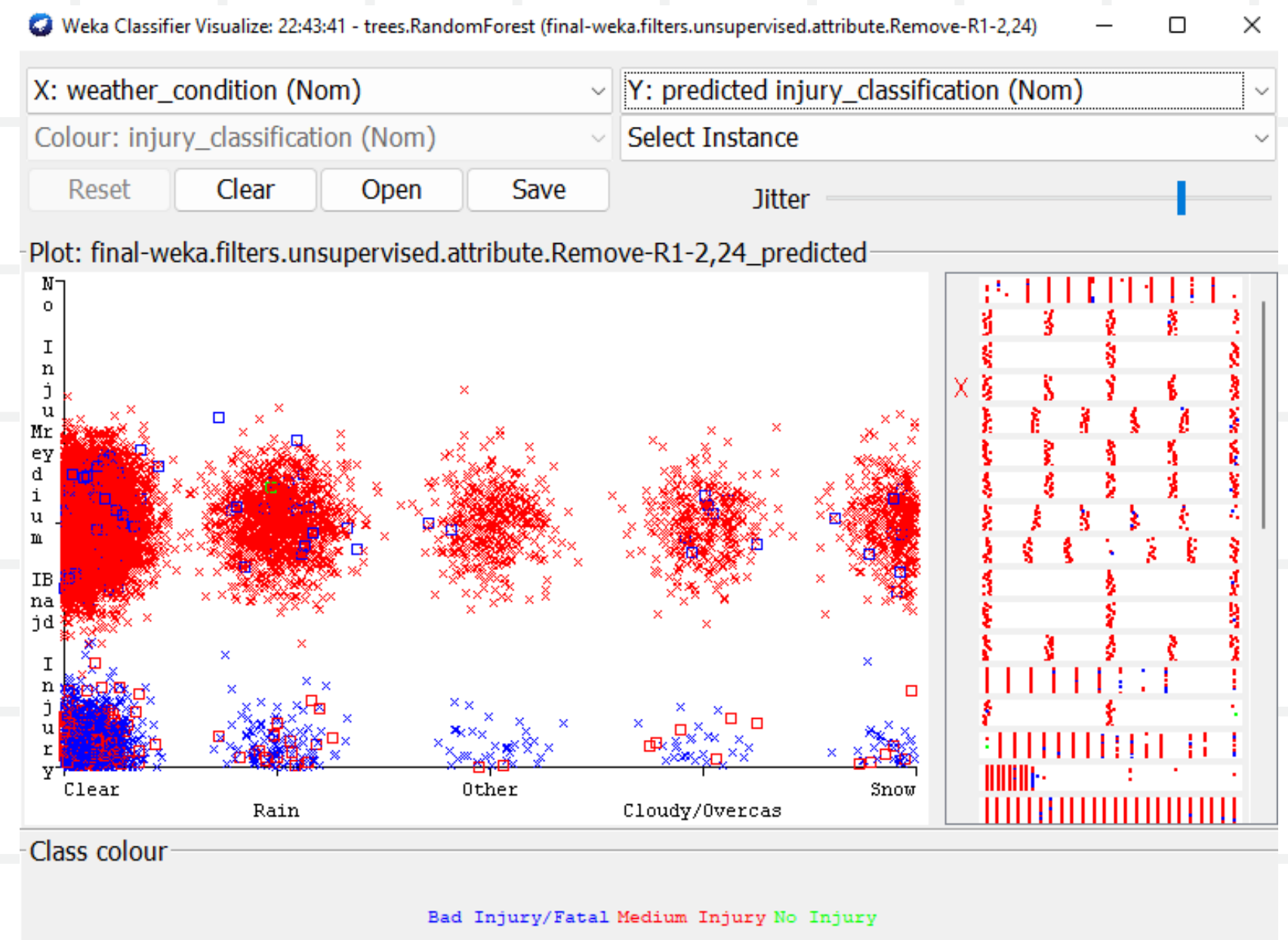
```
=== Confusion Matrix ===

 a   b   c  <-- classified as
666 232  0 |  a = Bad Injury/Fatal
248 8125  0 |  b = Medium Injury
 1    3  0 |  c = No Injury
```



Empty Boxes (Squares) = misclassifies

X's correctly classifies



Empty Boxes (Squares) = misclassifies

X's correctly classifies


I've created two separate scatter plots to illustrate the **correlation** between **weather** and **injury** classification.

