RECOGNIZING PATTERN BASED MANEUVERS OF TRAFFIC ACCIDENTS IN TORONTO

By

Waleed Saleem

A Major Research Project presented to

Ryerson University

In partial fulfillment of the requirements for the degree of

Master of Science in the

Program of

Data Science and Analytics

Toronto, Ontario, Canada, 2020

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PROJECT (MRP)

I hereby declare that I am the sole author of this Major Research Paper. This is a true copy of the MRP, including any required final revisions.

I authorize Ryerson University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.
Waleed Saleem

RECOGNIZING PATTERN BASED MANEUVERS OF TRAFFIC ACCIDENTS IN TORONTO

Waleed Saleem

Master of Science 2020

Data Science and Analytics

Ryerson University

ABSTRACT

This paper discusses the utilization of machine learning techniques to detect patterns of traffic accidents in Toronto. The primary and most fundamental purpose of carrying out this research is to identify and analyze the driving patterns and behaviors in Canada Toronto, as the main sample. The aim of this project paper is to examine the factors that contribute to road accidents in the country; and to evaluate statistically the effect of certain driver's personal characteristics on road accidents. This paper has proposed a model that trained multi class classification dataset through machine learning algorithms. This paper has used 4 classifiers that are used for supervised learning. Each classifier is implemented on the dataset in order to find the accuracy of the model. The classifiers are also compared to find out the best option for the given dataset. The accuracy of the algorithm showed more than 95% on the dataset which indicates the algorithm was a perfect fit for the given dataset.

Keywords: Traffic accident patterns, machine learning algorithm, detecting traffic accidents, Toronto city traffic accident patterns, multi class classification dataset.

ACKNOWLEDGEMENTS

I would first like to thank my supervisor Professor Dr. Saman Hassanzadeh Amin from the Mechanical & Industrial Engineering Department, at Ryerson University, for this Major Research Project. He has been a great support throughout the term to guide and direct my research and provide valuable feedback. For his magnificent support and assistance, help me to making this project a reality.

I would also like to acknowledge Dr. Ceni Babaoglu of the dept. of Mechanical and Industrial Engineering at Ryerson University as the second reader of this research project. I am gratefully indebted to her valuable lectures to prepare for Major research project that help me to build this step by step.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Driving is considered to be one of those activities in Canada in which the people take pride in and in an oil-driven economy and can see all sorts of cars starting from cheapest to the more luxury brands on the road. Driving is a necessity here where every outing, the household activity requires a car. The number of cars has increased significantly over the past decade.

The automobile is available for almost all households in Canada in recent years due to the increase in per capita income. This factor has led to an increase in the usage of cars, resulting in more accidents. Road traffic accidents occur as a result of three factors, namely: road users, road environment, and vehicles.

The rapid economic growth that Canada experienced in the last 40 years led to a remarkable increase in motorization and road network construction. Motor vehicles are considered as the primary method of transportation in Canada. Consequently, Motor Vehicle Accidents (MVA's) and accident-related injuries and deaths have become a major health hazard in the country.

Chaudhary et al. 2011**,** Bener and Bener 2007 found that Distracted driving behaviors (DDBs) such as talking on the cell phone, texting messages, eating, drinking, adjusting the car's radio, and interacting with passengers in the car while driving may create a potential traffic safety concern. Despite the potential harm associated with DDBs, as far as the literature review conducted, there were no studies found to determine the prevalence and consequences of these behaviors in Canada.

1.1 Purpose of Conducting this Research

Traffic accidents are a major problem in developing countries. Canada is no exception to this worldwide dilemma. In Canada, a high incidence of road accidents has been recorded in the last several years due to various factors. The primary and most fundamental purpose of carrying out this research is to identify and analyze the driving patterns and behaviors in Canada Toronto, as the main sample. Research on this issue is the time of need.

The aim of the proposed project paper is to examine the factors that contribute to road accidents in the country; and to evaluate statistically the effect of certain driver’s personal characteristics on road accidents. Issa, (2016) found that young drivers (less than 30 years) are involved in around 60% of the accidents and more than 80% of the accidents related to human factors. By identifying the demographics, the skill set, the driving patterns and behaviors of the local drivers and the road environment and other factors to which they are exposed to while driving, this proposed research project aims to pinpoint and identify the main causes of the road accidents that take place in the country which unfortunately lead to injuries and deaths. Consequently, findings based on the research are given and the ways of how these unwanted events can be tackled and reduced using conventional methods as well as modern and technological solutions such as eco-driving, future cars.

1.2 Surrounding Environment

Over speeding is common, especially by the owners of high-end sports cars and luxury vehicles. ‘Stop’ and ‘Yield’ signs are often ignored. Drivers pass at any time from any direction, and turn signals are rarely used. Passing on blind curves from both directions is also common. Pedestrians and livestock in the road can be a hazard; in some cases, shepherds have bedded their sheep near major highways at night, resulting in collisions between vehicles and livestock that stray onto the road. Motorists should drive defensively, use extreme caution, and wear seatbelts at all times.

Speeding, disobeying traffic signals, sudden lane change, and driver errors are frequent causes of road traffic collisions. The environment associated with driving is of several dimensions and can include infrastructures such as roads network structure and road conditions; climate and weather. In addition, poor road conditions are blamed for the high level of accidents in some cities.

1.3 Accidents

Canadian officials are well aware of the need to improve driving standards. The causes of traffic accidents can be divided into following three main reasons high speed, impaired driving, and distraction. According to the statistics revealed in 2013 by ICBC 28% of accidents occurred due to high speeding, 23% of accidents happened because of drunk driving & 29% of the accidents caused due to driver distraction**.** Klein Lawyers LLP, 2020 website has mentioned, Canada is known for its harsh climate where people are not used to hot weather. In summers, the temperatures soar to nearly a staggering 35-40 degrees. As summer temperatures rise, so do people’s tempers. For decades, researchers have observed a correlation between hot weather and increases in violent, aggressive behavior. Similarly, over speeding in snowy and icy weather is risky and can cause tragic results. Figure 1 illustrates a road accident in bad weather conditions.

Figure 1: Road accidents in bad weather conditions.

In one classic study from 1984, APS Fellow Douglas Kenrick and Steven MacFarlane observed that drivers get more aggressive on hotter days. The team had a research confederate purposely irritate other drivers by remaining stopped all the way through a 12-second green light at an intersection. Aggressive behavior was calculated by tracking the total amount of time that other drivers spent honking.

“Results indicated a direct linear increase in horn honking with increasing temperature. Stronger results were obtained by examining only those subjects who had their windows rolled down (and presumably did not have air conditioners operating),” Kenrick and MacFarlane report **[3]**. Al Turki (2014) In addition to honking, the co-relational link between high temperatures and aggressive behavior, called the *heat hypothesis*, has been shown in behaviors ranging from murder and assault to car thefts.

Due to different weather and road conditions, 30 fatal car accidents are recorded in 2013 with an average of 46 a year over the most recent 5-year span. Furthermore, texting while driving is another main cause of traffic accidents. Alongside over speeding, weather conditions, and driver distraction; unfortunately, many drivers and passengers are not obeying simple traffic rules such as wearing seat belts can also cause serious traffic accidents. According to Canadian Motor Vehicle Traffic Collision Statistics, 2018 From 2014 to 2018 the drivers who were not wearing seat belts; the fatality rate has been increased from 26.2% to 29.4%. Moreover, passengers who are not using seat belts are also at risk of getting serious injuries in accidents. According to Canadian Motor Vehicle Traffic Collision Statistics, 2018 from 2017 to 2018 alone the rate of passenger serious injuries increased from 16.9 % to 18.4 % for not wearing seat belts. It can be noted the traffic rules should be revised not only for drivers but also for passengers riding the vehicle. Figure 2 shows a serious road accident caused by consuming alcohol and over speeding.

Figure 2: Serious road accident due to consuming alcohol and speed

According to Canadian Motor Vehicle Traffic Collision Statistics, 2018 the number of road deaths per year has risen during recent years to a rate equal to an average of 5 deaths per day. According to WHO the most common human factors contributing towards traffic accidents include speeding (in 65% of accidents), driver error (in 80% of accidents), violation of traffic signals at intersections (in 50% of accidents), and illegal U-turns. Other causes are related to vehicles, the road, and the environment (e.g., road layout, which contributes to 20% of accidents). Excessive speeding was the most common cause reported in all recent and past studies. According to Al-Ibrahim et al. (2010), driver error was identified as the main contributing factor in about two-thirds of all RTAs, mainly characterized as reckless driving and excessive speeding.

1.4 Problem Statement

The problem of accidents, over speeding, road rage has always been known widely in Canada, ***affects*** the government, citizens, residents as public transport is not yet introduced, and the driving standards are not up to mark until this time, everyone is handicapped to a vehicle for moving from one place to another. ***The impact of which is an*** increase in accidents, road rage must be decreased and make the roads of Canada safe and secure for everyone to travel on.

A successful solution would be to come out with a dataset using a Machine Learning Algorithm like Logistics Regression, which can be further used by the government of Canada. It will excel in the government's part and citizens so that the relevant changes could be made to overcome the fear people have.

1.5 Objectives

To improve the standard of driving in Canada by profiling and analyzing the pattern of driving using Machine Learning Algorithms

1.6 Scope and Limitations of the Study

The research tries to take into account the events across the entire country but has its limitations. Underreporting and disorganized data are considered to be the major limitations of RTA data. Such as this research will be using data that is being generated by the Toronto police. Although this data is not considered enough, it can certainly give a keen insight into the current driving situation prevailing in the country.

1.7 Drawbacks of the Current State

The current system has many loopholes and gaps that can be rectified in order to smoothen the traffic system. This will not only make it easier for drivers to ride a vehicle but also decrease the number of accidents to a great extent. As a result, the number of injuries and deaths can also be minimized. Some of the drawbacks and loopholes are mentioned next.

One of the most significant drawbacks of the current transportation system is inadequate human behavior and judgment. Focusing on how to eliminate driver distractions and decrease the number of decision points a motorist needs to make will help a more stable variable to design roads for.

Secondly, a tightening of the driver's license requirements shall be considered. By doing so, the chances to hand over licenses to immature and mediocre drivers will be lessened. This can be done by implementing strict measures such as well thought out driving tests carried out with transparency, increased age limit, restrictions, or suspensions of licenses in case of repetitive traffic violations, etc.

Thirdly, spreading awareness about following driving rules, the importance of safety and how it can be achieved, and the consequences of violating traffic laws is extremely crucial. The government should take the necessary steps to embed these important topics into its citizens eligible for driving. This can be started as early as possible like from college-level or by carrying out seminars/workshops.

2. LITERATURE REVIEW

This section focuses on the review of the literature based on the same or similar research topics. It includes a comprehensive review of previously existing studies on traffic accident prediction, driver behavior profiling & analysis of the factors responsible for traffic accidents.

