Data Analytics Route Navigation in Toronto Based on Accidents Data

Final Report: EECS 6414

Karan Deep Singh
singkara@yorku.ca
York University
Toronto, Canada

Mahima Chaudhary
cmahima@cse.yorku.ca
York University
Toronto, Canada

Payal Goyal
payalg7@yorku.ca
York University
Toronto, Canada

ABSTRACT

The driving conditions in the city of Toronto are complex and could sometimes even be dangerous, especially due to rise in traffic in recent years. This paper discusses 3 approaches to the Toronto Accidents Data which we gathered from the City of Toronto Police Data Portal. First, we explored the data-set by applying several Data Analytic Visualization using the Tableau Software, which helped us gain us several insights such as– the determining friendly weather conditions and best time of the day to drive in Toronto. Second, we applied the two machine learning strategies of Clustering and Classification and used various models within each strategy to determine the severity of an accident as well as the accident prone neighbourhoods. Finally, we prepared a navigation application which avoids the accident prone areas on top of the APIs provided by HereMaps[19] and the data results from Machine Learning, in addition to other functions.

CCS CONCEPTS

• **Human Centred Computing** → **Visualization**; • **Computer systems organization** → *Real Time Systems*; • **Theory of Computation** → Data Integration; • **Networks** → Network algorithms; • **Computing Methodologies** → Machine Learning; • **Applied Computing** → Transportation

.

KEYWORDS

Accident Prediction, Data Visualization, Data Mining, Feature Engineering, Machine Learning, KSI (killed or Seriously Injured) Data set

1 INTRODUCTION AND MOTIVATION

Our project is about exploring several machine learning strategies combined with the data analytics on the collisions data-set in Toronto. In recent research, Data Visualization with Machine Learning has been used by several researchers [20] to essentially predict the trends and derive insights. On the similar lines, our project is a thorough combination of visual representation and ML models that we have applied on the KSI Dataset(Killed or Seriously Injured).

The Data Domain of our dataset is clean, rich and contains a range of datatypes such as- Boolean, numerical, categorical, time as well as geospatial (latitudes, longitudes). Overall data quality of good, and there are not much missing values in essential features. Since the data is for 10 years(2007-2017), we have also demonstrated numerous trends in our visualizations.

The goal of our project is 3 fold:

- Data Analytics: Firstly, have derived percipient information from the analytics to suggest improvements for better and safer roads. This could be information such as trend in accidents, road conditions related to accidents and weather conditions friendly to drive.
- Machine Learning: We focused on classification of accidents in order to find their severity. In addition, we have used clustering to find the accident prone neighbourhoods. This will be discussed in detail below.
- Navigation Application: In the end, we have used to data to build an application which not only will highlight accident prone neighbourhoods, but also will help users navigate through the safe transit in Toronto.

Our main motivation behind this project was the recent surge in the Toronto Collisions due to traffic and population increase, especially after 2010. The traffic conditions have exacerbated and this calls for detailed analytics about how to reduce accidents due to increased traffic. Therefore, this project is our humble attempt to figure the major reasons which cause cause collisions and give feedback to users on how to remain safe. This feedback to the users is given through the navigation application developed by us using the clustering algorithms to find out the most accident prone areas on the map of Toronto and avoiding them while navigation.

Our Data Analytics and ML yield the answers to the following questions:

- Machine Learning:
  - List of neighbourhoods sorted according to the accident prone severity.
  - Classify how severe an accident is, given its input features.
- Data Analytics:
  - Prime causes for accidents.
  - Time/day/month/types with high collision frequency
  -Interesting time/Age based trends.
  -Other Contributing factors like Speeding, Visibility etc.
- Navigation Application: Avoid collision by suggesting alternative navigation routes to Torontonians.

The analysis is important as we need to use clustering to identify the accident prone neighbourhoods. Our clustering algorithm compares the Neighbourhood's density in collisions and also takes

area into account. A simple sort of accidents according to neighbourhoods will not suffice, since the data is temporal.

We tested our functionality by dividing our data-set into 2 parts i.e Training set and Test set. In addition, we have also applied cross-validation strategy to determine the test set, whose details are in the kernel. The test set size we have taken is 10% of the total size. In order to determine the accuracy, our ML model predicts whether an accident is fatal/not based on the features on the separated out test data. In addition we have also tested the navigation app based on the data that whether the app is successfully able to suggest alternate routes through manual testing by routing through the collision prone areas.

We have also used clustering techniques to find the accident prone neighborhoods in Toronto. This information is used by our navigation application to suggest alternate route that are safer to travel.

A few potential applications of this project could be for the Toronto Police/Government to figure out the major causes of accidents and alleviate the situation. The citizens could also use the navigation application to ensure a safe maneuver when travelling through the city.
In future, our infrastucture can be integrated with IOT sensors of vehicles such that they provide analytics based on real time data. We could also integrate the Navigation Application with the users historical data and provide him insights about the safety comparison in other route.