This approach allows us to understand the differences between **actual** and **predicted classifications** more clearly and efficiently



# Cluster Analysis

The role of **cluster analysis** in revealing underlying patterns in the Chicago Traffic Crashes data. The importance of clustering for understanding **crash characteristics and severity levels**.





# K-Means Clustering with 3 Clusters

Attribute	Full Data (13250.0)	0 (3383.0)	1 (4416.0)	2 (5451.0)
posted_speed_limit	30.2364	30.2149	30.1515	30.3185
traffic_control_device	Signal Controls	Signal Controls	No Controls	Signal Controls
device_condition	FUNCTIONING PROPERLY	FUNCTIONING PROPERLY	NOT CONTROL	FUNCTIONING PROPERLY
weather_condition	Clear	Clear	Clear	Clear
lighting_condition	DAYLIGHT	DARKNESS, LIGHTED ROAD	DAYLIGHT	DAYLIGHT
first_crash_type	Angle/Rear End/Turning	Angle/Rear End/Turning	Angle/Rear End/Turning	Angle/Rear End/Turning
trafficway_type	Not Divided	Not Divided	Not Divided	Intersection
alignment	STRAIGHT AND LEVEL	STRAIGHT AND LEVEL	STRAIGHT AND LEVEL	STRAIGHT AND LEVEL
roadway_surface_cond	DRY	DRY	DRY	DRY
road_defect	NO DEFECTS	NO DEFECTS	NO DEFECTS	NO DEFECTS
damage	OVER \$1,500	OVER \$1,500	OVER \$1,500	OVER \$1,500
prim_contributory_cause	Traffic_Rule_Disregard	Unknown	Unknown	Traffic_Rule_Disregard
num_units	2.2097	2.2643	2.176	2.2033
most_severe_injury	Medium Injury	Medium Injury	Medium Injury	Medium Injury
injuries_total	2.1012	2.5279	1.8105	2.0719
injuries_no_indication	1.2069	1.1546	1.123	1.3073
crash_hour	12.8094	12.4597	13.1596	12.7426
crash_day_of_week	4.0409	3.5788	4.2418	4.1649
crash_month	6.3248	6.4286	6.6911	5.9635
latitude	41.8364	41.8374	41.8426	41.8308
longitude	-87.6693	-87.6751	-87.6793	-87.6576
person_type	DRIVER	DRIVER	DRIVER	DRIVER
sex	M	M	M	F
safety_equipment	Unknown	Unknown	Unknown	Used
airbag_deployed	DID NOT DEPLOY	DEPLOYED, COMBINATION	DID NOT DEPLOY	DID NOT DEPLOY
ejection	NONE	NONE	NONE	NONE
injury_classification	Medium Injury	Medium Injury	Medium Injury	Medium Injury
driver_action	No Action	Other	No Action	No Action
driver_vision	NOT OBSCURED	UNKNOWN	NOT OBSCURED	NOT OBSCURED
physical_condition	NORMAL	NORMAL	NORMAL	NORMAL
bac_result	TEST NOT OFFERED	TEST NOT OFFERED	TEST NOT OFFERED	TEST NOT OFFERED
age	37.2089	33.7827	39.2489	37.6826
injuries_bad_fatal	0.2248	0.3757	0.2043	0.1477
injuries_medium	1.8765	2.1522	1.6062	1.9242

Weka Explorer

PreprocessClassifyClusterAssociateSelect

Clusterer

ChooseSimpleKMeans-init 0 -max-c

Cluster mode

☒ Use training set

☐ Supplied test setSet...