2.1 Vision-based highway traffic accident detection

Ni et al. (2019) has generated a system where they have acquired the data from live traffic surveillance video in order to detect the traffic status and the vehicle orbit via a machine learning algorithm. Through the machine learning background extraction algorithm, they have extracted the highway lane boundary lines, the left and right lane boundary lines, emergency lane marking lines, and also the fence on both sides of the road. In order to detect the vehicles on the road, they have trained the vehicle detection model using a convolutional neural network. A relationship between the vehicle orbit and the boundary lines has been studied in order to detect the rate of possible traffic accidents also to reduce the false alarm rate. To obtain the vehicle orbit they have identified vehicles in continuous frames.

In order to avoid the abnormal parking to detect as traffic accidents, they have set 4 criteria when the incident will be called a traffic accident. The paper mentions if the number of the vehicle is less than 2 and located in an emergency lane for a continuous period of 2 seconds it is abnormal parking otherwise is considered to be an accident. Although due to bad weather or any unpredictable situation any 2 vehicles can be in an emergency line. So, in that case, their traffic detection system will detect it as a traffic accident which is not true. The paper also mentions if a vehicle is in between the traffic lanes and median strip, if 2 vehicles are in a non-emergency line and didn't change the position in a continuous period of 2 seconds & if the vehicle is in highway boundary lanes is considered to be a traffic accident. However, the paper did not mention any criteria for a vehicle speeding.

Most accidents are occurred due to crossing speed limit or over speeding, frequent lane changing and pushing vehicles out of the road also causes accidents. If a vehicle has been pushed away and it happened to be in the emergency line it will consider it a traffic accident because it's in the emergency line. The paper has tested their model based on traffic monitoring recording of video in the Qinglin highway of Shandong Province from August 2017 to November 2018. The data they have used to test the model is really insufficient for concluding the system to be effective. Also, they have studied the model only on past data.

So, it is unsure whether the model will be effective in other traffic situations. They should have studied the model on a large data set and in real-time traffic situations. Also, they have studied their model only on one highway, They should have studied the model in other highways and large data set in order to test the model's effectiveness. Traffic accidents occur everywhere not only in highways. It is really important to study traffic accidents on other roads as well. They should have identified the busy roads and the roads where the accident is common. Based on those they should have tested the algorithm. Because the different area has a different situation and different driving behavior. The model could be effective in a particular area but not widely. In this paper, they have mentioned that the experimental video covers all types of traffic conditions including light, traffic flow, and weather conditions. As it was mentioned before the paper only studied a particular highway data, so it did not study all types of traffic condition but only the particular area's condition.

2.2 Driver behavior profiling: An investigation with different smartphone sensors and MLA

One of the most important factors for traffic accidents is the driver's behavior. Researchers believe that studying driver's behavior can combat traffic accidents. Many researchers have studied driver's behavior through machine learning techniques where these systems can predict traffic accident probability depending on the driver's behavior in a particular situation. According to Carvalho et al. (2017), MIT researchers published a system where the driver's behavior studied through 4 MLA's. They have used Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest (RF), and Bayesian Network (BN) in the study. Driving event types were examined through multi-level supervised learning. The main aspect of the study was to find out the best machine learning algorithm along with motion sensors and the number of frames in the sliding window in order to get precise driving event types. In this work a quantitative assessment of the exhibitions of 4 MLAs (BN, MLP, RF, and SVM) with various designs connected in the recognition of 7 driving occasion composes utilizing information gathered from 4 Android cell phone sensors accelerometer, linear acceleration, magnetometer, and gyroscope) were introduced. 69 tests of these occasions compose were gathered in a certifiable explore different avenues regarding 2 drivers. The execution was likewise looked at while applying changed sliding time window sizes. 15 executions with various irregular seeds of 3865 assessment gatherings of the frame EA ={1: sensor, 2: sensor axis(es), 3: MLA, 4: MLA configuration, 5: number of frames in sliding window} were completed. Therefore, the top 5 performing congregations for each driving occasion write were distinguished. With regards to the examination, these outcomes demonstrate that (I) greater window sizes perform better; (ii) the spinner and the accelerometer are the best sensors to recognize our driving occasions; (iii) as general manager, utilizing all sensor tomahawks perform superior to anything utilizing a solitary one, aside from forceful left turns occasions; (iv) RF is by a wide margin the best performing MLA, trailed by MLP; and (v) the execution of the main 35 blends is both agreeable and proportional, shifting from 0.980 to 0.999 mean zone under the ROC curve (AUC) values. This study is helpful for collecting data regarding the driver's behavior in order to detect traffic accidents and minimize the false alarm rate.

2.3 Rich Monitoring of Road and Traffic Conditions using Mobile Smartphones

Many traffic accidents occur due to the condition of the road and weather. Mohan et al. **(**2008), proposed a system which is a Windows Mobile smartphone application to monitor road and traffic conditions**.** is a smartphone-based system designed to detect a few variables regarding vehicle and roads such as bumps, speed breakers, horns and go and stop traffic and for this purpose, it uses a smart phone's accelerometer, microphone, global system of mobile (GSM) communications and GPS. What happened is that on a centralized server, Nericell aggregated sensed data from the smartphones that were connected and used here. Mohan et al. envisioned the system being used to annotate existing traffic maps with data and information that were necessary such as turbulent traffic and other updates. Nericell strives to use the sensors in a power-efficient manner. Only the accelerometer is sampled continuously with the GSM radio kept active, which is needed for communication anyway. Moreover, in order to conserve energy, the system comprising of the microphone and GPS would activate only when required. Before the data is sent for aggregation to the server, it is thoroughly filtered and locally processed.

2.4 Automatic Traffic Accident Detection and Notification with Smartphones

To automatically detect traffic accidents and for notification with smartphones, the WreckWatch system, developed by Dougherty et al. (2011)**,** uses smartphones and sensors to identify accidents as compared to the original manufacturer's system which uses readings from the vehicle's electronic control unit (ECU). White et al. disagree because they think it is not every time possible for the user to connect to the ECU on every road trip. Moreover, one area of concern is that not every vehicle is equipped with ECUs, so an accident detecting system which is independent of ECU is of benefit in the end. WreckWatch uses a soft real-time (close to realtime) approach sampling the accelerometer, microphone, and GPS of a smartphone. An accident is detected by threshold filtering of the sensor readings. So, the data that is recorded during and after the accident is sent to the server through the GSM. As a result, relevant and important information from the database server can be sent to the authorities about the accident. False positives (FPs) are more likely to occur with a system using only smartphone sensor data. So, there are chances that a mere drop of the phone or an unintentional knock/slap on the phone may be considered an accident. Therefore, context information obtained from filters must be used to prevent FPs. First, the determined acceleration is filtered by ignoring any values below 4g. Second, a user is assumed to be in a vehicle if they are moving faster than 25 km/h. The smart phone's GPS is used to determine the speed of the user. So, if the user is traveling at speeds faster than 25km/h, only then the readings are evaluated. This, as a result, decreases energy consumption and also prevents any incorrect false alerts in case the phone is dropped.

2.5 Mobile Phone-Based Drunk Driving Detection

Bai et al. (2010**)** developed such a system that can notice driving under influence by only using the accelerometer of the smartphone. Their inspiration for outlining such a framework is the way that more often than not alcoholic driving goes unnoticed by the experts, which puts numerous individuals' security in danger. They compressed alcoholic driving-related practices from an examination done by the United States National Highway Traffic Safety Administration. There are two classes of behavioral signs which compare to a high likelihood of alcoholic driving. The primary classification is identified with path situating issues, for example, floating and swerving. The second classification is identified with speed control issues, for example, sudden increasing speed or sporadic braking. Both these classes of signals can be recognized by

utilizing an accelerometer to delineate prompts into the parallel and longitudinal increasing velocities of a vehicle. The system is designed with four software components: a monitoring daemon module, calibration module, pattern matching module, and an alert module.

2.6 Accident Detection Using Convolutional Neural Networks

Ghosh et al. (2019) has emphasized that due to a lack of timely help at the accident spot often claim many lives in an accident. The purpose of the study was to detect accidents from video footage taken by a camera and also to send out help in a timely manner through the system at the same time. They have used advanced deep learning algorithms using Convolutional Neural Networks for analyzing the frames collect from video footage by a camera. A portable and remote computer Raspberry Pi 3 B+ Model was set up to act as a camera. In order to detect accidents, they have pre-trained an Inception v3 model on two different sets of images and sequences of video frames. As a result, it was able to detect an accident frame by up to 98.5% accuracy. About 10000 severe accidents and 10000 non-accident frames were used to train the model. Tensor Flow, OpenCV, and Keras have been used in order to implement the model in Raspberry Pi. The model functions on a given video frame by analyzing each frame of the video to detect whether it is an accident frame or not. If it matches or exceeds the threshold of 60% it considers the frame as an accident frame and sends a message to nearest hospital and police station with the timestamp, location, the frame where the accident detected for further analysis and an emergency light also lights up by using a GSM module setup.

In this paper, they have used CNNs to model spatial data such as image classification, object detection, etc. LSTMs are used for sequential data and prediction. CNN's and LSTMs architecture has been used to process a sequence of images and videos in order to make a precise prediction. This architecture is inspired by Inception v3 and used to train images. In order to extract meaningful information, the Convolutional layer preserved the relationship between the pixels of the image by learning features through small squares of input data. Furthermore, the pooling layer is used to cleansing the data and extracting the relevant information. Pooling layer functions based on 3 following layers Average pooling, Max pooling, and Sum pooling.

Their methodology is to run every frame of the video through Inception v3 and save the output to the final layer of the pool of the network. Then they passed the vector of features to RNN. Then they have trained the RNN saving the disk of sampled frames such that it can pass from CNN and train a new network architecture. Each frame looped in chronological order added to a queue of size N and the first frame pop off that was added. In the experiment, they have a queue size of 5 in order for Pi camera to record every 5 seconds video at 5 frames per second. The layers were trained through Image Data Generator in Keras which is used for image augmentation. It artificially expands the dataset so that the model can learn from as much data as it could. The image augmentation features include zoom, pre-processing function, shear, etc.