This report is organized as follows: Section 2 discusses the related work upon similar accidents data set. Section 3 demonstrates in detail about formal definition and discussed about various Prediction and clustering models we use. Section 4 outlines the methodology in detail. Section 5 contains the evaluation and results. Section 6 finally concludes the report with our final remarks and future recommendations.

2 PROBLEM DEFINITION

We have used 2 major Algorithmic Models in our work i.e. clustering [12, 22]and prediction models[3, 4] to derive useful information from the data-set. This section will discuss the two components(along with the hardness of each method) in detail starting with clustering.

Clustering models have been used to find the neighbourhoods that experienced highest number of Fatal accidents, which is done on the basis of top features that we get from Extra Trees Classifier model. Extra Trees Classifier is an ensemble type learning technique which aggregates the results of multiple uncorrelated decision trees collected in a “forest” to output it’s classification result. We have employed two types of clustering models namely KMeans Clustering and Agglomerative Clustering. The clustering models aim to minimize the following equation:

$$J(V) = \Sigma_{i=1}^{c}\Sigma_{j=1}^{c_i}(||x_i - v_j||)^2$$

where,
'$||x_i - v_i||$' is the euclidean distance between the centre of i^{th} cluster i.e. x_i and j^{th} point in that cluster i.e. v_i.
'c_i' is the number of datapoints in i^{th} cluster
'c' is number of cluster centres.

The above equation makes the clustering algorithm NP-hard. However, the complexity of clustering algorithm is O(tknd)for a fixed number t of iterations , for n (d-dimensional) points, where k is the number of centroids (or clusters).

The formal definitions of the two clustering models is as given below:

- k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
- Agglomerative Clustering is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. This is achieved with th ehelp of a dendrogram.

Prediction Models are used to predict whether or not an accident is fatal. For this we take the features returned from Extra Tree Classifier as independent features and the TARGET column is "FATAL"– which demonstrates fatality of the accident. This study has compared 5 Learning models–

- Logistic Regression is a regression analysis tool to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. The Logistic Regression model works on the equation below,it relates the independent variable, X, to the rolling mean P:

  $$P = 1/1 - e^{a+bX}$$

  where where P is the probability of a 1 (the proportion of 1s, the mean of Y), e is the base of the natural logarithm (about 2.718) and a and b are the parameters of the model. The time complexity of LR is O(dcnE), where d is features, c is classes, n is size of data and E is the number of epochs.
- K Nearest Neighbours belongs to the supervised learning domain and is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data. The KNN approach tries to optimize the similarity measure like the euclidean distance shown in the equation below:

  $$D = \sqrt{\Sigma_{i=1}^{k}(x_i - y_i)^2}$$

  where K is the number of neighbors. The complexity of KNN is O(ndk) where n is the size of dataset, d is dimension of data and k is the number of neigbors in consideration.
- Decision Trees is a flowchart-like tree structure where an internal node represents feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The decision tree method selects attributes on the basis of entropy and information gain. These are calculated by the equations shown below:

  $$Entropy(S) = \Sigma_{i=1}^{c} - p_i \log_2 p_i$$

  Where S ⟶ Current state, Pi ⟶ Probability of an event i of state S or Percentage of class i in a node of state S.
  Entropy of multiple attributes is given by:

  $$Entropy(T, X) = \Sigma_{c\varepsilon X} P(c) Entropy(c)$$

  where where T⟶ Current state, X ⟶ Selected attribute, c ⟶ Possible values of selected attricute, P(c) ⟶ Probability

of particular value of the attribute, Entropy(c) ⟶ Entropy for the value of attribute.
Finally Information gain is calculated as:

$$InformationGain(T, X) = Entropy(T) - Entropy(T, X)$$

The Attribute with highest Information gain is selected at a given node of the tree. The complexity of decision tree is O(Nkd) where N is number of examples, k is dimension of data and d is depth of tree.

- Random Forest is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees and it gets prediction from each tree and selects the best solution by means of voting. The Random forest might also take the average result of all the regression trees each of which is formed using a different sample of data. After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x,this is done according to the below equation:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

where B is the number of trees and f_b is regression tree model on seen data. The complexity of Random Forest is O(ntree * mtry * d * n) where ntree is number of trees, mtry is variables to sample at each node, d is depth and n is number of features.

- Neural Network is a network or circuit of neurons. These artificial networks may be used for predictive modeling, adaptive control and applications where they can be trained via a dataset. Self-learning resulting from experience can occur within networks, which can derive conclusions from a complex and seemingly unrelated set of information. Each neuron in the neural network applies an activation to its input according to the equation below:

$$a_j^l = \sigma(\Sigma_k w_{jk}^l a_k^{l-1} + b_j^l)$$

where a_i^j is output of current node, σ is the activation function, w_{jk}^l are weights from previous layer to current layer, a_k^{l-1} are inputs from previous layer and b_j^l is bias value of current layer.

