Q1 Intermediate Report

10/8/2024

Aaryan Sumesh and Sami Saleh

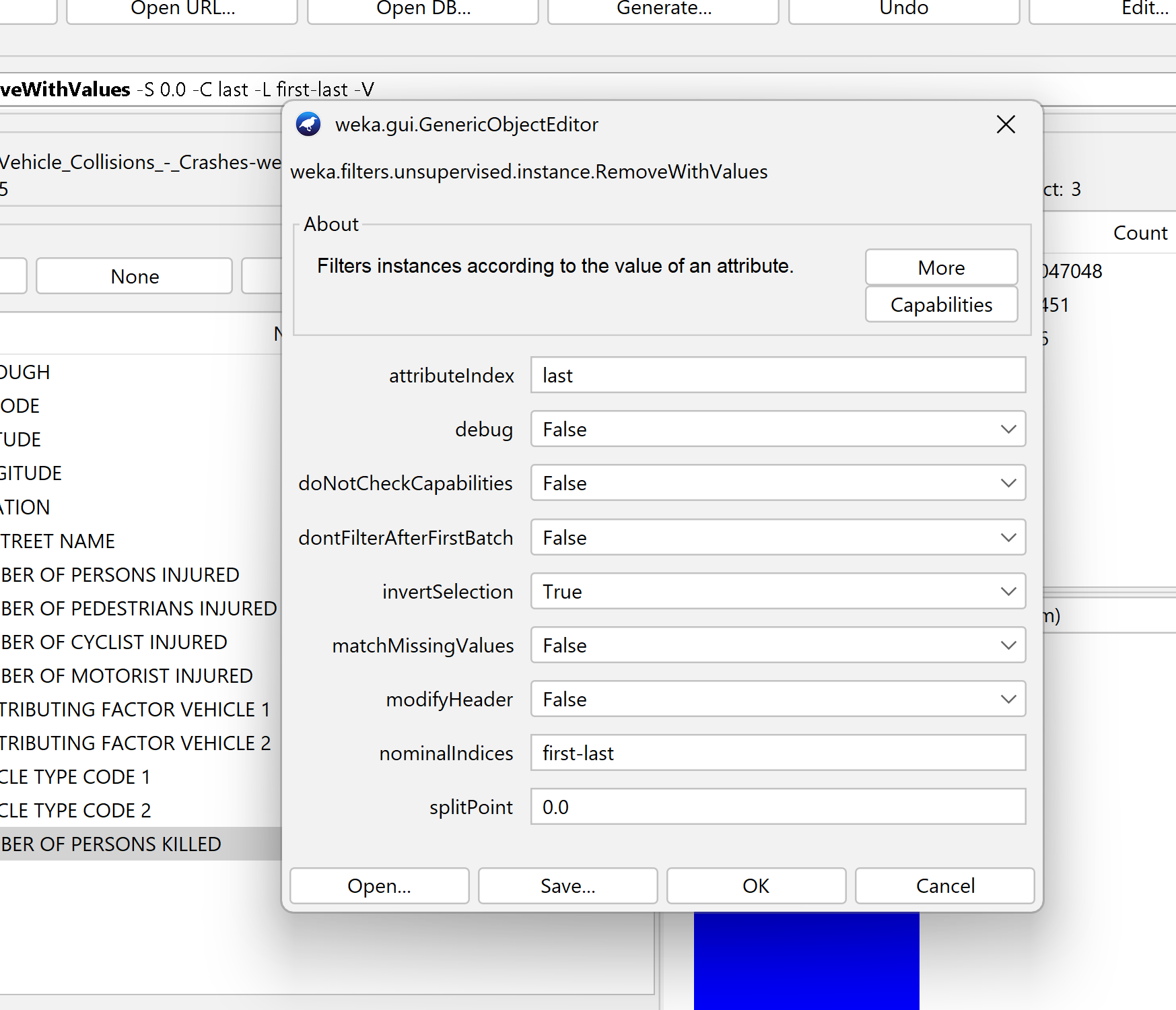
Project: Predicting and Classifying the Lethality of Car Crashes in New York Boroughs

**Project Goal:**

We will be predicting the severity of car crashes in New York Boroughs. We will group car crashes into three categories based on how lethal the crashes are. The three class labels will be NON-LETHAL, SOMEWHAT-LETHAL, and VERY-LETHAL.

**Missing Value Handling:**

To handle the missing values, we deleted all instances with a missing class value, which in this case is the NUMBER OF PERSONS KILLED class, removing a total of 30 instances. To do this, we go to choose > unsupervised > instance > RemoveWithValues. In the menu, we changed invertSelection to true, then clicked ok > apply. We will now use this attribute as our class label.