2.7 Real-time accident detection coping with imbalanced data

Derrible et al. (2019) has compared two popular machine learning models Support Vector Machine (SVM) and Probabilistic Neural Network (PNN) in order to detect real-time traffic accidents on Eisenhower express in Chicago. In this research 7 models were trained and tested on both learning models on traffic condition data from 1 to 7 min after the accident occurrence. Different traffic conditions were analyzed in this paper including weather conditions, accidents, and loop detector data. To up cope with imbalanced data they have used Synthetic Minority Oversampling Technique (SMOTE). This paper was first to apply such a technique for an accident detection system, The purpose of the research was to compare the efficiency between two supervised machine learning methods for accident detection on the urban expressway and to apply SMOTE to deal with imbalanced data. Furthermore, they have determined to find the optimal number of minutes between 1 to 7 to detect accidents more accurately. They have mentioned in this paper that the SVM model shows a relatively less false alarm rate in incident detection. It is also mentioned SVM is best to cope with small data samples however it cannot deal with big imbalanced datasets. The study has collected its data from the Illinois Department of Transportation (IDOT). They have selected the Eisenhower expressway which connects the Chicago loop to the north side of Chicago. 24 loop detectors have been located on this stretch. Then they have divided the section into 23 sections in order to get an average of 1km. Two loop detectors were located at the beginning and end of each section to capture the traffic conditions upstream and downstream of the accident location. This paper has studied loop detector, accident, and weather condition data from June 2017 to December 2017. A preliminary study was performed on 24 non-accident cases are selected from each of the 23 from June to December 2017. Then the dataset of 85,182 non-accident cases and 32 accident cases prepared after the elimination of erroneous non-accident cases. Since this is an imbalanced data, new synthetic data generated from 32 accident cases and the number of accident cases increased to similar to non-accident cases.

2.8 Hetero-ConvLSTM: A Deep Learning Approach to Traffic Accident Prediction on Heterogeneous Spatio-Temporal Data

Yang et al. (2018) has studied the traffic accident prediction problem using the Convolutional Long Short-Term Memory (ConvLSTM) neural network model. They have studied the big datasets of the state of Iowa across 8 years in order to extract dynamic traffic conditions such as weather, environment, road condition, and traffic volume. They performed an extensive experiment on 8-year data over the entire state of Iowa where their system shows successful predictions on traffic accidents and improves the prediction over the baseline approach. The paper proposed a framework where several ideas were implemented such as incorporating spatial graph features and spatial model ensemble. Hetero-ConvLSTM, deep learning approach to predict big heterogeneous Spatio-temporal data. They extracted urban and environmental features such as traffic volume, road condition, rainfall, temperature, and they have collected satellite images and map-matched with grid cells. The model was trained in such a way that it can predict the traffic accidents that will occur on each grid cell in the future time slot by the number of accidents as well as the other urban and environmental features at each location. They have incorporated spatial features into the Convolutional Long Short-Term Memory neural network model (ConvLSTM) to capture the temporal trends and heterogeneity of the data. The framework has learned different models and different regions of the study area and the results are assembled in order to generate the final prediction. Comprehensive experiments were performed on various parameter settings, feature sets and baseline approaches. The results of the study show that for rural areas weather and spatial graph features are important for traffic accident detection and road condition, traffic volume, and holiday/weekday information are playing an important role when it comes to urban areas. They constructed 31 features from their dataset and then they have grouped them into 7 categories are as follows road network, road condition, satellite image, rainfall, weather, traffic volume, calendar features, spatial graph features then converted into 3-dimensional (64×128×1) tensor.

2.9 A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction

Hu et al. (2018) has studied big traffic accident data and proposed a deep learning model for traffic accidents than can predict risk based on a recurrent neural network. The paper has analyzed the frequency of spatial-temporal patterns and spatiotemporal correlation was presented of traffic accidents. Based on the patterns found in the analysis they have proposed a deep learning model that was accurate enough to predict the risks of traffic accidents. The model was trained to learn the deep connections between traffic accidents and spatial-temporal patterns. For studying the pattern of risk of traffic accidents they have analyzed the records of Beijing in 2016 and 2017. The record contains the time and GPS coordinate to the accident event. They have preprocessed the data by discretization in order to get the proper data structure to build the machine learning model and analyze the pattern of accidents. After the preprocessing of the data, they obtained a matrix whose element is the count of traffic accidents happened within the region and the time slot. The study shows that the traffic accident frequency is highly related to the geographical position of a region and the highest traffic accident region usually lies in the major commercial and business areas. The paper proposed model is temporal related, the data was arranged chronologically. The traffic data of Beijing Jan 2016 to April 2017 was chosen to train the model, the data from April 2017 to August 2017 was chosen to test the trained model. To validate the data, they have chosen the last 20% of the data. They build the architecture of TARPML upon Keras a python deep learning library, RMSProp was chosen as an optimizer. They compared the Root Mean Square Error (RMSE) of the TARPML method with different input sequence length. The result of the study shows strong periodical temporal patterns and regional spatial correlation. The study only utilized traffic accident data without considering other related data such as traffic flow. Human mobility, road characteristics, and special events can be significant to a traffic accident. Since the prediction results are coarse-grained, they are not eligible for road level accident risk prediction.

3. DESCRIPTIVE ANALYTICS | EXPLORATORY DATA ANALYSIS

3.1 Data Source and data file

This paper has used the dataset collected from Toronto police service, Public Safety Data Portal, Open Government Licence – Ontario, (2019) an open-source & open-licensed website**.** The dataset includes all the patterns & traffic accident-related factors responsible for the traffic accident. Furthermore, the dataset contains killing and seriously injured data from the year 2006-2019 in Toronto. Moreover, the dataset has 60 columns in total and has been processed further to cleansing the data.

3.2 Text Analysis

Figure 3 shows the classes with respect to each feature. The X-axis shows the features and Y-axis shows the number of classes. In total there are 37 features out of them this graph shows 31 categorical features. Comparative to all the categorical features VEHTYPE shows the highest number of classes; 25. Furthermore, AUTOMOBILE has the lowest number of classes among all the features.

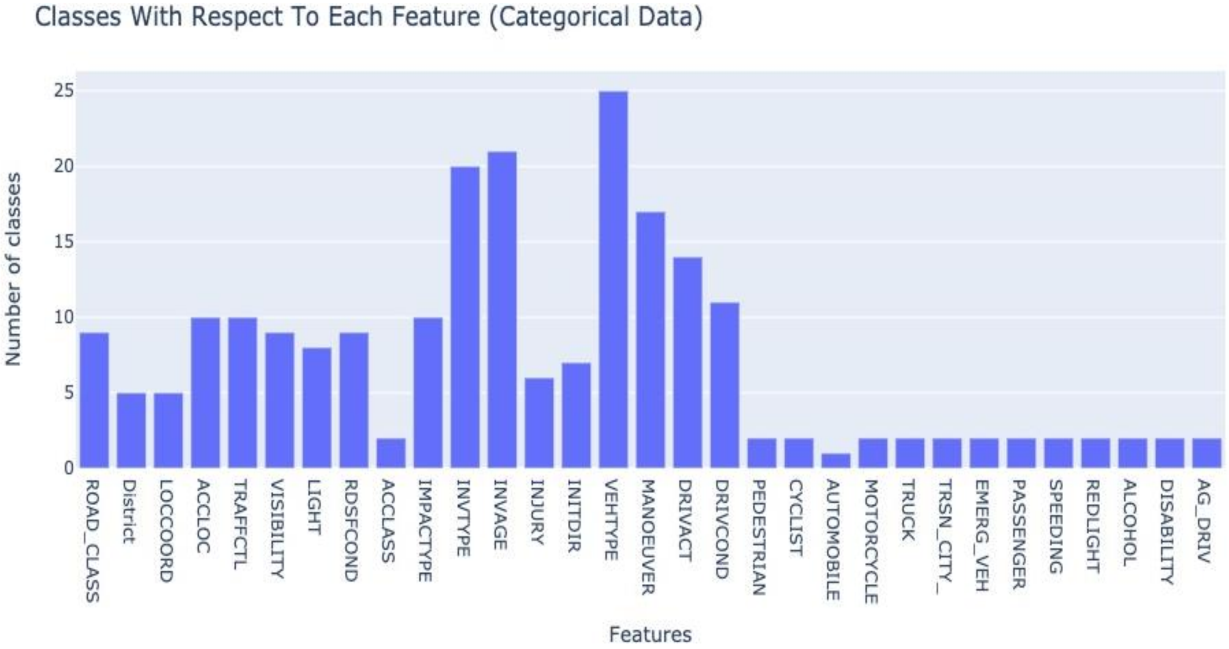


Figure 3: Classes with respect to each feature (Categorical Data)

Figure 4 shows the class label distribution of Impact Type. Since it is the multiclass classification problem, It contains 10 classes with different impact types of accidents. It also shows the number of instances with respect to each class. By looking at the chart, it is clear that the data has an Imbalanced distribution of classes.

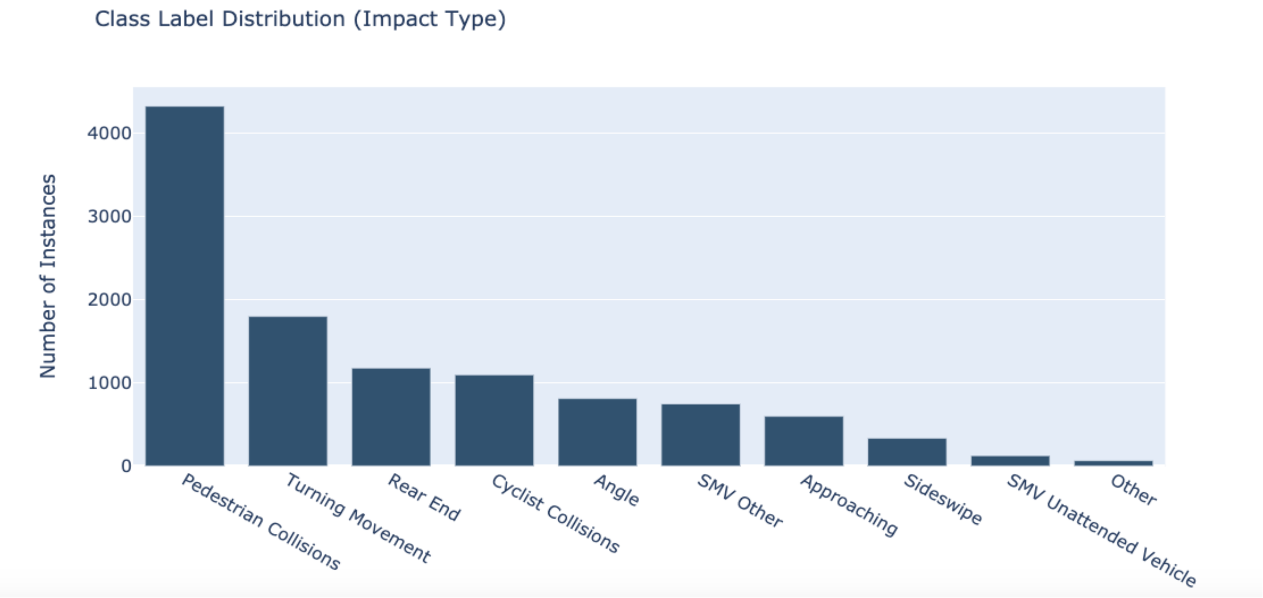


Figure 4: Class Label Distribution (Impact Type)

Figure 5 shows that traffic accidents occurred from the year 2008 till 2018. The X-axis shows the year of the accidents and the Y-axis shows the number of accidents. The graph shows the highest number of accidents recorded was 1200 in the year 2012. Furthermore, the graph shows the lowest number of accidents recorded was in 2018. Moreover, by analyzing the graph it can be concluded that from 2008 to 2012 the accidents were higher than it got a significant drop from 2012 to 2014 and lastly it increased till 2016 and got constant from 2017 to 2018.