☐ Percentage split%66

☐ Classes to clusters evaluation

(Num) injuries\_medium

☒ Store clusters for visualization

Ignore attributes

StartStop

Result list (right-click for options)

23:38:02 - SimpleKMeans

weka.gui.GenericObjectEditor

weka.clusterers.SimpleKMeans

doNotCheckCapabilitiesFalse

dontReplaceMissingValuesFalse

fastDistanceCalcFalse

initializationMethodRandom

maxIterations500

numClusters3

numExecutionSlots1


preserveInstancesOrderFalse

reduceNumberOfDistanceCalcsViaCanopiesFalse

seed10



## 3 Cluster Analysis Detailed

- Cluster 0 (26%): Crashes mainly at 30 mph, under functional 'Signal Controls'. The weather was **clear**, often during **lit darkness**. The majority of crashes were '**Angle/Rear End/Turning**' type with '**Medium**' injuries. The primary cause is unknown.
  - Cluster 1 (33%): Similar to Cluster 0, but with '**No Controls**' in place. Crashes mostly **during daylight on dry roads**. Injuries were typically '**Medium**' severity, with the cause being unknown.
  - Cluster 2 (41%): Most crashes at 30 mph under '**Signal Controls**', in **clear weather** conditions, either in **daylight or lit darkness**. The majority of crashes were '**Angle/Rear End/Turning**' type at intersections, with '**Medium**' severity injuries. The main cause was **disregard for traffic rules**.
- 



# K-Means Clustering with 5 Clusters

Attribute	Full Data (13250.0)	Cluster# 0 (2206.0)	1 (3151.0)	2 (1074.0)	3 (2322.0)	4 (4497.0)
posted_speed_limit	30.2364	30.1469	30.0816	30.2086	30.3867	30.3178
traffic_control_device	Signal Controls	Signal Controls	No Controls	Signal Controls	No Controls	Signal Controls
device_condition	FUNCTIONING PROPERLY	FUNCTIONING PROPERLY	NOT CONTROL	FUNCTIONING PROPERLY	NOT CONTROL	FUNCTIONING PROPERLY
weather_condition	Clear	Clear	Clear	Rain	Clear	Clear
lighting_condition	DAYLIGHT	DAYLIGHT	DAYLIGHT	DAYLIGHT	DAYLIGHT	DAYLIGHT
first_crash_type	Angle/Rear End/Turning	Angle/Rear End/Turning	Angle/Rear End/Turning	Angle/Rear End/Turning	Angle/Rear End/Turning	Angle/Rear End/Turning
trafficway_type	Not Divided	Intersection	Not Divided	Not Divided	Not Divided	Intersection
alignment	STRAIGHT AND LEVEL	STRAIGHT AND LEVEL	STRAIGHT AND LEVEL	STRAIGHT AND LEVEL	STRAIGHT AND LEVEL	STRAIGHT AND LEVEL
roadway_surface_cond	DRY	DRY	DRY	WET	DRY	DRY
road_defect	NO DEFECTS	NO DEFECTS	NO DEFECTS	UNKNOWN	NO DEFECTS	NO DEFECTS
damage	OVER \$1,500	OVER \$1,500	OVER \$1,500	OVER \$1,500	OVER \$1,500	OVER \$1,500
prim_contributory_cause	Traffic_Rule_Disregard	Unknown	Unknown	Traffic_Rule_Disregard	Unknown	Traffic_Rule_Disregard
num_units	2.2097	2.2684	2.1663	2.1462	2.2558	2.2028
most_severe_injury	Medium Injury	Medium Injury	Medium Injury	Medium Injury	Medium Injury	Medium Injury
injuries_total	2.1012	2.3164	1.6271	1.9786	2.7993	1.9967
injuries_no_indication	1.2069	1.1206	1.0292	1.4022	1.2407	1.3095
crash_hour	12.8094	12.0734	12.9702	14.0912	13.0866	12.6084
crash_day_of_week	4.0409	3.6296	4.1803	3.7551	3.7511	4.3629
crash_month	6.3248	6.2743	6.4783	5.4646	6.5556	6.3282
latitude	41.8364	41.8321	41.8446	41.8461	41.8378	41.8299
longitude	-87.6693	-87.671	-87.6807	-87.6746	-87.6775	-87.655
person_type	DRIVER	DRIVER	DRIVER	DRIVER	PASSENGER	DRIVER
sex	M	M	M	F	F	F
safety_equipment	Unknown	Unknown	Unknown	Unknown	Used	Used
airbag_deployed	DID NOT DEPLOY	DEPLOYED, COMBINATION	DID NOT DEPLOY	NOT APPLICABLE	DID NOT DEPLOY	DID NOT DEPLOY
ejection	NONE	NONE	NONE	NONE	NONE	NONE
injury_classification	Medium Injury	Medium Injury	Medium Injury	Medium Injury	Medium Injury	Medium Injury
driver_action	No Action	No Action	No Action	No Action	Other	No Action
driver_vision	NOT OBSCURED	UNKNOWN	UNKNOWN	UNKNOWN	NOT OBSCURED	NOT OBSCURED
physical_condition	NORMAL	UNKNOWN	NORMAL	NORMAL	NORMAL	NORMAL
bac_result	TEST NOT OFFERED	TEST NOT OFFERED	TEST NOT OFFERED	TEST NOT OFFERED	TEST NOT OFFERED	TEST NOT OFFERED
age	37.2089	36.2756	39.9946	36.1508	33.1219	38.0778
injuries_bad_fatal	0.2248	0.2302	0.2307	0.1443	0.2705	0.1394
injuries_medium	1.8765	1.9352	1.3964	1.8343	2.5289	1.8572

