**BIG DATA IN IMPROVING ROAD SAFETY OF A CITY**

Road and traffic accidents are uncertain and unpredictable incidents and their analysis requires the knowledge of the factors affecting them. Data analytics using big data helps in discovering patterns, correlation and finding trend. It works on all kinds of data. The existing road safety and flow prediction mainly uses traffic prediction models that are shallow and unsatisfying for many real-world applications. This situation drives me to consider this topic to create architecture models to analyse the large set of traffic data.

The analysis and prediction of such data will help us in improving the public safety by reducing accidents and drunk-driving incidents and may also help the police agencies to track and analyse crime-related communications, the accident-prone areas and reduce risk and the number of critical incidents to serve communities better. So, the big data analysis on accident data may help trends over time to develop an “early warning” systems to head off future problems.

Traffic datasets are massive with a huge amount of data in it and its very hard to efficiently process such data. Therefore, the clustering algorithms can be used to deal with the huge data by classifying the data into various cluster/groups. So, the traffic control planners must have all possible solutions to correlate, process and able to get the best possible results of big data. To get all these possible solutions one must apply big data algorithms and techniques.

Applying big data techniques can revolutionize traffic safety. In this project, I am using advanced analytics and predictive modelling that can make the path for more centralized decisions on traffic systems. Using predictive modelling and all the big data algorithms, I am more concentrating on bad weather, previous driver condition who is involved in the violations/ accidents and the major accident-prone areas, and help to reduce them to the maximum extent possible.

**RELATED DOMAIN OF STUDY:**

Traffic violation/control

**ALGORITHMS CURRENTLY CONSIDERING:**

* Pre-processing
* Pandas and Numpy
* MapReduce
* Clustering Algorithm
* Association rules (Apriori Algorithm)
* Data Visualization

**Pre-processing**: Eliminating the columns/attributes that are not essential and considering only the attributes that are required for the analysis.

After pre-processing, I am trying to import the csv into the panda data frame object and then remove the rows with the null/empty values from the data frame and store them in a new data frame and finally save the modified dataset to a new csv file.

I am using clustering algorithm to divide the datasets into different clusters based on the years when the violation has occurred i.e. I am creating six clusters one for each year (Date of Stop: 2012 to 2017) and each cluster having all the violation data for that year. After clustering the dataset, I am using the Apriori algorithm to mine the frequent item sets or association rule mining for each cluster. For example: cluster 1 will have all the violation data for the year 2012. So, we can create the association rules for the attributes such as Accident, time and gender. Therefore, the association rules can be used to identify the various traffic violation circumstances and accident-prone areas for every cluster.

**POSSIBLE DATA SOURCES:**

* https://data.montgomerycountymd.gov/Public-Safety/Traffic-Violations/4mse-ku6q
* <https://catalog.data.gov/dataset?tags=traffic>
* <https://catalog.data.gov/dataset/traffic-violations-56dda>
* <http://data.imap.maryland.gov/datasets/maryland-annual-average-daily-traffic-annual-average-daily-traffic-sha-statewide-aadt-lines>?

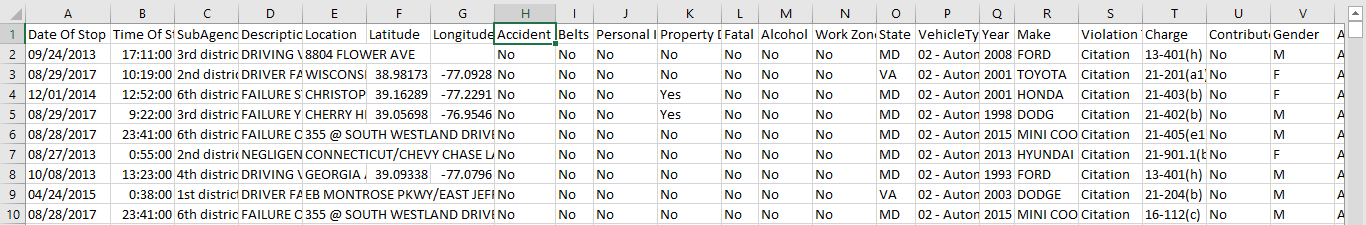
This dataset contains traffic violation information from all electronic traffic violations issued in the County. Any information that can be used to uniquely identify the vehicle, the vehicle owner or the officer issuing the violation will not be published.

**Samples of the data source:**

Date of Stop, Time of Stop, Sub Agency, Description, Location, Latitude, Longitude, Accident, Belts, Personal Injury, Property Damage, Fatal, Alcohol, Work Zone, State, Vehicle Type, Year, Make, Violation Type, Charge, Contributed To Accident, Gender, Arrest Type

09/24/2013, 17:11:00, "3rd district, Silver Spring", DRIVING VEHICLE ON HIGHWAY WITH SUSPENDED REGISTRATION,8804 FLOWER AVE, , , No, No, No, No, No, No, No, MD, 02 - Automobile, 2008, FORD, Citation, 13-401(h), No, M, A - Marked Patrol

08/29/2017, 10:19:00, "2nd district, Bethesda", DRIVER FAILURE TO OBEY PROPERLY PLACED TRAFFIC CONTROL DEVICE INSTRUCTIONS, WISCON SIN AVE@ ELM ST, 38.981725, -77.09275667, No, No, No, No, No, No, No, VA, 02 - Automobile, 2001, TOYOTA, Citation, 21-201(a1), No, F, A - Marked Patrol



**GRAPHICS:**

I am planning to display the accidents/ Violation type for each year (i.e. for each cluster) and use the bar graphs and pie charts to visualize

* Number of accidents/ violation type for each year.
* Percentage of accidents occurring in male and females for each year.
* For each year the no. of accidents for each state occurring due to alcohol.

**CURRENT CHALLENGES:**

* Difficult to clean the dataset with one million tuples approx.
* Clustering the data into different clusters based on the year (Date of Stop)
* Visualize the data to display them in graphs (bar graphs and pie charts) and to describe the end results.

**REFERENCE URL’S:**

* <http://www.ijritcc.org/download/browse/Volume_4_Issues/May_16_Volume_4_Issue_5/1465367456_08-06-2016.pdf>
* <https://link.springer.com/article/10.1186/s40537-015-0035-y>
* <https://www.researchgate.net/publication/284282571_A_data_mining_framework_to_analyze_road_accident_data>
* <http://www.ijritcc.org/download/browse/Volume_4_Issues/July_16_Volume_4_Issue_7/1468993772_20-07-2016.pdf>
* <http://www.ibmbigdatahub.com/blog/6-ways-analytics-can-help-manage-traffic-and-reduce-accidents>