

|  |
| --- |
|  |
| FINAL PROJECT |
| BI and DATA ANALYTICS |
| MOHIT NARGOTRA <040-817-484>| KRUSHANG PATEL <041-021-848> |

|  |  |
| --- | --- |
|  | Decorative |
| CONTENTS |
| INtRODUCTION AND DATAUNDERSTANDINGDATA Preparation  1. **DATA VISUALISATION AND CLASSIFICATION** 2. **MODELING AND**   **EVALUATION**  **4.1 KNN CLASSIFICATION**  **4.2 DECISION TREE**  **4.3 CLUSTERING**  **4.4 OUtlier DETECTION**   1. **RESULTS** 2. **CONCLUSION** |

### INtRODUCTION AND DATA UNDERSTANDING

We are using RIPA Police Stop Dataset. The San Diego Police Department has recorded information about all the stops they made while adhering to the guidelines outlined in Government Code section 12525.5. This code was established after the passing of the Racial and Identity Profiling Act of 2015 (AB 953). The department started collecting this data on July 1, 2018. However, data on stops made before this date is only available for vehicle stops and can be accessed through the police vehicle stops dataset. The data set contains the data from 2018 to 2022. However, we only used the data set for the year 2018. The whole data set contains three different datasets: basic details, result of stop, and action taken. They have been briefly described in the following passages.

Basic details: The dataset contains fundamental details about a police stop, such as when and where it occurred, how long it lasted, and some information about the individual stopped. Each row represents a person who was stopped and is assigned a pid, a unique identifier for that person, and a stop\_id, a unique identifier for the stop itself. Sometimes, one stop may involve multiple individuals, so a stop\_id can have several associated pids. Hence, we assumed that person with pid 1 was the driver, and we removed the other. This dataset has 29 columns and 653602 instances. A short description of all the attribute is given as below:

**Table 1 : Brief description of attributes in Basic Details dataset.**

|  |  |  |
| --- | --- | --- |
| **Field** | **Field\_type** | **Description** |
| **stop\_id** | int64 | unique identifer for stop |
| **ori** | str | agency originating identifier |
| **agency** | str | agency |
| **exp\_years** | int64 | officer years of experience in law enforcement |
| **stopdate** | datetime64[ns] | date stop occurred |
| **stoptime** | str | time stop began |
| **stopduration** | int64 | duration of time for stop in minutes |
| **stop\_in\_response\_to\_cfs** | int64 | was the stop made in response to a call for service? |
| **officer\_assignment\_key** | int64 | type of officer assignment at time of stop (code) |
| **assignment** | str | type of officer assignment at time of stop (description) |
| **intersection** | str | location of stop - intersecting street name |
| **block** | float64 | location of stop - hundred block |
| **land\_mark** | str | location of stop - landmark |
| **street** | str | location of stop - street name |
| **highway\_exit** | str | location of stop - highway exit |
| **isschool** | int64 | did stop occur at a school? |
| **school\_name** | str | name of school where stop occurred |
| **cityname** | str | name of city where stop occurred |
| **beat** | int64 | location of stop - SDPD beat |
| **beat\_name** | str | location of stop - SDPD beat/neighborhood name |
| **pid** | int64 | unique identifer for person on a stop |
| **isstudent** | int64 | was person stopped a student? |
| **perceived\_limited\_english** | int64 | officer's perception that the person stopped has limited or no fluency in English |
| **perceived\_age** | int64 | officer's perception of the approximate age of the person stopped |
| **perceived\_gender** | str | officer's perception of the gender of the person stopped (description) |
| **gender\_nonconforming** | int64 | officer's perception of whether the person stopped is gender nonconforming |
| **gend** | int64 | officer's perception of the gender of the person stopped (code) |
| **gend\_nc** | float64 | officer's perception of whether the person stopped is gender nonconforming |
| **perceived\_lgbt** | str | officer's perception of whether the person stopped is lesbian, gay, bisexual or transgender |

Result of Stop: The dataset records the results of a stop for an individual who was stopped by the San Diego Police Department, and this information was collected in accordance with the guidelines set forth in Government Code section 12525.5, which was established by the Racial and Identity Profiling Act of 2015 (AB 953), or RIPA. Each row in the file represents one outcome for a particular individual who was stopped by the police, and a stop may have multiple outcomes. The individual who was stopped is identified by a unique identifier in the PID field, and the stop itself is identified by a unique identifier in the stop\_id field. This data set has 6 columns and 707933 instances. A short description of all the attribute is given as below:

**Table 2 : Brief description of attribute in Stop Result Dataset**

|  |  |  |
| --- | --- | --- |
| **Field** | **Field Type** | **Description** |
| **stop\_id** | int64 | unique identifer for stop |
| **pid** | int64 | unique identifer for person on a stop |
| **resultkey** | int64 | outcome(s) of stop (code) |
| **result** | str | outcome(s) of stop (description) |
| **code** | float64 | specific violation if stop outcome is warning, citation or custodial arrest (code) |
| **resulttext** | str | specific violation if stop outcome is warning, citation or custodial arrest (description) |

Action Taken: The dataset includes the actions taken by a police officer towards an individual who was stopped by the San Diego Police Department, and this data was collected in accordance with the requirements outlined in Government Code section 12525.5, which was established by the Racial and Identity Profiling Act of 2015 (AB 953), or RIPA. Each row in the file represents one action taken toward a specific individual who was stopped by the police, and an officer may have taken multiple actions toward the same person. The individual who was stopped is identified by a unique identifier in the PID field, and the stop itself is identified by a unique identifier in the stop\_id field. This dataset has 4 columns and 1031391 instances. A short description of all the attribute is given as below:

**Table 3 : Brief Description of attributes in Action Taken dataset**

|  |  |  |
| --- | --- | --- |
| **Field** | **Field type** | **Description** |
| **stop\_id** | int64 | unique identifier for stop |
| **pid** | int64 | unique identifier for person on a stop |
| **action** | str | officer's actions toward the person stopped |
| **consented** | str | response by the person stopped if the officer asked for consent to search |

For the analysis, we removed the unnecessary columns from each dataset and modified some of the attributes. Later, we combined all the data sets to pre-process them into Weka. All the steps of data processing are described in detail in the forthcoming sections.

### DATA PREPARATION

The purpose of performing data preparation is to ensure that the data is suitable for analysis and to minimize errors that can occur during the analysis. Without proper data preparation, the analysis may produce inaccurate or incomplete results, which can lead to incorrect conclusions and decisions. The data preparation process helps to ensure that the data is clean, consistent, and accurate, making it easier to work with and analyze. Moreover, the process helps to identify and address any missing or erroneous data, which can affect the analysis results. By performing data preparation, we can ensure that the analysis is accurate, reliable, and based on a complete and accurate data set, which can help us to make informed decisions and generate valuable insights.

1. **Data Selection**

* We excluded several variables, including ori, agency, intersection, address block, landmark, address street, highway exit, school name, beat, beat name, gender nonconforming, gender, and gender non-conforming because they did not have any impact on the actions taken by the police. These variables were deemed irrelevant to the analysis and did not contribute to our understanding of the actions taken by the police during stops. As a result, we chose to exclude them from the basic details data set to streamline our analysis and focus on the variables that were more directly relevant to our research question.
* Using filters on the date column, we removed all the date entries except 2018. We only analyzed the data for 2018.
* In some cases, a single stop can involve multiple individuals, which can result in multiple p\_id values for a single stop\_id. However, for the purpose of our analysis, we were only interested in whether action was taken against the driver or not. Therefore, we assumed that the driver would always be assigned a p\_id of 1 for a given stop\_id. Consequently, we removed all instances of multiple p\_id values for a single stop\_id, as we only needed to consider the action taken against the driver.
* Similarly, in the action taken dataset, we removed all rows that did not belong to the year 2018. This was done to ensure that our analysis was focused on a consistent time period. We also removed all instances of multiple p\_id values for a single stop\_id, as we were only interested in the actions taken against the driver. Furthermore, we removed the 'consent' column from the action taken dataset, as approximately 95% of the data in this column was missing. We made the decision to exclude this variable from our analysis, as the high degree of missing data made it difficult to derive meaningful insights from it.
* We decided to remove the 'result' dataset from our analysis because we found that we had sufficient attributes in the other datasets to determine whether or not an action was taken during a stop. This decision was made to streamline our analysis and focus on the most relevant variables that would help us answer our research questions. By removing this dataset, we were able to simplify our analysis and reduce unnecessary complexity in our data.

1. **Data Cleaning**

* After selecting the relevant data for our analysis, we found that all the data in our datasets was unique, and we did not come across any duplicates. This ensured that our analysis was based on accurate and reliable data, as there were no instances of duplicate data that could skew our results or misrepresent the true nature of the stops made by the San Diego Police Department.