To calculate the loss of our prediction functions, we calculate the prediction algorithm is Accuracy Score, which is equal to the ratio of number of correct predictions and total number of predictions.

$$Accuracy = TP + TN/TP + FP + FN + TN$$

TP:True Positives TN:True Negatives FP:False Positives FN:False Negatives

3 RELATED WORK

Some of the related previous works[5, 7, 14] that we studied include the US Accidents data-set [9, 15, 16] which has a few research papers published upon, wherein they explore data analytics and tried to answer the questions such as– which state has highest accidents, which sates are safer to drive in US etc., Furthermore, similar work has been mentioned in the following papers– A Countrywide Traffic Accident Dataset [18] , Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights[17] and Unsupervised Traffic Accident Detection in First-Person Videos [23].

Although thorough, the existing papers did not use machine learning algorithms [21] excessively to predict fatality of the collisions nor currently any application exists which would suggest alternate navigation routes based on collision conditions in an area. Furthermore, the current research papers have not applied the clustering methodologies to figure the most accidental neighbourhoods. The existing work also lacks use of Data mining techniques to find important features that contribute to accidents.
In our analysis, we are adding value on the existing papers firstly by developing an application to suggest collision free routes, no other paper we know of discusses the possibility/idea of such an application. Secondly, our work is novel based on the derived analysis from clustering and classification techniques in Machine Learning.

4 METHODOLOGY

We have used 3 types of Data Analysis models i.e. Descriptive , Diagnostic and Predictive Analysis. We have used ML models for Diagnostic and Predictive analysis, while we have used Tableau for Descriptive analysis. In the subsequent sections, we will explain the steps we took for each methodology and how that was sufficient for our analysis.

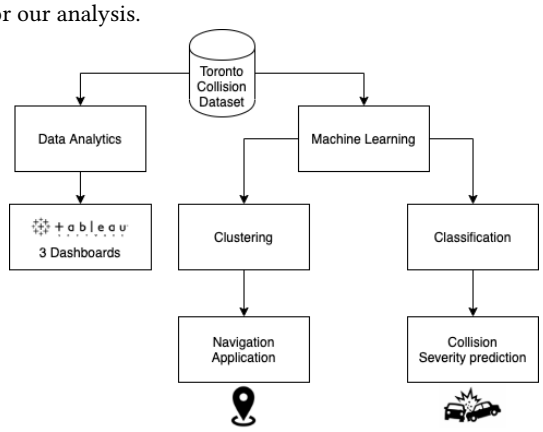

Figure 1: Data Analytics Architecture Diagram

Our methodology is separated into 3 major chunks as discussed below:

4.1 ML to derive accident prone neighbourhoods and fatal accidents

4.1.1 Data Cleaning and Prepossessing. We clean the data in the following ways:

- Disregarding columns with high percentage of missing values: The KSI Clean data set had certain columns that had many missing values, which have no significance in prediction. Therefore, the columns with more than 80% missing values were removed. Those columns were "Pedestrian Crash Type", "Pedestrian Action", "Condition of Pedestrian",

"Cyclist Crash Type", "Cyclist Action", "Cyclist Condition", "Distance and direction of the accident".

- Equalizing data : In this section we are replacing the entries in the of Accident Class column from "Property damage" to Non-Fatal. The KSI dataset has a few entries which are reported only as property damage in the accident and we have replaced the entries in column Accident class (which had three values to namely Non-Fatal injury, Property damage only and Fatal) from property damage to Non-Fatal. Since we are trying to predict Fatal/Non-Fatal Accidents, this conversion would help us categorize all elements into 2 distinct values.
- One-hot encoding for categorical data: To better compute the loss functions, we used one-hot encoding for categorical data.

4.1.2 Feature Engineering: Featuring Engineering in data set is important to find top features that help in the prediction of fatal/non fatal accidents through machine learning algorithm. We aim to use following techniques:

- Finding suitable Target column: The column which would form the best target is the one which demonstrates the fatality. 3 columns demonstrated the fatality:"ACCLASS" , "INJURY" and "FATAL". "INJURY" could not be used as a target column as it had many missing values(12%). On analysis we found that ACCLASS and FATAL were equal i.e. they showed the same pattern. Therefore any one of them could be used as Target class, however we used "FATAL" as it had binary values.
- As demonstrated in **Figure 2**: Finding most important features that contribute to fatal injuries: Given the target column we found out the most important features that led to fatal accidents. For this task we removed the features that were redundant or were of no use Accident Number, Location Coordinates, Minutes etc. We applied Extra Trees Classifier Model from Scikit Learn to find the top 10 features for the target class "FATAL". This Extra Trees Classifier Model implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

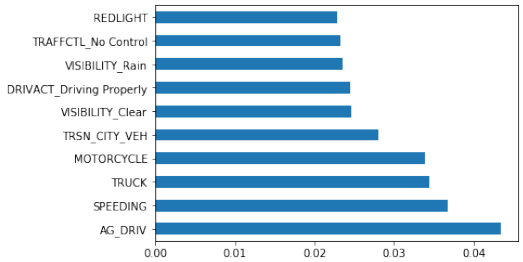

Figure 2: Top 10 Features from Extra Tree Classifier

4.1.3 Clustering to find Fatal accident prone neighbourhoods.

- Elbow Method to decide optimal number of clusters: The idea of the elbow method is to run k-means clustering on the data set for a range of values of k , and for each value of k calculate the sum of squared errors (SSE). The value that has the small k and a sufficiently small SSE is chosen. In our case the optimal value was 4.