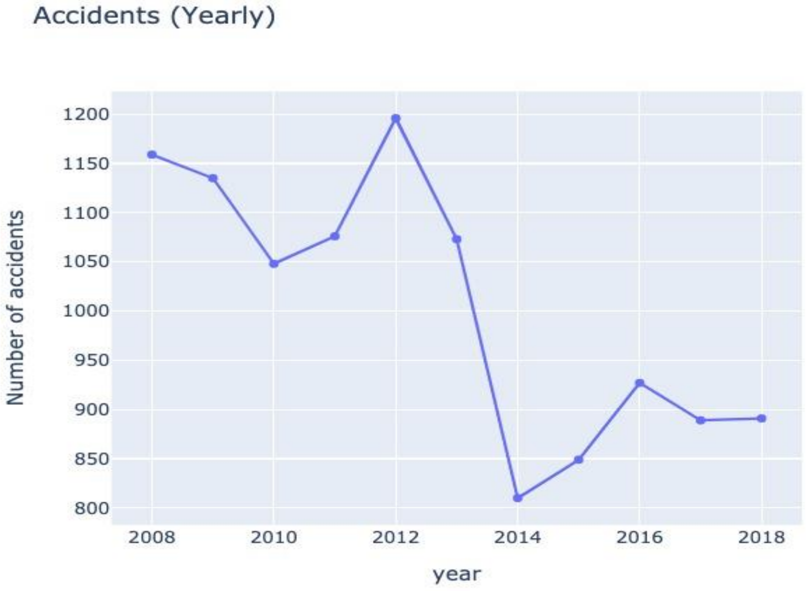


Figure 5: Traffic Accidents by Years

Figure 6 shows the traffic accidents that occurred in 24 hours. The X-axis shows the hour of the accidents and the Y-axis shows the number of accidents. The graph shows the highest number of accidents recorded was in between 15 to 20 hours of the day. Furthermore, the graph shows the lowest number of accidents recorder was in between 0 to 5 hours of the day. It concludes that traffic accidents are lower at the beginning of a day then it keeps increasing until the rush hour and then again keeps decreasing at the end of the day.

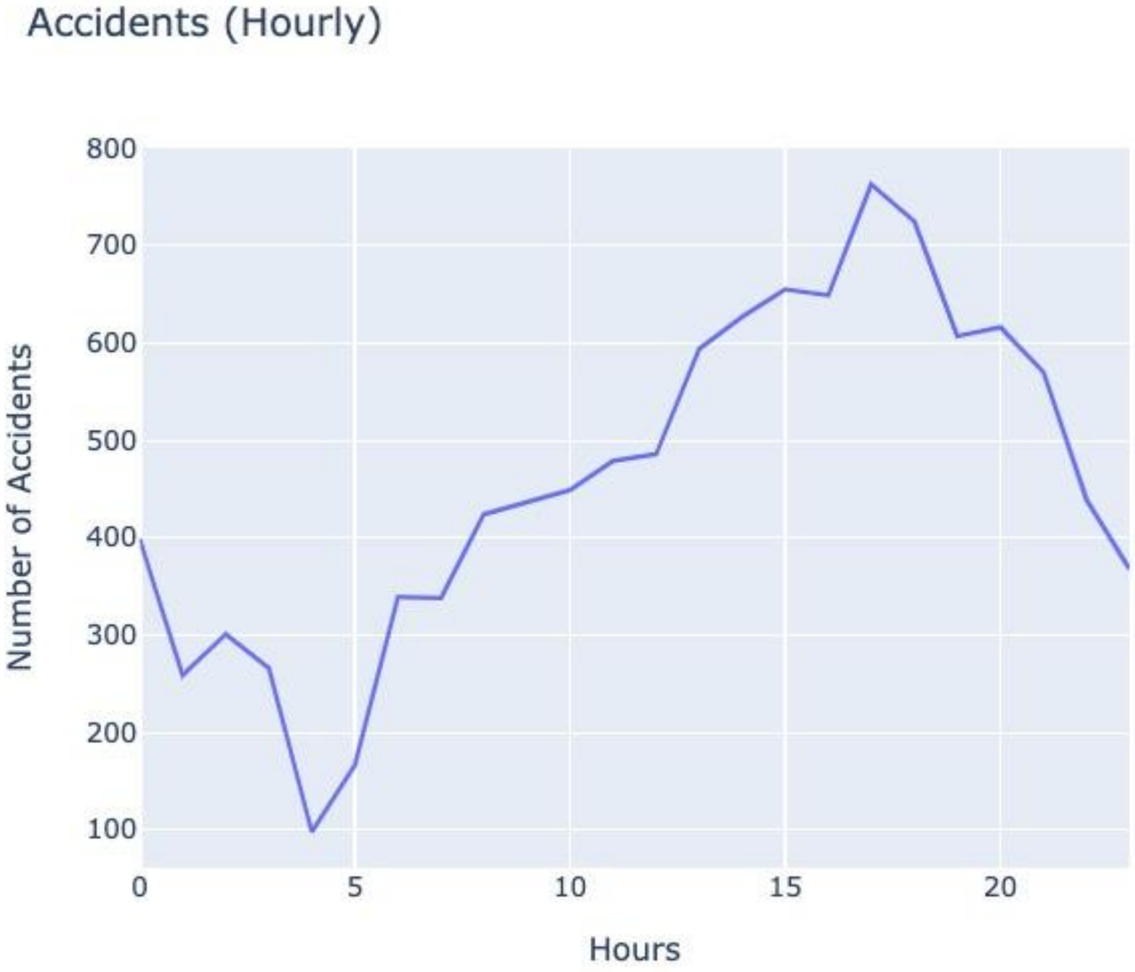


Figure 6: Traffic Accidents by Hours

Toronto is a huge city and is covered with many districts. Figure 7 shows the main districts of Toronto. The DT & EAST York has the highest coverage of Toronto city. Furthermore, North York has the lowest coverage of Toronto city. Since DT & East York have the highest coverage of the city, the over speeding regarding the district is as follows 14% of the response are YES and 86% of the response are No with respect to over speeding.

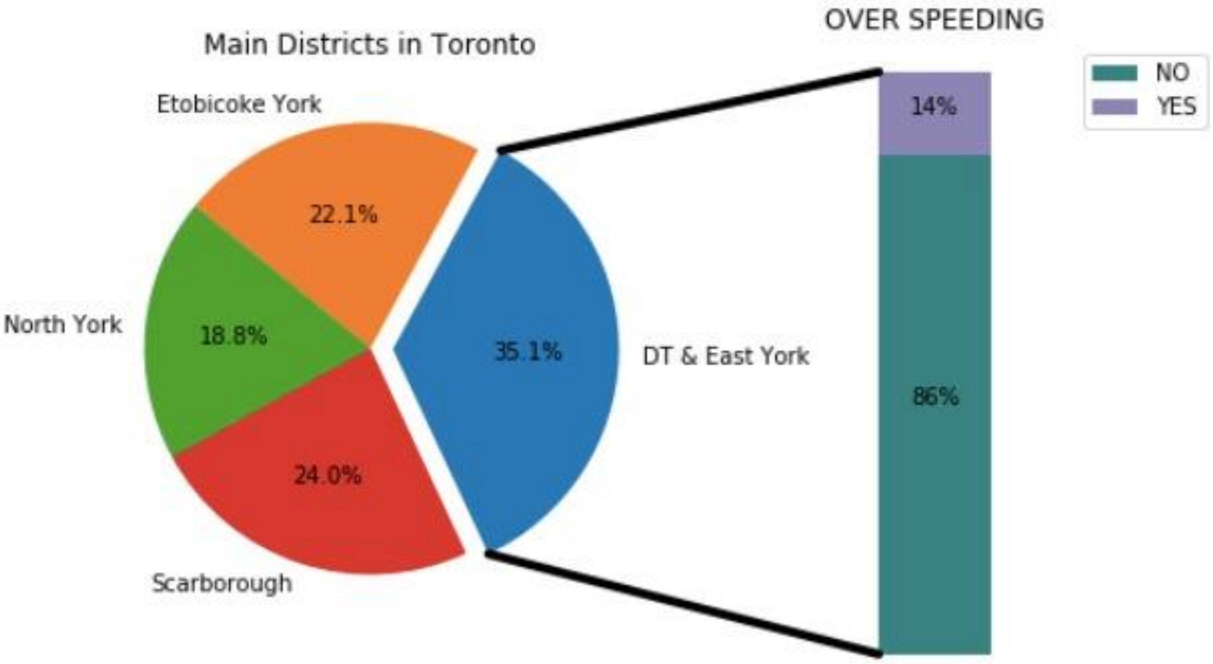


Figure 7: Main districts in Toronto & Over

Figure 8 shows the map of Toronto city. The map has been marked with red, yellow, and green zone according to the severity of traffic accidents. The red zone emphasizes the most accident zone, yellow zone emphasizes less severe accident zone than red, and the green zone emphasizes the least severe accident zone.