1. **Data Construction**

* Converting string attributes into nominal attributes was one of the most significant and time-consuming procedures we undertook for our project. We had to create several new columns to ensure that the data was relevant and consistent with our analysis. This involved transforming string attributes into categorical or nominal variables that could be more easily analyzed and interpreted. This process required a lot of effort and attention to detail, as we had to ensure that the data was accurate and reliable for our analysis.
* To begin with, we generated a column for the months and obtained the month in which the stop took place. The formula used in Excel for this purpose was = MONTH(A1).
* Next, we created a Time of the Day column using the existing time column. The stops were categorized into different time zones including Morning, Afternoon, Evening, and Night. To achieve this, an Excel formula was used: =IF(D2<=TIMEVALUE("03:59:59"),"Night",IF(D2<=TIMEVALUE("10:59:59"),"Morning",IF(D2<=TIMEVALUE("16:59:59"),"Afternoon",IF(D2<=TIMEVALUE("9:59:59"),"Evening","Night"))))

**Table 4: Time zone division table**

|  |  |
| --- | --- |
| **Time Zone** | **Time Period** |
| 4:00 AM to 10:59 AM | Morning |
| 11:00 AM to 4:59 PM | Afternoon |
| 5:00 PM to 9:59 PM | Evening |
| 10:00 PM to 3:59 AM | Night |

* Thirdly, we created a Stop Duration column that categorized the duration of each stop as Short, Moderate, Long, or Extended. To do this, an Excel formula was used: =IF(E2<61,"Short",IF(E2<300,"Moderate",IF(E2<720,"Long","Extended"))).

**Table 5: Duration division table**

|  |  |
| --- | --- |
| **Duration (In Minutes)** | **Period** |
| 0 to 60 | Short |
| 61 to 299 | Moderate |
| 300 to 719 | Long |
| 720 and over | Extended |

* Subsequently, we conducted research on how to divide the city and discovered that all cities are classified into different county regions. Therefore, a County column was created to list the top five counties. The excel formula used: =VLOOKUP(I2,Sheet1!$A$2:$B$38,2,TRUE), The cell I2 contains the City Column that needs to be updated. The table used for the VLOOKUP function is !$A$2:$B$38, where Column A represents the address city and Column B represents the corresponding county.
* Finally, we created an Age Group column, where all the age ranges were divided into categories including Baby, Child, Youth, Adult, and Senior. An Excel formula was used for this purpose: =IF(ISBLANK(N2),"N/A",IF(N2<2,"BABY",IF(N2<12,"CHILD",IF(N2<25,"YOUTH",IF(N2<=60,"ADULT",IF(N2>60,"SENIOR","ERROR")))))).

**Table 6: Age division table**

|  |  |
| --- | --- |
| **Age Division** | **Age Group** |
| 0 - 2 | BABY |
| 2 - 25 | YOUTH |
| 25 - 60 | ADULT |
| 60 and over | SENIOR |

1. **Data Formatting**

* To facilitate analysis, some data was formatted using "True" and "False". Specifically, for the Action Taken column, which is our class attribute, we formatted the data as "Yes" if any action was taken against the person, and as "No" if no actions were taken.

All the CRISP-DM steps have been completed, and now we have the final dataset for further analysis, which includes basic details and the Action Taken column data set identified by the Stop ID column.

### DATA VISUALISATION AND CLASSIFICATION

After completing preprocessing and following all the appropriate steps in the CRISP-DM process, we are now ready to import our data into Weka. The data set contains 79774 unique instances with 15 columns. Here is a brief overview of each column.