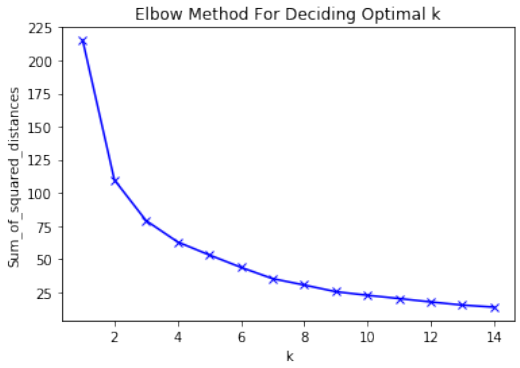

Figure 3: Elbow Method to select optimal k

- Clustering Methods: Two types of clustering methods were applied to find out accident prone neighbourhood:

  K-Means Clustering: The K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids. We used the top 10 features that we got from Extra Trees Classifier Model to cluster the data on the basis of District and Hood_Names. Scikit-learn was used to implement K means clustering.

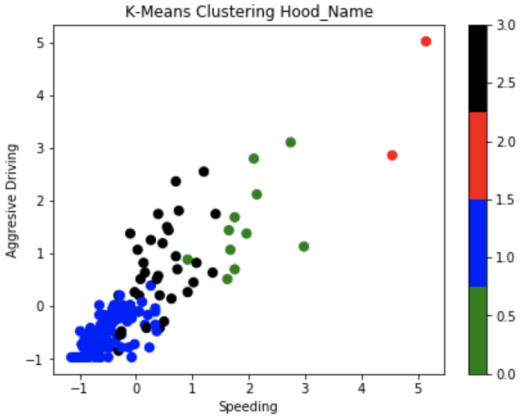

Figure 4: Clusters in KMeans Clustering

Agglomerative Clustering: The agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. The algorithm starts by treating each object as a singleton cluster. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects. The result is a tree-based representation of the objects, named dendrogram. Agglomerative clustering works in a "bottom-up" manner, i.e. each object is initially considered as a single-element cluster (leaf). At each step of the algorithm, the two clusters that

are the most similar are combined into a new bigger cluster (nodes).

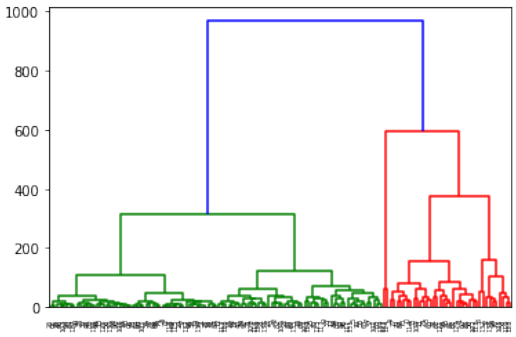

Figure 5: Dendrogram for Agglomerative Clustering

The clusters we got after Agglomerative Clustering are shown in Figure 6. We can see that we get similar kinds of clusters that we got from KMeans clustering algorithm i.e.the topmost neighbourhoods remain the same in both cases.

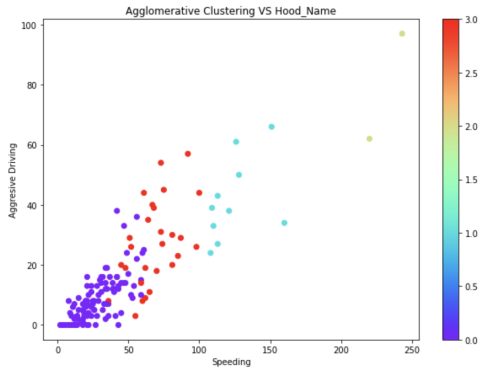

Figure 6: Clusters in Agglomerative Clustering

From both the clustering methods the neighbourhoods with maximum fatal accidents were 'Waterfront Communities-The Island (77)' and 'West Humber-Clairville (1)'. This demonstrates the correctness of the method.

4.1.4 Machine Learning Models for prediction: Sklearn library was used to apply five models on the data namely Random forest, Decision trees, K nearest neighbours and Logistic Regression . The data consisted of the independent features that comprised of important features that we got from Extra Tree classifier model.