Weka Explorer

PreprocessClassifyClusterAssociateSelect

Clusterer

ChooseSimpleKMeans-init 0 -max-c

Cluster mode

☒ Use training set

☐ Supplied test setSet...

☐ Percentage split% 66

☐ Classes to clusters evaluation

(Num) injuries\_medium

☒ Store clusters for visualization

Ignore attributes

StartStop

Result list (right-click for options)

23:38:02 - SimpleKMeans

weka.gui.GenericObjectEditor

weka.clusterers.SimpleKMeans

doNotCheckCapabilitiesFalse

dontReplaceMissingValuesFalse

fastDistanceCalcFalse

initializationMethodRandom

maxIterations500

numClusters5

numExecutionSlots1


preserveInstancesOrderFalse

reduceNumberOfDistanceCalcsViaCanopiesFalse

seed10



## 5 Cluster Analysis Detailed

- **Cluster 0 (17%):** Accidents primarily in intersections under '**Signal Controls**', **medium injuries**, cause unknown.
  - **Cluster 1 (24%):** Similar to Cluster 0, but on '**Not Divided**' roads with '**No Controls**', cause unknown.
  - **Cluster 2 (8%):** Accidents on '**Not Divided**' roads under '**Signal Controls**' in rain, caused by **traffic rule disregard**.
  - **Cluster 3 (18%):** Similar to Cluster 2 but with '**No Controls**', **daylight** accidents **involving passengers**, cause unknown.
  - **Cluster 4 (34%):** Most frequent cluster; accidents at intersections under '**Signal Controls**' on **clear** days, caused by **traffic rule disregard**.
- 



# K-Means Clustering Sum of Squared Errors Analysis

Our K-Means clustering model was analyzed using the  
The sum of Squared Errors (SSE) for a number of clusters from 1 to 10.  
Here's a brief overview of our findings:

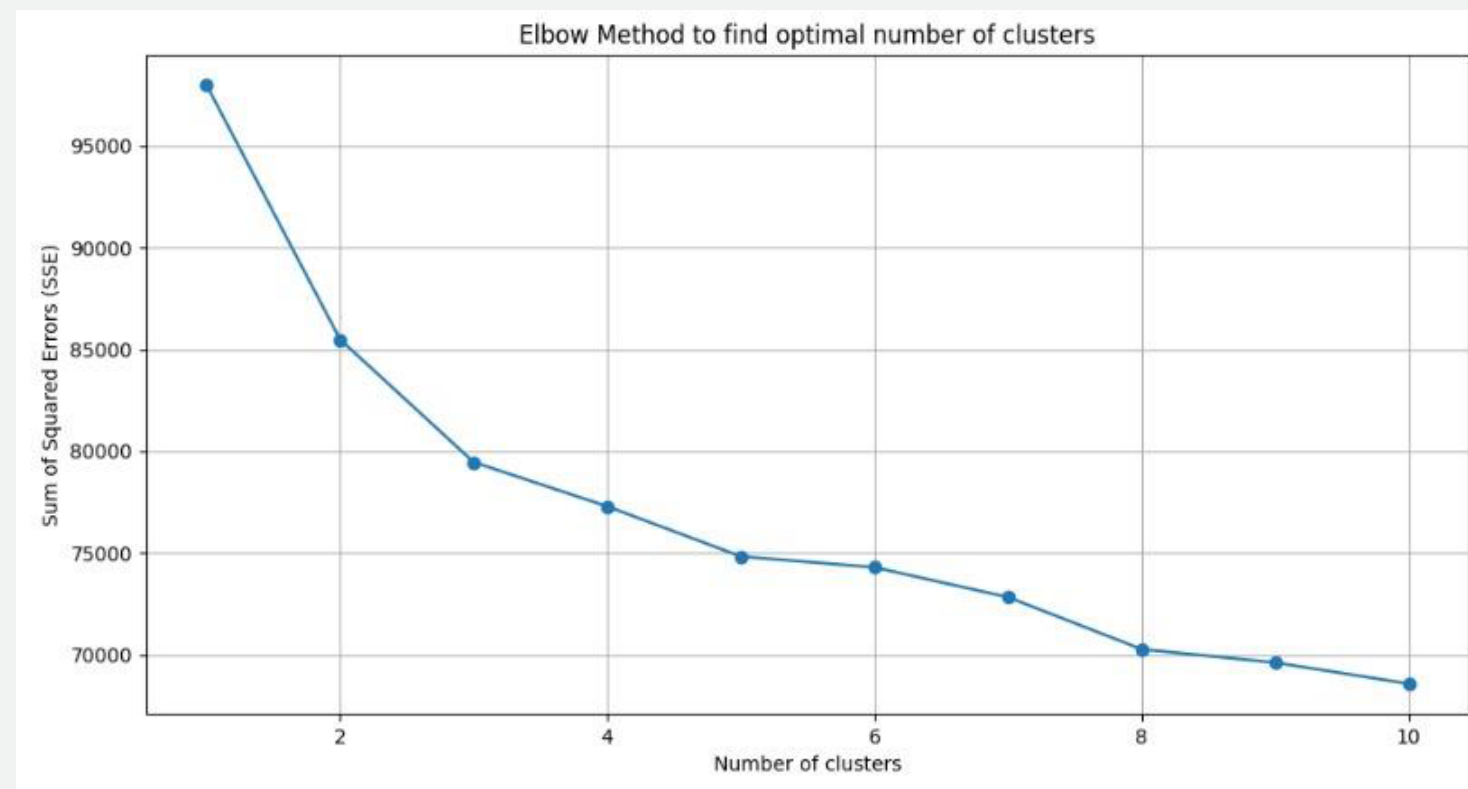
For 1 cluster, SSE = 98006.652  
For 2 clusters, SSE = 85494.19  
For 3 clusters, SSE = 79460.17  
For 4 clusters, SSE = 77287.31  
For 5 clusters, SSE = 74825.31

For 6 clusters, SSE = 74290.27  
For 7 clusters, SSE = 72821.24  
For 8 clusters, SSE = 70264.58  
For 9 clusters, SSE = 69607.73  
For 10 clusters, SSE = 68567.00



# Elbow Graph

We can see that the reduction in SSE becomes less pronounced after **4 clusters**. Between 4 and 5 clusters, the reduction in SSE is **smaller** compared to the decrease between 3 and 4 clusters. Between 5 and 6 clusters, the SSE reduction is even less and continues to slow down thereafter






# Outlier Detection

Outlier detection plays a critical role in our data analysis process. For this, we deployed two models: the Isolation Forest Model (ISF) and the Local Outlier Factor (LOF). Our first step was to use the visualizer in Weka to get an initial overview of potential outliers. With 38 attributes at hand, it was imperative to identify potential areas and the volume of outliers before running our models.

We excluded non-essential variables such as 'crash\_record\_id', 'person\_id', 'crash\_date\_x', 'crash\_type', 'alignment', 'device\_condition', 'crash\_month', 'latitude', and 'longitude'. Following this, we applied one-hot encoding to the remaining categorical attributes.