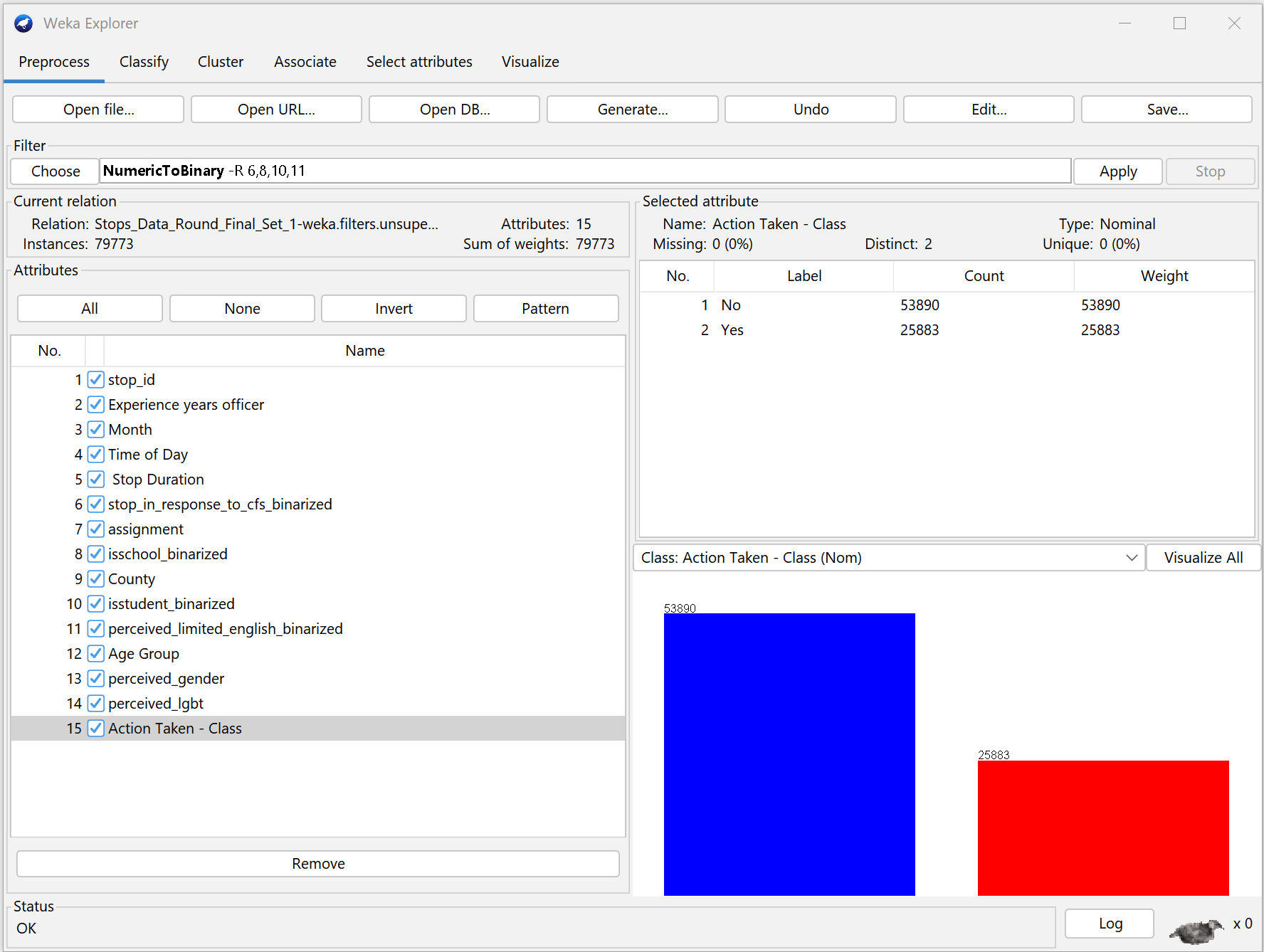
**Table 7: Brief Description of attributes of the final dataset**

|  |  |  |
| --- | --- | --- |
| **Field** | **Field\_type** | **Description** |
| Stop\_id | id | Stop\_id of the stop |
| Experience Years Officer | int | Number of years of experience the officer had, who stopped the vehicle |
| Month | Nominal | The month of the year in which the stop occurred. (July to December 2018) |
| Time of the Day | Nominal | The time zone of the day when stopping occurred |
| Stop Duration | Nominal | How long did the stop occur |
| Stop in response to cfs | Binary | was the stop made in response to a call for service? |
| Assignment | Nominal | type of officer assignment at time of stop (description) |
| IsSchool | Binary | Did the stop occurred at/near school |
| County | Nominal | County of the region where stop occurred |
| IsStudent | Binary | Is the driver a student |
| Perceived\_limited\_english | Binary | officer's perception that the person stopped has limited or no fluency in English |
| Age Group | Nominal | Age group in which the person belong to. |
| Perceived Gender | Nominal | officer's perception of the gender of the person stopped (description) |
| Perceived LGBTQ | Nominal | officer's perception of whether the person stopped is lesbian, gay, bisexual or transgender |
| Action Taken | Nominal | Was an action taken against the driver or no |

After loading the data into Weka, all the attributes were verified if they were correctly converted. Some of the attributes such as Stop in response to cfs, IsStudent, IsSchool, Perceived\_Limited\_English etc. were converted to binary using Filter 🡪 Weka 🡪 UnSupervised 🡪 Attribute 🡪 NumerictoBinary.