- KNN classifer :KNN with 6 nearest neighbours was used in our case and the accuracy we got was 0.898
- Decision Tree classifier : Decision tree of depth 8 was used and entropy was used as the information criteria. The accuracy score we got was 0.916
- Random Forest Classifier : The number of estimators used was 100. The accuracy score was 0.931
- Logistic Regression : The accuracy with logistic regression algorithm was 0.912.
- Neural Network: This model consisted of an input layer, two fully connected hidden layers with relu activation and an output layer. The optimizer used for Adam. The accuracy with logistic regression algorithm was 0.934.

4.2 Application to suggest alternate routes

: The prime use-case of classification algorithms discussed above is to be used in the navigation application which we call $NavigationSAFE^{TM}$ and is publicly available on Github[10]. Perhaps the highlight of the project, this navigation application is developed on Angular 8 and uses several Javascript APIs by Here Maps [11, 19] for re-routing. It is a browser enabled web application and uses the A* algorithm[6, 8] behind the scenes to route and re-route, whose implementation is provided also by HereMaps. The A* algorithm to take into account the accident based neighbourhood by increasing the weight of the routes passing through the accident heavy areas. This approach is discussed in detail in the subsequent sections.

4.2.1 The A Algorithm.* : Implementation of Navigation on Here Maps[19] is implemented through A* Algorithm, which is a rerouting algorithm and is one of the most popular technique used in graph traversals. One of the basic functionality of A* algorithm is to Avoid Areas, which is easily achievable by increasing the weights on the edges of the graph which pass through the collision prone areas.
Here Maps [19] provides a very convenient API[13] to avoid the specific areas, which offers to avoid a pair of latitude and longitude, which in our case is the Accident Frequent Neighbourhoods. Our understanding of this functionality is - when we call the Avoid Areas API with a particular area's coordinates i.e. Latitude and Longitude pair, it would increase the weight on the Edges of the graph of that particular area and eventually avoid that area.
Therefore in order to avoid all collision prone areas, we called the Avoid Areas API several times through here maps for each neighbourhood and then integrated the results with our application.

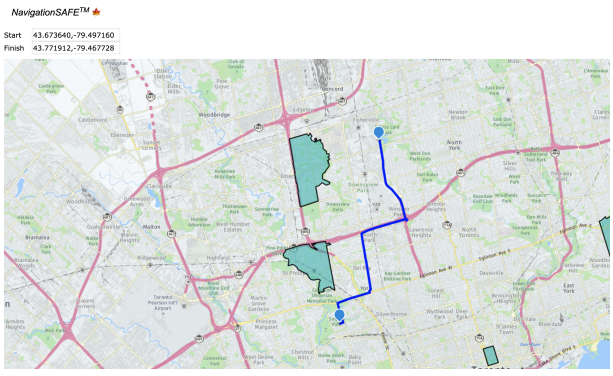

Figure 7: Navigation directions to a given place avoiding the green patches representing collision prone neighbourhoods.

To conclude, for the areas which are accident prone, the edge weight in the graph of A* is deliberately made high and therefore the path with high accidents are not suggested. We were successfully able

to build a this navigation application and presented it to the cohort as well [19].

4.2.2 The Navigation Application: As demonstrated in the figure 7, we have developed the web application - NavigationSAFE. There are 2 main functions of this application in addition to a typical Navigation application. First, this application highlights all the accident prone areas which is demonstrated by the green patches as shown in the figure. Second, all the navigation routes suggested are typically avoiding the Accident Prone areas. In case there is no path which exists apart from the accident prone areas, our application will demonstrate the risk by overlapping Navigation blue lines with the green patches as a warning. Considering a worst case scenario for our application as demonstrated in Figure 8, where there is no other route available apart from the Collision Prone Hoods, then the application will issue a warning to the user as shown below. As we can see that the green patches highlighting risky areas are overlapping with the Navigation Directions which is a basic warning to the users.

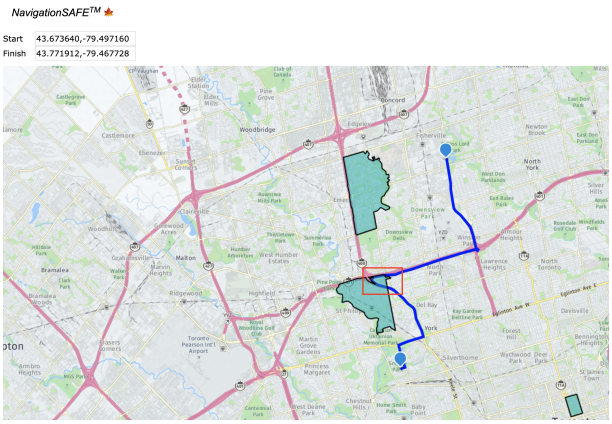

Figure 8: Worst case scenario of Navigation, when there is no other route apart from accident prone area.