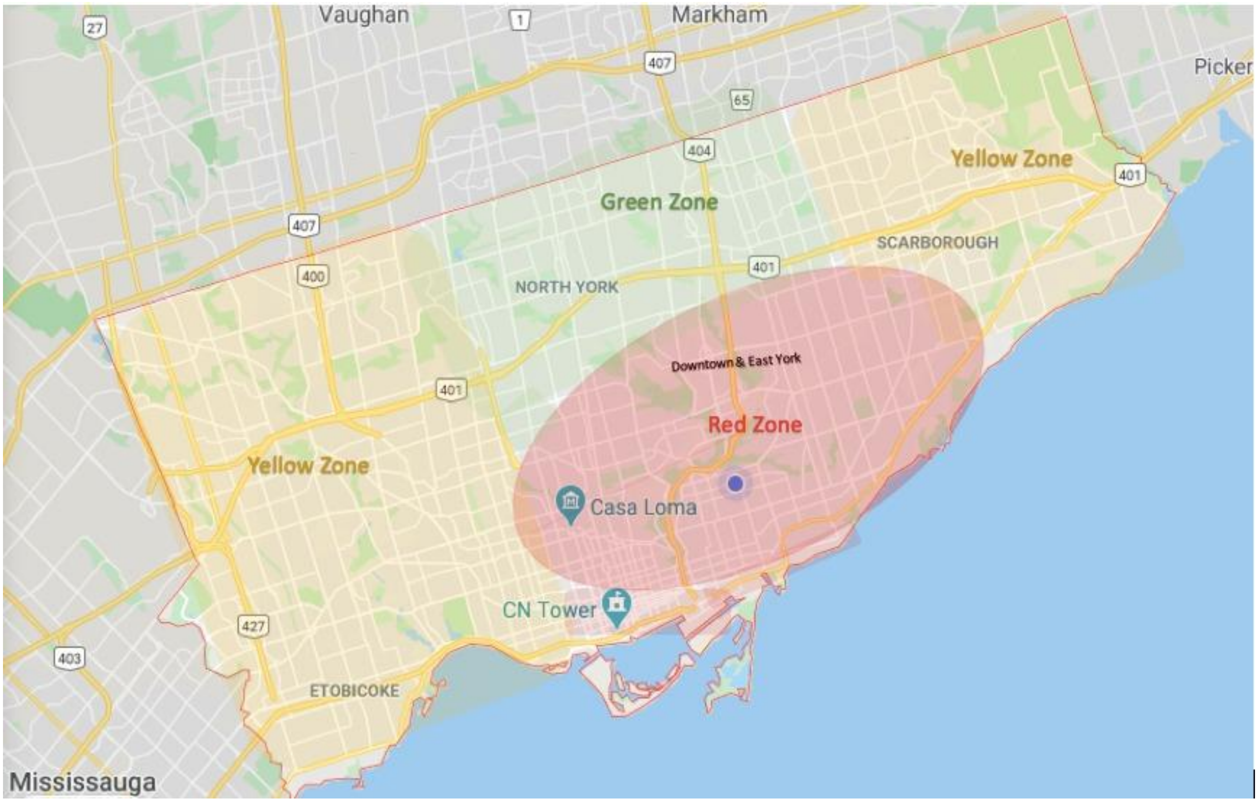


Figure 8: The map of Toronto city marked with accident severity spots

Figure 9 shows the top 5 street types of Toronto with most accidents. The X-axis shows the street type and the Y-axis shows the number of accidents. Moreover, the major arterial has recorded the greatest number of accidents which is more than 7000. Furthermore, the local street type has the lowest number of accidents which is less than 1000.

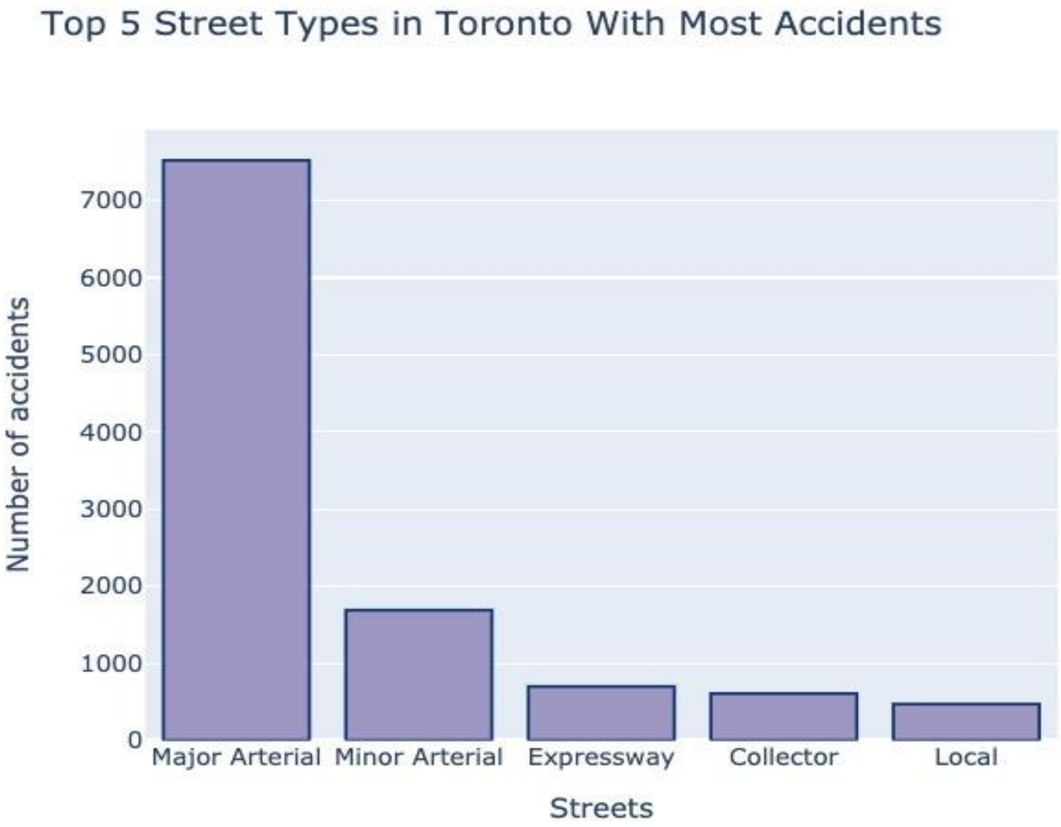


Figure 9: Top 5 Street Types of Toronto with Most accidents

Figure 10 shows the top 5 streets of Toronto with most accidents. The X-axis shows the street name and the Y-axis shows the number of accidents. Moreover, Bathurst St. has recoded the greatest number of accidents which are more than 200. Furthermore, Lawrence St. has the lowest number of accidents comparatively which is 160.

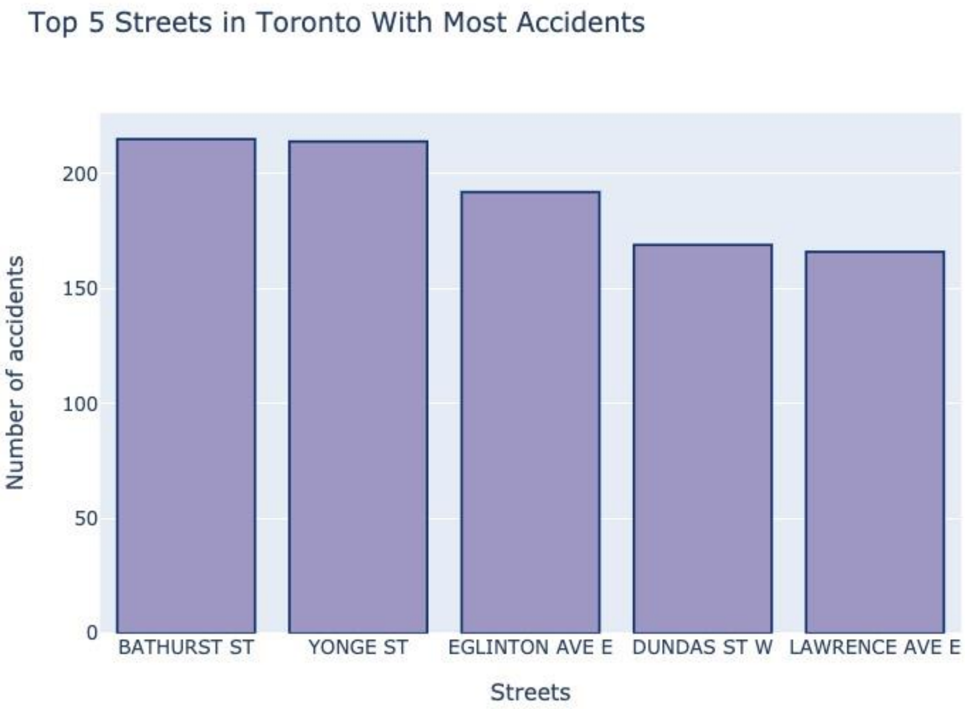


Figure 10:Top 5 Streets of Toronto with Most Accidents

Figure 11 shows the over speed factor with respect to the age group. From the graph below it can be interpreted that from 20 to 29 age group has committed the most traffic accidents. Furthermore, from 60 to 64 age group has committed the least traffic accidents.

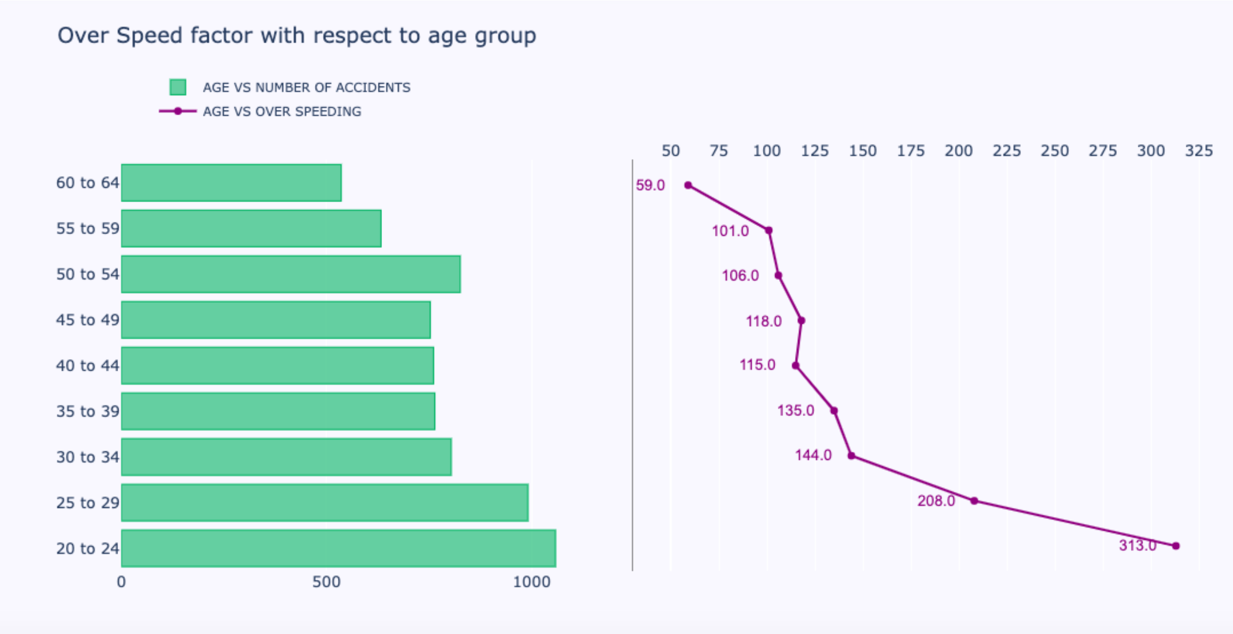


Figure 11: Over speed factor with respect to age group

4. METHODOLOGY | EXPERIMENTS

4.1 Aim of the Study

The aim of the study is to predict multi-class classification and to train the model with patterns responsible for traffic accidents. The model is trained in such a way that it is able to learn all the patterns of the traffic accident and is able to detect the impact type such as turning movement, pedestrian movement, rear- end, approaching sidewalk, or any other impact. Furthermore, the model is able to detect the accident patterns based on the impact type it detects.