1. **Classification**

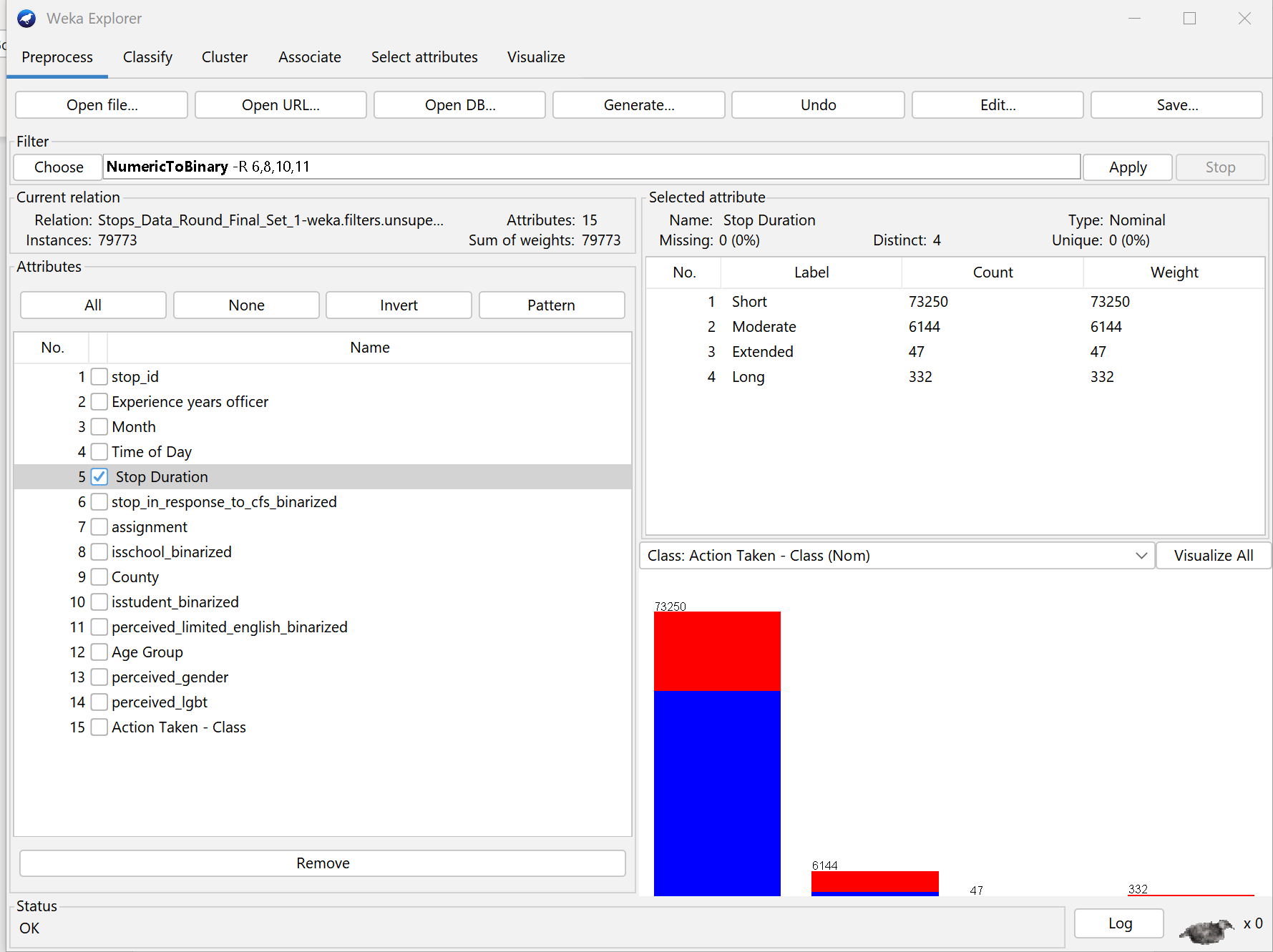
The screenshot given below shows the proper classification of data in terms of whether an action was taken or not against the driver. Among 79774 stops occurred between July 2018 to December 2018, action was taken against only 25883 drivers, which is less than even 50% of total stops.



**Figure 1: Classification whether an Action was Taken Against the Driver or not.**

1. **Visualization**

Upon classifying the data on whether the action was taken (red graph) or not (blue graph), we tried visualizing some bar graphs and plots to see if we can determine some factors that affected the decision of action being taken against the driver.