4.3 Visualization

Data visualization makes big data easier to understand and detect underlying patterns. Primarily, we used Tableau Software[2] to visualize the various aspects of our dataset. One of the main motivation to use Tableau was that several kinds of Network Visualizations are readily available in Tableau, and given that our data-set is Geo-spatial, Tableau makes a perfect fit. We have used an abundance of charts such as– Line Chart to show trends over the years, Bar Charts for categorical data and Map chart to show geographically the positions of neighborhoods as clusters. In addition, other advanced features of Tableau have been utilized to increase user interaction such as: Providing check-boxes in filter to view the districts individually or together for comparison.Clickable hyperlinks/URLs which links to the real world articles that supports our analysis. Dynamic map chart where user can zoom in or zoom out to see better view of their queries.

Our Tableau Project delivers 3 Dashboards:
– Dashboard 1 illustrates Fatal Accidents v/s Factors
– Dashboard 2 shows the overall collision stats.
– Dashboard 3 demonstrates the time trend of accide

Below is the detailed description of each dashboard with accompanying figures.

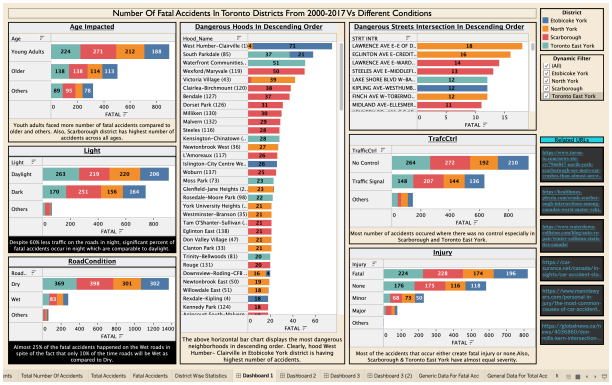

Figure 9: Dashboard 1: Showing Fatal Acc. Vs Factors

The number of fatal accidents are plotted against different factors leading to accidents, using stacked bar charts in Figure 9. Interesting **stories** can be inferred from the above graphs regarding the trends during night/daylight and the effect of road conditions on accidents. For example, we derived that night time is not as safe to drive perhaps due to visibility conditions. Another story we determined was - although winters might not be as suitable for driving, the number of collisions are lesser. Also, the dashboard shows the top dangerous neighbourhoods and street intersections in descending order where a user can customize the view using given filter.

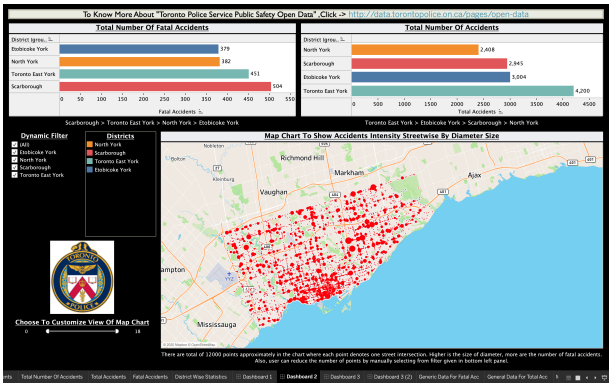

Figure 10: Dashboard 2: Showing Overall Statistics Of Acc.

Figure 10 demonstrates the overall statistics of the data-set. It shows that although Scarborough district is at number third in case of total number of accidents but it ranks first when it comes to fatal accidents count. Hence, it implies that whatever number of accidents are happening in Scarborough, most of them are fatal.

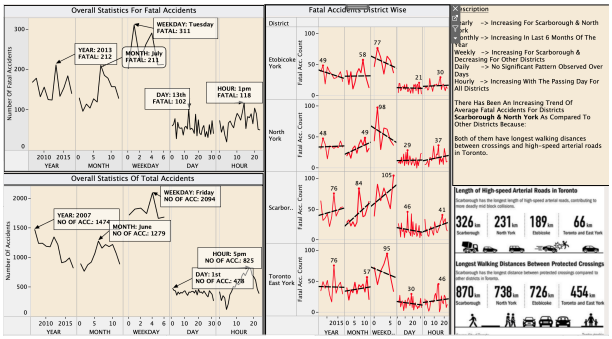

Figure 11: Dashboard 3: Showing Trend Of Acc. Over Time

Finally, we visualize the trends in accidents over the years from 2007-2017 using line charts in Figure 11. Left panel shows overall numbers whereas right panel shows the distinction based on the districts. Clearly,[1] districts Scarborough and North York lands on top because of the reasons that these districts have the longest length of high speed aerial roads and walking distances between protected crossings.

5 EVALUATION/RESULTS

5.1 Data Set

The KSI data set is a cleaned version of the traffic accident reports from the City of Toronto Police Open Data portal. It consists of accidents which occurred in Toronto from 2007 to 2017. The types of data range from numeric, boolean and categorical values with a total of 56 features for each data entry in KSI_CLEAN.csv and with a total of 12557 entries in the data set. The insights that we drew from data exploration are as follows.