# Insights from the Isolation Forest Model




we utilized a model-agnostic method for assessing feature importance, such as permutation importance or SHAP





# Ensemble Outcome: 9 Outliers Identified

1. Outlier 1 & 2: For these two incidents, the number of units involved was recorded as eight. However, each crash only involved two people who both tragically suffered severe or fatal injuries. Interestingly, these incidents occurred in zones with a relatively low-speed limit of just 20 miles/hr.
  2. Outlier 3: The third outlier featured a crash involving six units but only one individual, a young person, who was mildly injured. This occurred in a 30 mile/hour speed limit zone.
  3. Outliers 4, 5 & 6: These three incidents each involved 14 individuals who were injured, with two people suffering severe or fatal injuries in each case. Remarkably, these incidents involved only two units.
  4. Outlier 7: This outlier involved a driver in poor physical condition who hit a pedestrian, resulting in medium-level injuries. The incident took place at 1 am.
  5. Outlier 8: An incident involving a three-year-old girl who, fortunately, was not injured when the car she was in collided with an object. Despite the obscured vision of the driver and the deployment of the airbag, the child remained unharmed.
  6. Outlier 9: This outlier featured a female driver who, regrettably, disregarded safety precautions and did not have any safety equipment in place. This resulted in severe or fatal injuries for the driver.
- 





# Association Model

For this model specifically, in order to obtain any rules that contains our interest (e.g. injuries\_classification\_Medium) , we pull the most recent **200000** records from people dataset and **filtered out the ones with no injury** and then join the crash dataset to people dataset using crash\_record\_id. I then redid the same feature engineering and missing value handling procedures. The data records available here for association rule is 18737.

for the apriori algorithm we narrowed down the attributes selected for the model to 'posted\_speed\_limit', 'traffic\_control\_device', 'lighting\_condition', 'age', 'sex', 'num\_units', 'crash\_hour', 'safety\_equipment', 'prim\_contributory\_cause', 'first\_crash\_type', 'airbag\_deployed', 'injury\_classification'. And performed data binning and one-hot encoding for attributes needed.

We performed association rule modeling. The **minimum support threshold** is set to 0.2 and **minimum threshold for confidence** is set to 0.7.