****

**Figure 2: Stop Duration over the Action Taken**

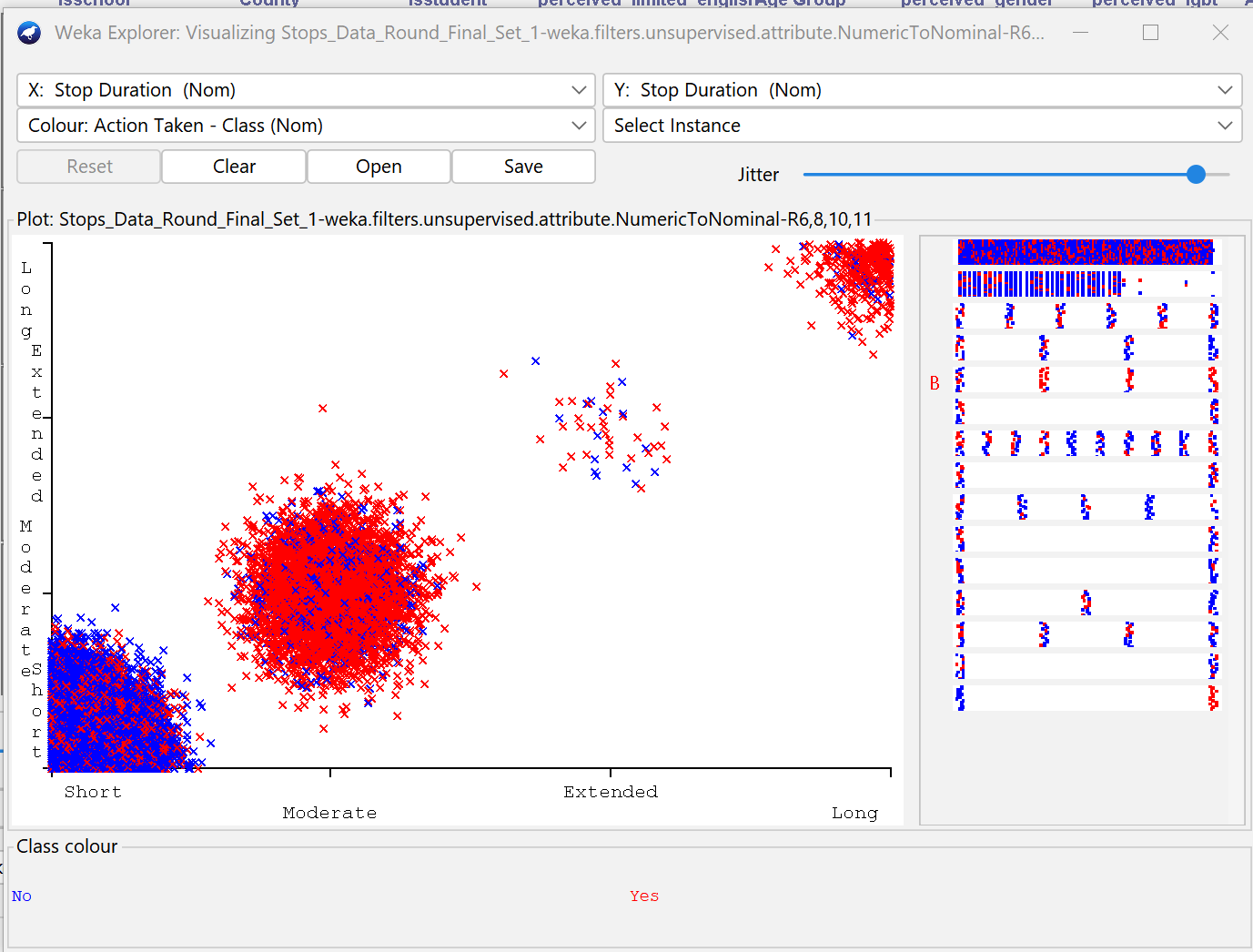
Based on Figure 2, we can observe that there is a correlation between stop duration and the actions taken against drivers. When the stop duration is short, the majority of officers did not take any action. However, when the stop duration increased to moderate or long, the majority of office did take appropriate actions. It is important to note that for rare instances with extended stop duration, it is difficult to visualize the data on the graph.