Analysis of Features in Data set:

- 7 columns have more than 80% missing values
- ACCLASS(type of accidents) feature is same as Fatal feature.
- FATAL is the most suitable column for target variable for prediction.
- The data types are as follows 23 category, 2 float64, 24 int64 and the memory usage is 3Mb
- The top 10 features that we get for class "FATAL" are– Aggressive Driving, Speeding, Truck, Motorcycle, Transition Vehicle, Visibility, Driver's Act, Traffic Control and Redlight

	District	Ward	Hood
Min(T)	Etobicoke York	Beaches East York	NA
Max(T)	Toronto East York	Toronto Centre-Rosedale	Waterfront Communities-The Island
Min(F)	North York	NA	NA
Max(F)	Scarborough, East York	Scarborough Centre	Victoria Village

Figure 12: Min and Max counts of accidents(T) vs fatality at hood level

Trends Over the Data set:

- Fatal Injuries were highest in 2016.
- Out of Speeding, Redlight, Aggressive driving and Alcohol, Aggressive driving led to 62.9% of accidents.
- The number of Automobiles involved in the accidents outnumbered other vehicles like Truck, Motorcycle, Transition Vehicle and Emergency Vehicle. Automobiles were present in 81.3% of the accidents.
- Passengers and Pedestrians were major victims in the accidents. Passengers were victimized in 42.2% accidents whereas Pedestrians were victimized in 45.5% of accidents. Cyclists were the least hurt i.e. in 12.3% of accidents.
- 82.2% of Accidents resulted in Fatality while 17.2% resulted in Disability.

5.2 Final Results

5.2.1 Prediction Results. The accuracy scores of various predictive models(to predict fatal accidents) on the Data set is summarized in the Table 1.

Algorithms	LR	KNN	DT	RF	NN
Accuracy Scores	0.912	0.898	0.916	0.931	0.934

Table 1: Accuracy scores of Logistic Regression, K nearest neighbours, Decision trees, Random forest and Neural Network respectively

From the Table 1 we conclude that Neural Network model worked the best on our Data set.

5.2.2 Clustering Results. On performing clustering on various neighbourhoods based on important features we got top two accident prone neighbourhoods namely 'Waterfront Communities-The Island (77)' and 'West Humber-Clairville (1)'.

5.2.3 Navigation Application to Avoid Routes.

- The navigation application is built in Here Maps[19].
- Network model is employed to avoid the accident prone areas/neighbourhoods.
- HereMaps provide access to API which can help find tune their Map's Network Model to avoid certain areas.
- These areas could be parameterized from our analysis of features and clustering.

5.2.4 Visualization Results. The below table highlights the peak number of accidents that had occurred at different levels of time.

	MIN(T)	MAX(T)	MIN(F)	MAX(F)
Year	2017	2007	2012	2013
Month	Feb	June	Feb	July
Weekday	Sat	Fri	Mon	Tues
Day	Last	First	Third	Thirteenth
Hour	5am	7pm	5am	3pm

Table 2: Minimum and Maximum values defined for total number of accidents(T) and fatal accidents(F)

6 CONCLUSION

In this work we have applied the machine learning algorithms for prediction of Fatal accidents and clustering of accident prone neighbourhoods. We have also done sufficient data exploration to

understand the structure of data and clean the data accordingly. As proposed in our proposal we have done the Data Exploration and completed machine learning analysis and finished the application development. We have the coordinates of the accident prone neighbourhoods and we have successfully integrated this data in our application in order to avoid those areas and make navigation of Torontonians safe. Our machine learning model works with an accuracy of 93.34% and it is used to predict the severity of an accident. This analysis can help police and paramedics to prioritize accidents according to their severity. Any possible amendments in our current work are suggested in Future Work section.

7 FUTURE WORK

Our work can be improved in many ways, be it applying more complex data or expanding the application to various other cities or the world as a whole. The future work can also take into account the presence of street intersection and apply analytics/models on a more granular level i.e. street intersections to achieve better navigation. The application right now only considers the GTA region, it can be expanded to other areas as well but this is subject to the availability of dataset. The application can use more complex algorithms to analyse the data, in order to get better results. This can be achieved once a larger dataset is available. An important variant of our application can be, it's use as a personalized accident risk predictor, that will use historic data of an individual and predict their proneness to accident if they follow their current routing routine. This can be used by people worldwide to opt for safer travel routes.