# Top 10 Association Rules for Records with Medium Injury Incurred

antecedents	consequents	antecedent s	consequent s	support	confidence	lift
frozenset({'crash_hour_Afternoon', 'posted_speed_limit_21-30 mps', 'num_units_Two Units'})	frozenset({'first_crash_type_Angle/Rear End/Turning', 'injury_classification_Medium Injury', 'lighting_condition_DAYLIGHT'})	0.26781401	0.47977053	0.20410628	0.7621195	1.5885083
frozenset({'crash_hour_Afternoon', 'num_units_Two Units'})	frozenset({'first_crash_type_Angle/Rear End/Turning', 'injury_classification_Medium Injury', 'lighting_condition_DAYLIGHT'})	0.30827295	0.47977053	0.23460145	0.76101861	1.5862137
frozenset({'crash_hour_Afternoon', 'first_crash_type_Angle/Rear End/Turning'})	frozenset({'injury_classification_Medium Injury', 'lighting_condition_DAYLIGHT', 'num_units_Two Units'})	0.31310386	0.47524155	0.23460145	0.74927676	1.5766230
frozenset({'crash_hour_Afternoon', 'posted_speed_limit_21-30 mps', 'first_crash_type_Angle/Rear End/Turning'})	frozenset({'injury_classification_Medium Injury', 'lighting_condition_DAYLIGHT', 'num_units_Two Units'})	0.27596618	0.47524155	0.20410628	0.73960613	1.5562741
frozenset({'crash_hour_Afternoon', 'first_crash_type_Angle/Rear End/Turning'})	frozenset({'posted_speed_limit_21-30 mps', 'injury_classification_Medium Injury', 'lighting_condition_DAYLIGHT'})	0.31310386	0.51509662	0.24728261	0.78977821	1.5332622
frozenset({'crash_hour_Afternoon', 'first_crash_type_Angle/Rear End/Turning', 'num_units_Two Units'})	frozenset({'posted_speed_limit_21-30 mps', 'injury_classification_Medium Injury', 'lighting_condition_DAYLIGHT'})	0.26086957	0.51509662	0.20410628	0.78240741	1.5189527
frozenset({'crash_hour_Afternoon', 'posted_speed_limit_21-30 mps'})	frozenset({'first_crash_type_Angle/Rear End/Turning', 'injury_classification_Medium Injury', 'lighting_condition_DAYLIGHT'})	0.34269324	0.47977053	0.24728261	0.7215859	1.5040229
frozenset({'crash_hour_Afternoon', 'num_units_Two Units'})	frozenset({'posted_speed_limit_21-30 mps', 'injury_classification_Medium Injury', 'lighting_condition_DAYLIGHT'})	0.30827295	0.51509662	0.23762077	0.77081293	1.4964433
frozenset({'crash_hour_Afternoon', 'first_crash_type_Angle/Rear End/Turning', 'num_units_Two Units'})	frozenset({'injury_classification_Medium Injury', 'lighting_condition_DAYLIGHT'})	0.26086957	0.6044686	0.23460145	0.89930556	1.4877622
frozenset({'crash_hour_Afternoon'})	frozenset({'posted_speed_limit_21-30 mps', 'injury_classification_Medium Injury', 'lighting_condition_DAYLIGHT'})	0.39613527	0.51509662	0.30283816	0.76448171	1.4841520
frozenset({'crash_hour_Afternoon', 'first_crash_type_Angle/Rear End/Turning'})	frozenset({'injury_classification_Medium Injury', 'lighting_condition_DAYLIGHT'})	0.31310386	0.6044686	0.2807971	0.89681774	1.4836465
frozenset({'crash_hour_Afternoon', 'posted_speed_limit_21-30 mps', 'first_crash_type_Angle/Rear End/Turning', 'num_units_Two Units'})	frozenset({'injury_classification_Medium Injury', 'lighting_condition_DAYLIGHT'})	0.227657	0.6044686	0.20410628	0.89655172	1.4832064
frozenset({'crash_hour_Afternoon', 'posted_speed_limit_21-30 mps', 'first_crash_type_Angle/Rear End/Turning'})	frozenset({'injury_classification_Medium Injury', 'lighting_condition_DAYLIGHT'})	0.27596618	0.6044686	0.24728261	0.89606127	1.4823950
frozenset({'crash_hour_Afternoon'})	frozenset({'injury_classification_Medium Injury', 'first_crash_type_Angle/Rear End/Turning', 'lighting_condition_DAYLIGHT'})	0.39613527	0.47977053	0.2807971	0.70884146	1.4774593
frozenset({'crash_hour_Afternoon', 'num_units_Two Units'})	frozenset({'injury_classification_Medium Injury', 'lighting_condition_DAYLIGHT'})	0.30827295	0.6044686	0.27445652	0.89030362	1.4728699
frozenset({'crash_hour_Afternoon', 'posted_speed_limit_21-30 mps', 'num_units_Two Units'})	frozenset({'injury_classification_Medium Injury', 'lighting_condition_DAYLIGHT'})	0.26781401	0.6044686	0.23762077	0.88726043	1.4678354
frozenset({'crash_hour_Afternoon', 'posted_speed_limit_21-30 mps'})	frozenset({'injury_classification_Medium Injury', 'lighting_condition_DAYLIGHT'})	0.34269324	0.6044686	0.30283816	0.88370044	1.4619459