Graphical user interface, application

Description automatically generated

**Figure 3: Plot of Experience of Stop Officer v/s Action Taken**

Figure 3 demonstrates a decrease in the number of stops as the number of experiences increases by 30 years. This trend may be attributable to various factors, such as a reduction in the number of officers with 30 years of experience due to retirement. Nevertheless, it is possible that an officer with over or equal to 50 years of experience might exist and could serve as an outlier in the data.



**Figure 4: Plot of Time Period for which the stop occured**

In Figure 4, the duration of each stop is shown, and it is evident that the action taken by the officers varies depending on how long the car was stopped. The graph indicates that there is a higher likelihood of action being taken by the officers when the stop lasted for a longer time. This could be because more severe cases require a longer period of investigation. It is worth noting that there were only 47 instances where the stop lasted for an extended period, and these may be considered as outliers in the project.

### MODELING AND Evaluation

1. **KNN CLASSIFICATION**

We evaluated the performance of our dataset in classifying actions using the K-nearest neighbors (KNN) algorithm. We tested the algorithm with various values of k and found that for our large dataset with over 70,000 instances, the optimal value of k is approximately the square root of the total number of instances, which is around 282.44.

As a result, we considered odd values of k between 273 to 285 and used the IBK algorithm from the Classify tab in Weka. Then, we applied the J48 decision tree classifier from the Trees option. We split the dataset into training and testing sets using a percentage split of 75% and tested the algorithm's performance with different values of k. The results obtained from this experiment are presented below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Values of K** | **True Positive** | **False Positive** | **Correctly Classified Instances** |
| 273 | 0: 0.979  1: 0.207 | 0: 0.793  1: 0.021 | 72.9478% |
| 275 | 0: 0.979  1: 0.208 | 0: 0.792  1: 0.021 | 72.988% |
| 277 | 0: 0.980  1: 0.207 | 0: 0.793  1: 0.020 | 72.988% |
| 279 | 0: 0.980  1: 0.206 | 0: 0.794  1: 0.020 | 72.968% |
| 281 | 0: 0.980  1: 0.206 | 0: 0.794  1: 0.206 | 72.968% |
| 283 | 0: 0.980  1: 0.205 | 0: 0.795  1: 0.020 | 72.9579% |
| 285 | 0: 0.981  1: 0.205 | 0: 0.795  1: 0.019 | 72.9629% |

From above tabulated data, we got to know the best and worst case of k by analyzing the squared error.

1. **DECISION TREE**

We tried generating the decision tree for our dataset which will help us understand the factors which were responsible for decision-making whether the action was taken or not.

To generate the decision tree on a very large scale of data, we needed to increase the minimum number of objects to a significantly larger value so that we can get an appropriate characterized decision tree.

Using Classify tab, Choose Weka 🡪 Classifier 🡪 Trees 🡪 J48. Keeping the cross-validation folds set to 15 and the unpruned property to ‘True’, we ran the model by changing the minimum number of objects to 1000, 1500, and 2000. The results are tabulated as:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MinNumOfObjects** | **No. of Leaves** | **Size of tree** | **Correctly Classified Instance** | **Confusion Matrix** |
| 1000 | 21 | 27 | 74.3685% | == Confusion Matrix ===  a b <-- classified as  50318 3572 | a = No  16875 9008 | b = Yes |
| 1500 | 9 | 13 | 74.4124% | === Confusion Matrix ===  a b <-- classified as  50883 3007 | a = No  17405 8478 | b = Yes |
| 2000 | 13 | 19 | 74.4525% | === Confusion Matrix ===  a b <-- classified as  50286 3604 | a = No  16776 9107 | b = Yes |
| 2500 | 6 | 9 | 74.4487% | === Confusion Matrix ===  a b <-- classified as  50471 3419 | a = No  16964 8919 | b = Yes |

We are getting the best results when the minimum number of objects is 2000 and the minimum number of objects is 2500. A brief description of the decision tree is given below:

Diagram

Description automatically generated

**Figure 5: Decision Tree with minimum number of objects as 1500**

According to the information presented in Figure 5, the attribute that had the highest level of entropy was the duration of the stop. If the stop lasted for a longer period of time, then action was usually taken against the driver. However, if the stop was shorter, the decision made by the officer depended on their level of experience. If an officer had more than 9 years of experience, then action was usually not taken. On the other hand, if the officer had less than 9 years of experience, the decision was based on whether the stop was in response to a service call or not. If it was, then action was taken, but if not, then the decision was based on the gender of the individual. Generally, the action was taken if the individual was male, female, or transgender male/boy, but if the perceived gender was transgender girl/woman, the action was less likely to be taken.