4.2 Software & Hardware Requirements / Technologies Used

4.2.1 Software

4.2.1.1 Operating System

- Mac OS Cataline Version 10. 15

4.2.1.2 Programming Language

- Python

4.2.1.3 Integrated Development Environment

- Jupyter
- Google Colab

4.2.1.4 Libraries

- TensorFlow
- Pandas
- Numpy
- Matplotlib
- Scikit learn
- Plotly
- Seaborn
- Keras

4.3 Architecture Diagram

Figure 12 shows the architecture and the main flow of the entire project. Starting from exploratory data analysis until the output, it shows what techniques have been used to achieve the best results. The diagram also shows the best classifiers which were selected based on their performance.

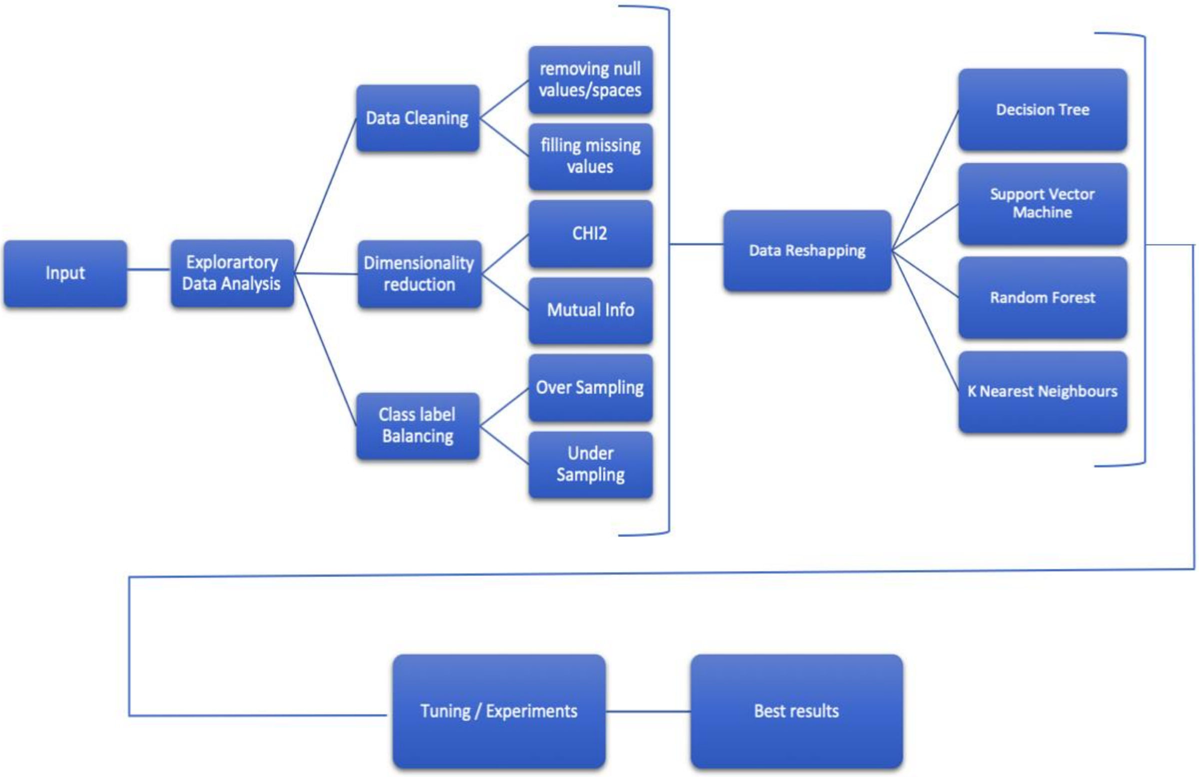


Figure 12: Architecture Diagram

4.4 Data Pre-processing

The raw data that has been collected for the study, consists of 60 features. However not all 60 features have been used in this study. Due to its intense number of features, some of the features have been removed manually in order to get rid of data that was not relevant to the study. The remaining features of the raw data have been used to find the correlation between the features. In order to find the relationship of the features, dimensional reduction technique has been used called the Chi-Square Statistic technique. A Chi-square statistic is a technique to study the relationship between categorical data. The technique has a null hypothesis which indicates in the test there is no relationship exists between the categorical data in the population meaning they are stand-alone or independent data.

$$X^2 = \sum \frac{(o-e)^2}{e}$$

Where *O* is the observed frequency and *E* is the Expectation Frequency. Since the study has used categorical data the Chi-Square technique has been used to assess the relationship between the categorical data. However, during the experimental results, the results of chi-square and mutual information were compared since both techniques are used to extract features from the categorical data. After applying the dimensional reduction technique, the number of features reduced to 20 features depending on the algorithms. However, in some cases for better learning, the dimensions had to be reduced even more than 20 features.

Since the complete dataset was in the categorical form, there was no need to normalize the data while fitting into the model. However, most algorithms require the data to be labeled in numerical form for the training process. For this purpose, the whole dataset has been transformed to numbers by using pandas data frame conversion function.

4.5 Classifiers & Experimental Analysis

4.5.1 Classifier Name: Decision Tree

Approach: Machine Learning | Supervised Learning | Classification

Idea: Decision Trees are the simplest to understand. A decision tree by name is easy to identify due to binary trees. The technique though is supervised learning of Machine Learning Algorithms. Given a problem the outcomes depend on the attributes and the data input. The results will be either in the classifications made for the problem or a Boolean answer is true or false.

Advantages: The advantage of the classifier decision tree is that it can be performeds with little hard data. Furthermore, it's useful because scaling or normalization is not required. Since it is a white box model, it's easily understandable.

Disadvantages: There are several disadvantages while working with the classifier decision tree. The minor effect of the data can bring a significant impact as a whole. Furthermore, the classifier takes time to train. Also, it is inadequate for applying regression and predicting continuous values.

Why a Decision tree is used in this project: Decision tree is a classifier that fits best for the supervised learning problems. In addition, DT is also being used widely for categorical data where each feature consists of multiple classes. Since the dataset that has been used in this project is completely labeled and have categorical variables.

Implementation on the dataset:

Table 1: The experiments of the classifier Decision Tree

Experiments	No. of Features	Over Sampling	Under Sampling	Criterion	FS technique	Accuracy on test	F1 Score	MSE Score	Log Loss Score
Experiment 1	**5**	**no**	**no**	**Gini**	**Chi2**	**0.958**	**0.916**	**0.981**	**4766.351**
Experiment 2	**5**	**no**	**no**	**entropy**	**mutual**	**0.811**	**0.683**	**5.127**	**21485.891**
Experiment 3	**5**	**Yes**	**no**	**Gini**	**Chi2**	**0.958**	**0.915**	**1.007**	**4731.812**
Experiment 4	5	no	Yes	Gini	mutual	0.561	0.354	11.173	50184.842

Experiment 5	**10**	**no**	**no**	**entropy**	**Chi2**	**0.906**	**0.784**	**1.825**	**10707.020**
Experiment 6	**10**	**yes**	**no**	**Gini**	**Chi2**	**0.896**	**0.790**	**2.214**	**11881.339**
Experiment 7	10	no	yes	entropy	mutual	0.624	0.403	9.145	43000.776

4.5.2 **Classifier Name:** Support Vector Machine

Approach: Machine Learning | Supervised Learning | Classification

Idea: SVM is used for both the Regression and Classification techniques of Machine Learning Algorithms. According to Gunn, (1998), It is just like single-layer or multi-layer nets**.** SVM will build a model that will represent a boundary on a region of points which represent data **[21].** The boundary is traditionally known as "hyperplane". Same as Bayesian Network, using SVM with a huge amount of data is not possible so overfitting should be avoided.

Advantages: The advantage of the classifier Support Vector machine is, it avoids overfitting using the regularization parameter. Moreover, another advantage is that the problem can be solved using the kernel. Also, its use since there are no local minima.

Disadvantages: The disadvantage of the classifier is that it solves the problem of overfitting data by modifying the parameters. Also, the kernel is extremely sensitive to overfit however it depends on the model section criterion.

Why SVM is used in this project: Support Vector Machine is a classifier that is used for supervised learning problems. It is a good performing algorithm when it comes to the classification of the categorical variables. Another reason to use this algorithm is that it has been widely used in the literature review, where most problems were relevant to the classification.

Implementation on the dataset:

Table 2: The experiments of the classifier Support Vector Machine

No. of Features	No. of Features	Over Sampling	Under Sampling	Criterion	FS Technique	Accuracy on test	F1 Score	MSE Score	Log Loss Score
Experiment 1	5	no	no	rbf	Chi2	0.278	0.0795	14.427	8395.294
Experiment 2	5	no	no	linear	mutual	0.147	0.081	15.231	6190.303

Experiment 3	5	yes	no	rbf	mutual	0.073	0.066	16.013	7603.339
Experiment 4	5	no	yes	linear	Chi2	0.097	0.077	17.932	7625.912
Experiment 5	10	no	no	rbf	mutual	0.278	0.088	17.289	7583.233
Experiment 6	10	no	no	linear	Chi2	0.117	0.076	17.808	6176.998
Experiment 7	10	yes	no	rbf	Chi2	0.077	0.066	17.670	7574.748

4.5.3 **Classifier Name:** K-Nearest Neighbor

Approach: Machine Learning | Supervised Learning | Classification

Idea: KNN is an algorithm that classifies the data based on a similarity measure. For instance, distance functions such as Euclidean, Manhattan, and Minkowski. However, these three distances are used only in the case, where the data is continuous. In terms of the categorical data, a hamming distance has been used to measure the distance.

Advantages: The advantage of the classifier K-Nearest Neighbor is that it does not require training unlike the decision tree. Also, it's useful since it's easy and simple to understand.

Disadvantages: The disadvantage of the classifier K-Nearest Neighbor is that it's not suitable for large datasets. Furthermore, scaling is necessary, unlike the decision tree. Also, it's hard to avoid noisy data, missing values while using K-Nearest Neighbor.

Why KNN is used in this project: The main reason to use KNN in this project is that it is supervised learning and comes under the classification approach. And it works well when you do not have to deal with large datasets.

Implementation on the dataset:

Table 3: The experiments of the classifier K-Nearest Neighbor

Experiments	No. of Features	Over Sampling	Under Sampling	K-Neighbors	FS Technique	Accuracy on test	F1 Score	MSE score	Log Loss Score
Experiment 1	**20**	**non**	**no**	**1**	**Chi2**	**0.901**	**0.882**	**1.281**	**11294.179**

Experiment 2	**20**	**yes**	**no**	**2**	**mutual**	**0.845**	**0.840**	**1.952**	**11579.181**
Experiment 3	20	no	yes	3	Chi2	0.140	0.134	15.875	76505.016
Experiment 4	25	no	no	2	Chi2	0.757	0.737	3.605	11429.280
Experiment 5	25	yes	no	3	mutual	0.766	0.786	2.990	11740.303
Experiment 6	30	no	no	3	Chi2	0.632	0.610	5.749	10554.418
Experiment 7	**30**	**yes**	**no**	**1**	**mutual**	**0.901**	**0.882**	**1.281**	**11294.179**

4.5.4 **Classifier Name:** Random Forest

Approach: Machine Learning | Supervised Learning | Classification

Idea: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees

Advantages: The advantages of the classifier Random Forest is, it does not require any feature scaling. Another advantage is that its useful for handling non-linear parameters effectively. Also, it can solve both classifications as well as regression problems.