**Removing Unnecessary Columns and Attributes**

Attributes unique to a specific car crash such as COLLISION\_ID, CRASH DATE, and CRASH TIME were removed as they will not help predict the lethality of car crashes.

Bin into 4 seasons for date, and 3 for times

In addition, attribute columns having a majority of missing values were removed (CROSS STREET NAME, OFF STREET NAME, CONTRIBUTING FACTOR VEHICLE 3, CONTRIBUTING FACTOR VEHICLE 4, CONTRIBUTING FACTOR 5, VEHICLE TYPE CODE 3, VEHICLE TYPE CODE 4, VEHICLE TYPE CODE 5)

More than 2

Finally, we removed attribute columns that were redundant for the class attribute: NUMBER OF PEDESTRIANS KILLED, NUMBER OF CYCLISTS KILLED, NUMBER OF MOTORISTS KILLED. These attributes were redundant and don’t provide any valuable information when classifying crashes.

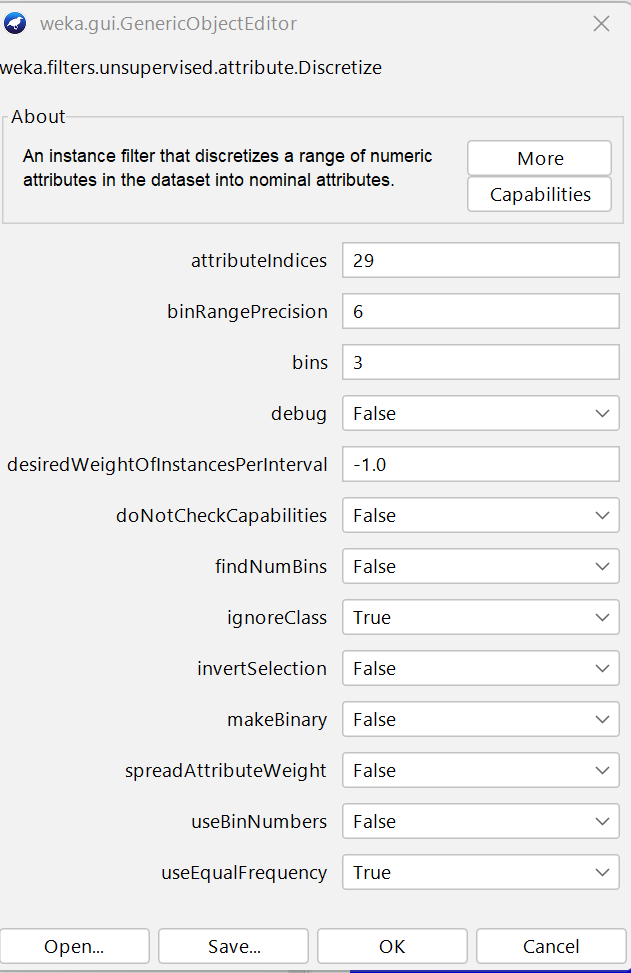
We filled in the remaining missing values using the WEKA ReplaceMissingValues function which replaces missing values with the mean of the other values.

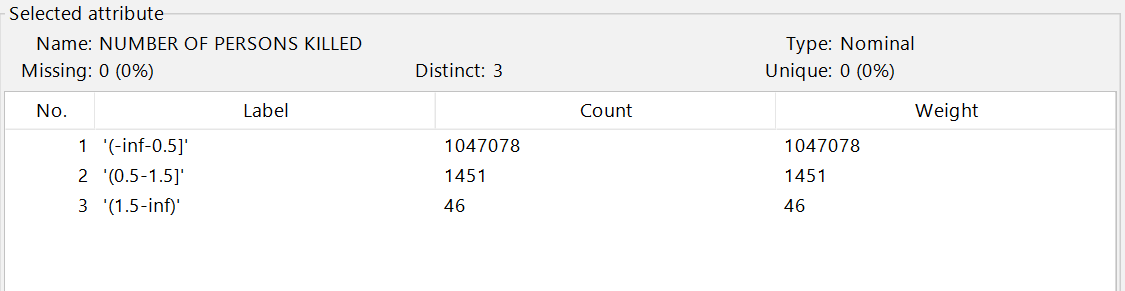
**Train-Test-Splitting:**

We used a 70-30 train-test split. To do this, we go to the weka filters in the preprocess tab. Then, we go to choose > instance > unsupervised > instance > remove percentage. Then, in the options menu we changed the percentage to 70. Then, we clicked ok > apply. We then saved this data as testing\_data, as this filter deleted 70% of the data, but also conserved the ratio of class labels. Then, we reopened the preprocessed data set and then did the same steps, but after we changed the percentage to 70, we selected “invert selection” > ok > apply. Then, we saved this dataset as traning\_data.