Diagram

Description automatically generated

**Figure 6: Decision Tree when the minimum number of objects is 2000**

Figure 6 demonstrates that the stop duration has the highest degree of entropy. The initial level of the decision tree in Figure 6 is the same as in Figure 5. From the second level, if the officer has more than 9 years of experience, the decision is based on the time of day that the stop occurred. If the stop took place in the night, afternoon, or evening, no action is taken. However, if the stop occurred in the morning, the decision is based on the location of the stop. If the stop took place on a stoop ID that is greater than 44189, no action is taken. On the other hand, the decision is based on gender again. However, in contrast to Figure 5, the final level of the decision tree in Figure 6 is entirely different. It indicates that action was only taken against transgender women/girls, whereas no action was taken against male, female, or transgender boys/men.

In the conclusion of the decision trees, we ran the random forest algorithm against all the generated trees and we got the following results.

Number of Iterations: 100

Correctly identified instances: 71.8012%

Confusion Matrix:

a b <-- classified as

43289 10601 | a = No

11894 13989 | b = Yes

1. **CLUSTERING**
   1. **Clustering by Farthest Fast**

The team attempted to cluster a dataset by varying the number of clusters to determine the optimal value for k. However, for clustering to be performed, all attributes must be numeric, except for the class attribute. The team has provided a table with the results for several models.

|  |  |
| --- | --- |
| **Value of K** | **Incorrectly classified Clusters** |
| 2 | 32.348% |
| 3 | 40.8196% |
| 4 | 43.422% |
| 5 | 57.1058% |
| 6 | 62.5136% |

From the results, it appears that the number of clusters that are incorrectly classified increases as the value of k increases. Therefore, for the given dataset, it seems that the optimal value for k is 2.

Detailed Results are:

Cluster centroids:

|  |  |
| --- | --- |
| Cluster 0 | 7699.0 6.0 July Afternoon Moderate 0 Other 0 Central Sandiego 0 0 ADULT Male No |
| Cluster 1 | 78939.0 6.0 December Morning Moderate 1 Patrol, traffic enforcement, field operations 1 Central Sandiego 1 0 YOUTH Transgender man/boy Yes |
| Cluster 2 | 50449.0 1.0 October Night Short 1 Patrol, traffic enforcement, field operations 0 Central Sandiego 0 1 SENIOR Female No |
| Cluster 3 | 44476.0 26.0 September Evening Short 0 K1-12 public school inlcuding school resource officer or school police officer 1 Central Sandiego 1 0 YOUTH Male No |
| Cluster 4 | 42070.0 24.0 September Morning Short 0 Patrol, traffic enforcement, field operations 0 South Bay 0 0 ADULT Female Yes |
| Cluster 5 | 70265.0 10.0 November Evening Short 0 Investigative/detective 0 Central Sandiego 0 1 YOUTH Transgender man/boy Yes |
| Cluster 6 | 79405.0 4.0 December Evening Short 0 Task force 0 East County 0 0 SENIOR Male No |

0 1 2 3 4 5 6 <-- assigned to cluster

25577 32 6447 1036 13300 675 6823 | No

14425 105 4327 300 3111 307 3308 | Yes

Cluster 0 <-- No

Cluster 1 <-- No class

Cluster 2 <-- Yes

Cluster 3 <-- No class

Cluster 4 <-- No class

Cluster 5 <-- No class

Cluster 6 <-- No class

* 1. **Clustering by K-Means**

The team conducted a k-means analysis to explore how varying the value of k would impact the results for their dataset. The following outcomes were obtained.

|  |  |
| --- | --- |
| **Value of K** | **Incorrectly classified Clusters** |
| 2 | 48.5753% |
| 3 | 61.7953% |
| 4 | 70.7019% |
| 5 | 70.8097% |
| 6 | 71.1005% |

Model and evaluation of the training set

* Clustered Instances

0 38214 ( 48%)

1 41559 ( 52%)

Class attribute: Action Taken - Class

Classes to Clusters:

0 1 <-- assigned to cluster

26677 27213 | No

11537 14346 | Yes

Cluster 0 <-- No

Cluster 1 <-- Yes

Incorrectly clustered instances: 38750.0 48.5753 %

From the provided table, it can be observed that the number of clusters that are incorrectly classified increases as the value of k increases. The optimal value for k appears to be the point at which the increase in incorrectly classified clusters begins to level off, which is referred to as the elbow point. In this case, it seems that k = 2 is the elbow point for the given dataset.

1. **Outlier Detection**