Disadvantages: The disadvantage of the classifier Random Forest is that its a complex approach also it requires a longer training period.

Why Random Forest is used in this project: Random Forest is a classifier that fits best for the supervised learning problems. In addition, RF is also being used widely for categorical data where each feature consists of multiple classes. Since the dataset that has been used in this project is completely labeled and have categorical variables

Implementation on the dataset:

Table 4: The experiments of the classifier Random Forest

Experiment	No. of Features	Over Sampling	Under Sampling	Depth/ state	FS Technique	Accuracy on test	F1 Score	MSE Score	Log Loss Score
Experiment 1	5	no	no	2,0	Chi2	0.648	0.248	11.930	3701.259
Experiment 2	5	no	no	2,100	Chi2	0.664	0.286	11.318	3847.590
Experiment 3	5	yes	no	1,0	mutual	0.592	0.243	12.315	5395.890
Experiment 4	5	no	yes	2,0	Chi2	0.555	0.268	14.847	4159.584
Experiment 5	10	no	no	2,100	mutual	0.628	0.239	11.407	4147.345
Experiment 6	10	yes	no	1,0	Chi2	0.660	0.400	9.313	5318.906
Experiment 7	10	no	yes	2,100	mutual	0.614	0.382	8.873	4709.617

RESULTS

Classifier Name: DECISION TREE

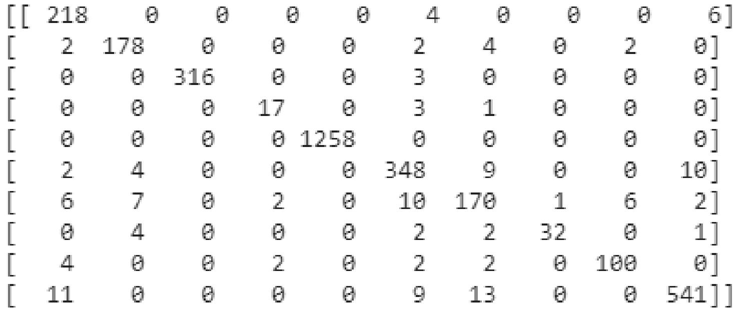
Results:

Table 1 shows the experiments of the classifier decision tree. Total 7 Experiments were conducted on the dataset where 5 Experiments show accuracy over 0.8. Among them, 3 Experiments show more than 0.9 accuracies on the test. One of the most accurate tests is Experiment no.1 which shows 0.958 accuracies were the no. of features was 5, no under or oversampling was recorded, the criterion was gini, and FS technique was Chi2. Furthermore, Experiment no.3 shows 0.958 accuracies were the no. of features was 5, oversampling was recorded, the criterion was gini and the FS technique was Chi2. Moreover, Experiment no.5 shows 0.906 accuracies were the no. of features was 10, sampling was neither over nor under, the criterion was entropy and the FS technique was Chi2. Also, Experiment no. 4 & 7 shows less accuracy respectively 0.561 & 0.624 on the test where no. of features respectively was 5 & 10, both recorded under sampling & FS technique was mutual, but criterion was gini & entropy respectively.

For Decision Tree following is the best configurations:

No. of Features	Over Sampling	Under Sampling	Criterion	FS technique	Accuracy on test	F1 Score	MSE Score	Log Loss Score
5	no	no	gini	Chi2	0.958	0.916	0.981	4766.351

Following is its Confusion Matrix:



```
[[ 218    0    0    0    0    4    0    0    0    6]
 [   2  178    0    0    0    2    4    0    2    0]
 [   0    0  316    0    0    3    0    0    0    0]
 [   0    0    0   17    0    3    1    0    0    0]
 [   0    0    0    0 1258    0    0    0    0    0]
 [   2    4    0    0    0  348    9    0    0   10]
 [   6    7    0    2    0   10  170    1    6    2]
 [   0    4    0    0    0    2    2   32    0    1]
 [   4    0    0    2    0    2    2    0  100    0]
 [  11    0    0    0    0    9   13    0    0  541]]
```

Classifier Name: SUPPORT VECTOR MACHINE

Results:

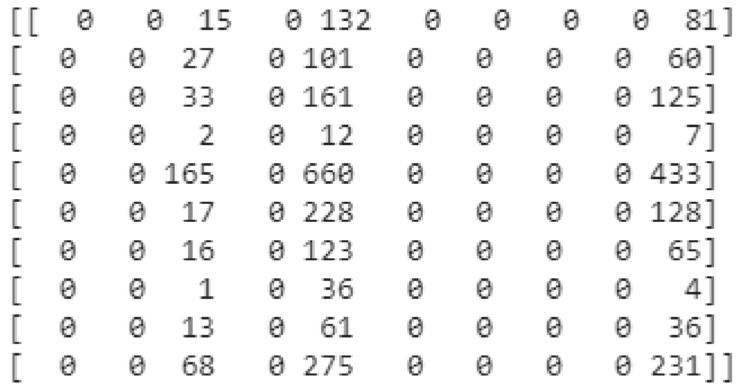
Table 2 shows the experiments of the classifier Support Vector Machine. Total 7 Experiments were conducted on the dataset where 4 Experiments show accuracy over 0.1. Among them, 2 Experiments show more than 0.278 accuracies on the test. One of the most accurate tests was Experiment no.1 which shows 0.278 accuracies were the no. of features was 5, no under or oversampling was recorded, the criterion was rbf and FS technique was Chi2. Furthermore, Experiment no. 5 shows 0.278 accuracies where the no. of features was 10, no under or oversampling was recorded, the criterion was rbf and the FS technique was mutual. Moreover, Experiment no.2 shows 0.147 accuracies where the no. of features was 5, sampling was neither over nor under, the criterion was linear and the FS technique was mutual. Also, Experiment no. 3 & 4 show less accuracy respectively 0.073 & 0.097 on the test where no. of features was 5, Experiment 3 was oversampled and Experiment 4 was under-sampled & FS technique was rbf & linear respectively, the criterion was mutual & Chi2 respectively.

For SVM following were the best configurations:

No. of Features	Over Sampling	Under Sampling	Criterion	FS Technique	Accuracy on test	F1 Score	MSE Score	Log Loss Score
5	no	no	rbf	Chi2	0.278	0.0795	14.427	8395.294

Following is its Confusion Matrix:

- As we can see it cannot predict the data correctly



```
[[  0   0  15   0 132   0   0   0   0  81]
 [  0   0  27   0 101   0   0   0   0  60]
 [  0   0  33   0 161   0   0   0   0 125]
 [  0   0   2   0  12   0   0   0   0   7]
 [  0   0 165   0 660   0   0   0   0 433]
 [  0   0  17   0 228   0   0   0   0 128]
 [  0   0  16   0 123   0   0   0   0  65]
 [  0   0   1   0  36   0   0   0   0   4]
 [  0   0  13   0  61   0   0   0   0  36]
 [  0   0  68   0 275   0   0   0   0 231]]
```

Classifier Name: K-NEAREST NEIGHBOR

Results:

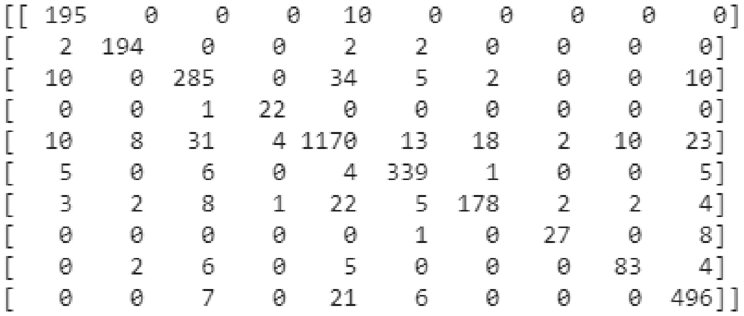
Table 3 shows the experiments of the classifier K-Nearest Neighbor. Total 7 Experiments were conducted on the dataset where 3 Experiments show accuracy over 0.8. Among them, 2 Experiments show more than 0.9 accuracies on the test. One of the most accurate tests is Experiment no.1 which shows 0.901 accuracies were the no. of features was 20, no under or oversampling was recorded, K-Neighbor was 1 and the FS technique was Chi2. Furthermore, Experiment no.7 shows 0.901 accuracies were the no. of features was 30, oversampling was recorded, K-Neighbor was 1, and the FS technique was mutual. Moreover, Experiment no.2 shows 0.845 accuracies were the no. of features was 20, oversampling was recorded, K-Neighbor was 2, and the FS technique was mutual. Also, Experiment no.3,4,5,6 shows accuracy under 0.7 respectively whereas the least accurate experiments were experiment no. 3 & 6. So, Experiment no.3 & 6 accuracy was 0.140 & 0.632 respectively, no. of features respectively was 20 & 30, both recorded no oversampling however Experiment no. 3 recorded under-sampled & FS technique was Chi2 & K-Neighbor was 3 respectively. Point to be noted only one in Experiment no. 3 was under sampled and Experiment no. 3 brought the least accurate result on the test.

For KNN following were the best configurations:

No. of Features	**Over Sampling**	**Under Sampling**	**K-Neighbors**	**FS Technique**	**Accuracy on test**	**F1 Score**	**MSE score**	**Log Loss Score**
20	non	no	1	Chi2	0.901	0.882	1.281	11294.179

Following is its Confusion Matrix

- It can predict 2 to 4 classes but other than that it’s also not good



```
[[ 195    0    0    0   10    0    0    0    0    0]
 [   2  194    0    0    2    2    0    0    0    0]
 [  10    0  285    0   34    5    2    0    0   10]
 [   0    0    1   22    0    0    0    0    0    0]
 [  10    8   31    4 1170   13   18    2   10   23]
 [   5    0    6    0    4  339    1    0    0    5]
 [   3    2    8    1   22    5  178    2    2    4]
 [   0    0    0    0    0    1    0   27    0    8]
 [   0    2    6    0    5    0    0    0   83    4]
 [   0    0    7    0   21    6    0    0    0  496]]
```

Classifier Name: RANDOM FOREST

Results:

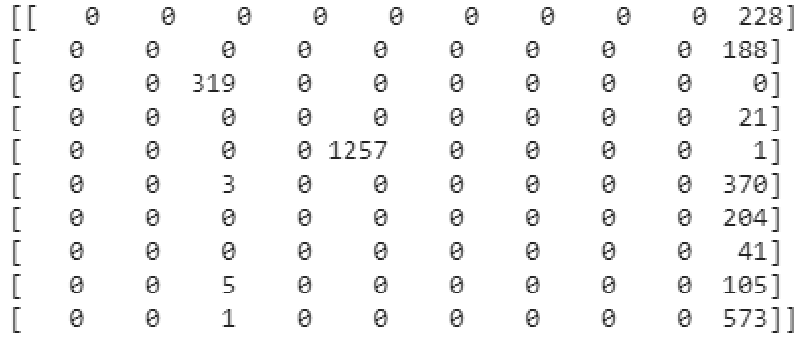
Table 4 shows the experiments of the classifier Random Forest. A total of 7 Experiments was conducted on the dataset. Most experiments show accuracy over 0.8. The top 3 experiments for the random classifier was experiment no. 2, 6 & 1 accordingly. Moreover experiment 2 shows 0.923 accuracies on the test where the number of features was 5, No oversampling and under sampling recorded, depth/state was 2100 and the FS technique was Chi2. Furthermore, experiment 6 shows 0.660 accuracies on the test where the number of features was 10, Oversampling recorded, depth/state was 1,0, and the FS technique was Chi2. Lastly, experiment no. 1 shows 0.648 accuracies on the test where the number of features was 5, No oversampling and under sampling recorded, depth/state was 2,0 and the FS technique was Chi2. The least accurate experiments were 4,3 & 7 where their accuracy on the test was 0.555,0.592 & 0.614.