frozenset({'traffic_control_device_Signal Controls', 'crash_hour_Afternoon'})	frozenset({'injury_classification_Medium Injury', 'lighting_condition_DAYLIGHT'})	0.23309179	0.6044686	0.20591787	0.88341969	1.4614815
frozenset({'crash_hour_Afternoon'})	frozenset({'injury_classification_Medium Injury', 'lighting_condition_DAYLIGHT'})	0.39613527	0.6044686	0.34903382	0.88109756	1.4576399
frozenset({'traffic_control_device_Signal Controls', 'prim_contributory_cause_Traffic_Rule_Disregard'})	frozenset({'posted_speed_limit_21-30 mps', 'injury_classification_Medium Injury', 'first_crash_type_Angle/Rear End/Turning'})	0.31914251	0.57971014	0.24335749	0.76253548	1.315373
frozenset({'traffic_control_device_Signal Controls', 'prim_contributory_cause_Traffic_Rule_Disregard', 'num_units_Two Units'})	frozenset({'posted_speed_limit_21-30 mps', 'injury_classification_Medium Injury', 'first_crash_type_Angle/Rear End/Turning'})	0.26871981	0.57971014	0.20471014	0.76179775	1.3141011
frozenset({'traffic_control_device_Signal Controls', 'posted_speed_limit_21-30 mps', 'prim_contributory_cause_Traffic_Rule_Disregard', 'num_units_Two Units'})	frozenset({'injury_classification_Medium Injury', 'first_crash_type_Angle/Rear End/Turning'})	0.23671498	0.67089372	0.20471014	0.86479592	1.2890207
frozenset({'traffic_control_device_Signal Controls', 'prim_contributory_cause_Traffic_Rule_Disregard'})	frozenset({'injury_classification_Medium Injury', 'first_crash_type_Angle/Rear End/Turning', 'num_units_Two Units'})	0.31914251	0.56400966	0.23158213	0.7256386	1.2865712
frozenset({'traffic_control_device_Signal Controls', 'prim_contributory_cause_Traffic_Rule_Disregard', 'num_units_Two Units'})	frozenset({'injury_classification_Medium Injury', 'first_crash_type_Angle/Rear End/Turning'})	0.26871981	0.67089372	0.23158213	0.86179775	1.2845518
frozenset({'traffic_control_device_Signal Controls', 'posted_speed_limit_21-30 mps', 'prim_contributory_cause_Traffic_Rule_Disregard'})	frozenset({'injury_classification_Medium Injury', 'first_crash_type_Angle/Rear End/Turning', 'num_units_Two Units'})	0.28442029	0.56400966	0.20471014	0.71974522	1.2761221
frozenset({'traffic_control_device_Signal Controls', 'posted_speed_limit_21-30 mps', 'prim_contributory_cause_Traffic_Rule_Disregard'})	frozenset({'injury_classification_Medium Injury', 'first_crash_type_Angle/Rear End/Turning'})	0.28442029	0.67089372	0.24335749	0.85562633	1.2753530
frozenset({'traffic_control_device_Signal Controls', 'lighting_condition_DAYLIGHT', 'num_units_Two Units'})	frozenset({'posted_speed_limit_21-30 mps', 'injury_classification_Medium Injury', 'first_crash_type_Angle/Rear End/Turning'})	0.31853865	0.57971014	0.23490338	0.73744076	1.2720853
frozenset({'traffic_control_device_Signal Controls', 'prim_contributory_cause_Traffic_Rule_Disregard'})	frozenset({'injury_classification_Medium Injury', 'first_crash_type_Angle/Rear End/Turning'})	0.31914251	0.67089372	0.27204106	0.85241249	1.2705626
frozenset({'traffic_control_device_Signal Controls', 'sex_F'})	frozenset({'posted_speed_limit_21-30 mps', 'injury_classification_Medium Injury', 'first_crash_type_Angle/Rear End/Turning'})	0.27928744	0.57971014	0.20561594	0.73621622	1.2699729
frozenset({'traffic_control_device_Signal Controls', 'num_units_Two Units'})	frozenset({'posted_speed_limit_21-30 mps', 'injury_classification_Medium Injury', 'first_crash_type_Angle/Rear End/Turning'})	0.47463768	0.57971014	0.34812802	0.73346056	1.2652194



# Top 1 Association Rules Implication

The implication of this rule is that if a crash occurs in the afternoon, with a posted speed limit of 21-30 mps, and involves two units, there is a high likelihood (76.21% confidence) that the crash will be of type angle, rear end, or turning, occur in daylight, and result in a medium injury. This suggests that these factors (time of day, speed limit, number of units involved) may contribute to the severity of injuries in a crash. As you can see, this rule effectively summarizes the factors that are recurring in the first top 19 rules. Starting rule no. 20, we start to observe new factors such as `prim_contributory_cause = traffic rule disregard` and `sex=F`.



**THANK YOU**