REFERENCES

[1] Mike Adler. 2019. *Toronto's Vision Zero 2.0 plans.* https://www.toronto.com/news-story/9514533-will-lower-speed-limits-on-scarborough-roads-reduce-pedestrian-deaths-//

[2] Lyn Bartram Melanie Tory Alper Sarikaya, Michael Correll and Danyel Fisher. [n.d.]. *What Do We Talk About When We Talk About Dashboards?* https://research.tableau.com/sites/default/files/DashboardsConspiracy_final.pdf

[3] Tristan Glatar Antoine Hebert , Timothee Gu and Brigitte Jaumard. (2019). *High-Resolution Road Vehicle Collision Prediction for the City of Montreal.* https://arxiv.org/pdf/1905.08770.pdf

[4] Sakham Nagendra Babu and Jebamalar Tamilselvi. (2019). *Generating Road Accident Prediction Set with Road Accident Data Analysis Using Enhanced Expectation-Maximization Clustering Algorithm and Improved Association Rule Mining.* http://www.iieta.org/journals/jesa/paper/10.18280/jesa.520108

[5] Juan Li Chunjiao Dong, Chunfu Shao and Zhihua Xiong. 2018. *An Improved Deep Learning Model for Traffic Crash Prediction.* https://www.hindawi.com/journals/jat/2018/3869106/

[6] Liang Zhao et. al. 2008. *A* Algorithm for the time-dependent shortest path problem.* https://www.researchgate.net/publication/253129361_A_Algorithm_for_the_time-dependent_shortest_path_problem.

[7] Amir farrokh Iranitalaba and Aemal Khattakb. 2017. *Comparison of four statistical and machine learning methods for crash severity prediction.* https://www.sciencedirect.com/science/article/abs/pii/S0001457517302865/

[8] Adeel Javaid. 2008. *Understanding Dijkstra Algorithm.* https://www.researchgate.net/publication/273264449_Understanding_Dijkstra_Algorithm;DOI:10.2139/ssrn.2340905

[9] Kaggle.com. 2019. *US Accidents Data set.* https://www.kaggle.com/sobhanmoosavi/us-accidents

[10] Payal Goyal Karan Singh, Mahima Chaudhary. (2020). *NavigationSAFE.* https://github.com/singkara/NavigationSAFE

[11] Davies C. Klippel A, Hirtle S. 2010. *You-are-here maps: Creating spatial awareness through map-like representations. Spatial Cognition Computation.* 10(2\T1\textendash3):83\T1\textendash93.doi:10.1080/13875861003770625.

[12] Salah Taamneh Sharaf Alkheder. Madhar Taamneh. (2017). *Clustering-based classification of road traffic accidents using hierarchical clustering and artificial neural networks.* InternationalJournalofInjuryControlandSafetyPromotion,24:3,388-395,DOI:10.1080/17457300.2016.1224902

[13] Here Maps. [n.d.]. *Requesting a Route Avoiding an Area.* https://developer.here.com/documentation/routing/dev_guide/topics/example-route-avoiding-an-area.html

[14] Ajith Abraham Miao Chong and Marcin Paprzycki1. 2019. *Traffic Accident Analysis Using Machine Learning Paradigms.* http://www.informatica.si/index.php/informatica/article/download/21/15

[15] Mohammad Hossein Samavatian Srinivasan Parthasarathy Moosavi, Sobhan and Rajiv Ramnath. 2019. *Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.* https://arxiv.org/abs/1909.09638

[16] Mohammad Hossein Samavatian Srinivasan Parthasarathy Moosavi, Sobhan and Rajiv Ramnath. 2019. *A country wide traffic accident data set.* https://arxiv.org/abs/1906.05409

[17] Sobhan Moosavi. 2019. *Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.* https://arxiv.org/abs/1909.09638

[18] Sobhan Moosavi. 2019. *A Countrywide Traffic Accident Dataset.* https://arxiv.org/pdf/1906.05409.pdfs

[19] Nokia. 2000. *Here Maps.* https://www.here.com/

[20] Martin Theus. 2000. *Visualisation of Categorical Data.* https://pdfs.semanticscholar.org/98be/677f02b52b708e5bf69835493440b04c60c3.pdf

[21] Kirsten Vallmuur. 2015. *Machine learning approaches to analysing textual injury surveillance data: A systematic review.* https://doi.org/10.1016/j.aap.2015.03.018

[22] Jiang H Wahab L. (2019). *A comparative study on machine learning based algorithms for prediction of motorcycle crash severity.* PLoSONE14(4):e0214966.https://doi.org/10.1371/journal.pone.0214966

[23] Yu Yao. 2019. *Unsupervised Traffic Accident Detection in First-Person Videos.* https://arxiv.org/pdf/1903.00618.pdf

APPENDIX

User's Manual

3 Major Components of the Project:

(1) Kaggle Kernel demonstrating all data analytics and methodology:
-https://www.kaggle.com/mahimachaudhary/ksi-data-exploration-and-analysis

(2) The application illustrating the Navigation Application's link:
-https://github.com/singkara/NavigationSAFE

(3) Link To Visualization Showing Data Analysis:
-https://public.tableau.com/profile/payal.goyal#!/vizhome/DataAnalyticsOnCityOfTorontoPoliceAccidentsData/Story?publish=yes