**Binning Class Attribute**

We use the WEKA Discretize function on our new class attribute NUMBER OF PERSONS KILLED to create three, equal frequency bins to categorize the lethality as seen below.



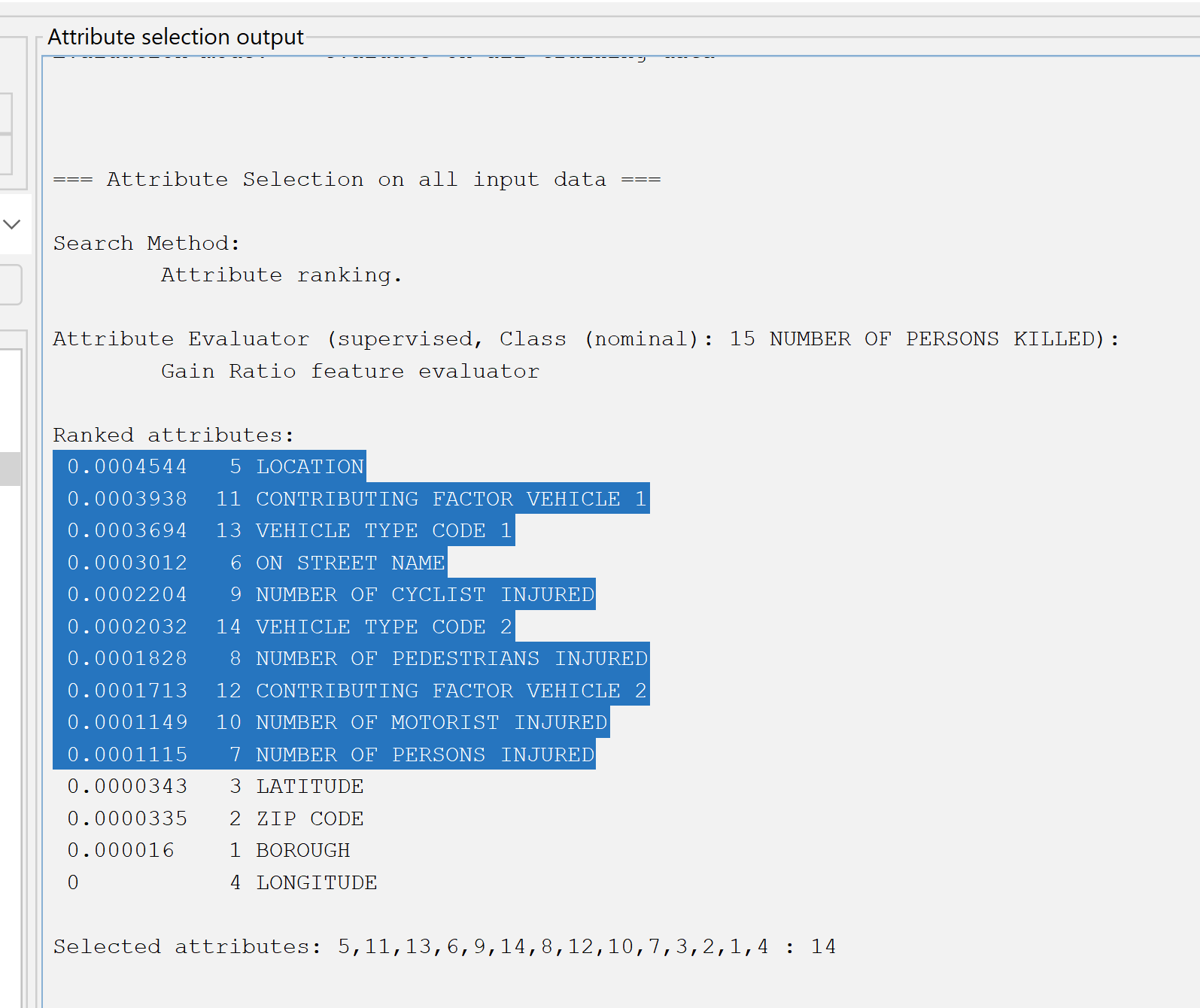


Note that the first bin must only contain 0.0 as a value, as we cannot have a negative or decimal number of persons killed. This bin will be our “non-lethal” category.

The second bin must only contain 1.0 as a value, as we cannot have a decimal number of persons killed. This bin will be our “slightly lethal” category.

The third bin contains integer values greater than 1.0. This bin will be our “highly lethal” category.

**Attribute Selection**:



**GainRatioAttributeEval**

We start attribute selection using WEKA GainRatioAttributeEval with a cutoff value for information gain of 0.0001. The selected values are highlighted in the image above.