For Random Forest following were the best configurations:

No. of Features	**Over Sampling**	**Under Sampling**	**Depth/ state**	**FS Technique**	**Accuracy on test**	**F1 Score**	**MSE Score**	**Log Loss Score**
5	no	no	2,0	Chi2	0.923	0.555	11.930	3701.259

Following is its Confusion Matrix:

- It can predict 2 to 4 classes but other than that it’s also not good



```
[[   0    0    0    0    0    0    0    0    0  228]
[   0    0    0    0    0    0    0    0    0  188]
[   0    0  319    0    0    0    0    0    0    0]
[   0    0    0    0    0    0    0    0    0   21]
[   0    0    0    0 1257    0    0    0    0    1]
[   0    0    3    0    0    0    0    0    0  370]
[   0    0    0    0    0    0    0    0    0  204]
[   0    0    0    0    0    0    0    0    0   41]
[   0    0    5    0    0    0    0    0    0  105]
[   0    0    1    0    0    0    0    0    0  573]]
```

Conclusion

Decision Tree is the Best Option in the given multi-classification problem

CONCLUSIONS

This paper has collected multi class classification dataset. Since the dataset has multi class classification, it is obvious that the classes were imbalanced. It was a challenge to prepare the data for machine learning algorithm with multi-class classified data. Appropriate measures have been taken to balance the dataset. After preparing the data the paper was able to extract its outcomes according to its classes. Although the dataset was multi-class classification, the traffic accident patterns for this paper were not much random since the accuracy of the algorithm was more than 95%. This shows that the algorithm did not either under-fit or over-fit. The dataset was trained and was double-checked to learn that is the algorithm being efficient or not. Furthermore, the data set has different patterns of traffic accidents which made this paper possible to cover most patterns of traffic accidents and to propose an algorithm that is able to detect the possible patterns of traffic accidents. Since the data set has been taken out from government analysis it has different patterns of traffic accidents and that made relations, patterns with one another based on seasonality, age factor, street types etc.

Moreover, this paper has used 4 classifiers that are purposely used for supervised learning. Later the outcomes of these classifiers are compared in order to find the best option for the given multi-classification problem. This paper has emphasized machine learning algorithms instead of deep learning models since the data set was limited for deep learning models however it was enough for generating machine learning algorithms. For the future implementation of this project, this paper is hopeful to implement deep learning models by gathering the required amount of data as deep learning models need in order to learn the data better. Based on the outcomes of this paper it can be concluded that it is giving the impact type analyzing the patterns of traffic accidents.

The research provided a great initiative for the upcoming researchers in the country to explore the patterns of traffic accidents. For future implementation, this paper is proposing to work with other city's traffic accident data since this paper used only Toronto city's data. If other city's data are gathered and analyzed across the country, it is possible to implement a model nationwide. The proposed model of this paper can be utilized in the future analysis of other cities in Canada. Although this paper has extracted the impact type from different patterns of traffic accidents however for future implementation it is determined to extract the details of an impact type such as speeding, percentages, intensity, depth of an individual impact type for a certain situation.

REFERENCES

1. Tison, J., Chaudhary, N., Cosgrove, L., & Preusser Research Group. (2011). National phone survey on distracted driving attitudes and behaviors (No. DOT HS 811 555). The United States. National Highway Traffic Safety Administration.

2. Bener, A., & Bener, O. F. (2007). Mobile phone use while driving and risk of road traffic injury: applying the Lorenz Curve and associated Gini Index. Adv Transp Stud, 13, 77-82.

3. a. Issa, Y. (2016). Effect of driver's personal characteristics on traffic accidents in Tabuk city in Saudi Arabia. Journal of transport literature, 10(3), 25-29.

   b. World Health Organization. (2013). Road safety in the Eastern Mediterranean region: Facts from the global status report on road safety 2013 (No. WHO-EM/HLP/075/E). http://www.who.int/violenceinjuryprevention/roadsafetystatus/2013

   c. Langton, James. “As the Spotlight Falls on Saudi's Roads, Its Safety Record Comes into Sharp Focus.” The National, The National, 28 Sept. 2017, www.thenational.ae/world/mena/as-the-spotlightfalls-on-saudi-s-roads-its-safety-record-comes-into-sharp-focus-1.662138.

4. Klein Lawyers LLP, Common Causes of Car Accidents, Causes of Car Accidents in Canada,2020 https://www.callkleinlawyers.com/car-accident-lawyers/types/

5. Al Turki, Y. A. (2014). How can Saudi Arabia use the decade of action for road safety to catalyse road traffic injury prevention policy and interventions?. International journal of injury control and safety promotion, 21(4), 397-402.

6. Government of Canada, Canadian Motor Vehicle Traffic Collision Statistics, 2018 https://www.tc.gc.ca/eng/motorvehiclesafety/canadian-motor-vehicle-traffic-collision-statistics-2018.html

7. World Health Organization. Country Cooperation Strategy for WHO and Saudi Arabia 2006–2011. 2006.

8. Al-Naami, M. Y., Arafah, M. A., & Al-Ibrahim, F. S. (2010). Trauma care systems in Saudi Arabia: an agenda for action. Annals of Saudi Medicine, 30(1), 50-58.

9. Wang, P., Ni, C., & Li, K. (2019, December). Vision-based highway traffic accident detection. In *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing* (pp. 1-5)

10. Ferreira, J., Carvalho, E., Ferreira, B. V., de Souza, C., Suhara, Y., Pentland, A., & Pessin, G. (2017). Driver behavior profiling: An investigation with different smartphone sensors and machine learning. PLoS One, 12(4), e0174959.

11. Mohan, P., Padmanabhan, V.N., Ramjee, R.: 'Nericell: rich monitoring of road and traffic conditions using mobile smartphones'. Proc. of the Sixth ACM Conf. on Embedded Network Sensor Systems, 2008, pp. 323–336

12. White, J., Thompson, C., Turner, H., Dougherty, B., Schmidt, D.C.: 'WreckWatch: automatic traffic accident detection and notification with smartphones', Mob. Netw. Appl., 2011, 16, (3), pp. 285–303

13. Dai, J., Teng, J., Bai, X., Shen, Z., Xuan, D.: 'Mobile phone based drunk driving detection'. Fourth Int. Conf. on Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2010, pp. 1–8

14. Ghosh, S., Sunny, S. J., & Roney, R. (2019, March). Accident Detection Using Convolutional Neural Networks. In 2019 International Conference on Data Science and Communication (IconDSC) (pp. 1-6). IEEE.

15. Parsa, A. B., Taghipour, H., Derrible, S., & Mohammadian, A. K. (2019). Real-time accident detection: coping with imbalanced data. Accident Analysis & Prevention, 129, 202-210.

16. Yuan, Z., Zhou, X., & Yang, T. (2018, July). Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 984-992).

17. Ren, H., Song, Y., Wang, J., Hu, Y., & Lei, J. (2018, November). A deep learning approach to the citywide traffic accident risk prediction. In 2018 21st International Conference on Intelligent Transportation Systems (ITSC) (pp. 3346-3351). IEEE.

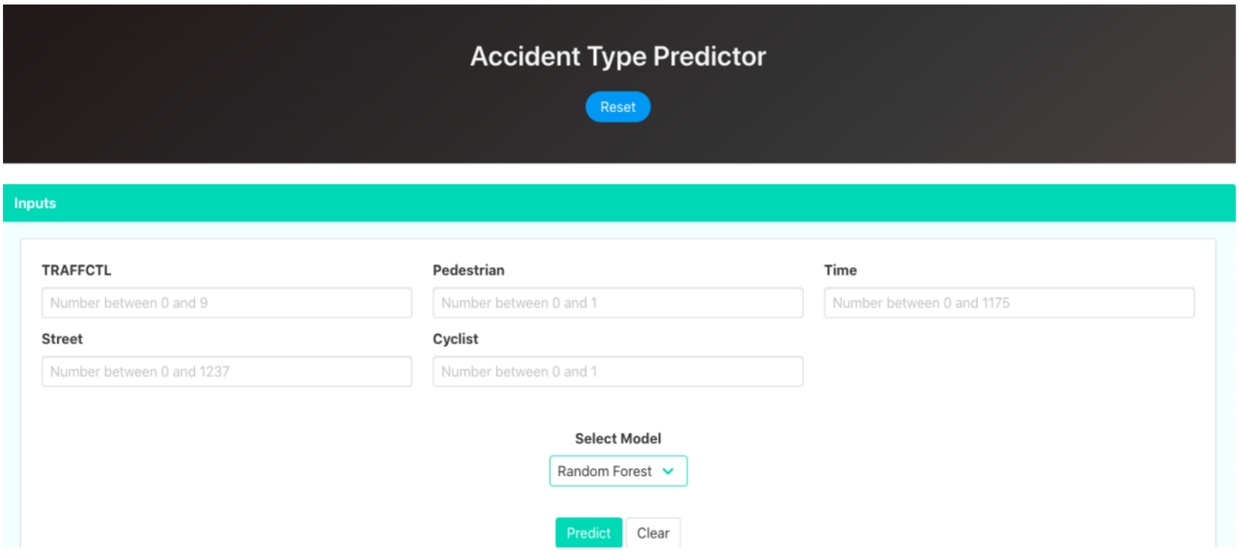
18. Toronto police service, Public Safety Data Portal, Automobile, 2020
https://data.torontopolice.on.ca/datasets/automobile

19. Toronto police service, Public Safety Data Portal, Open Government Licence – Ontario, 2019
https://data.torontopolice.on.ca/pages/licence

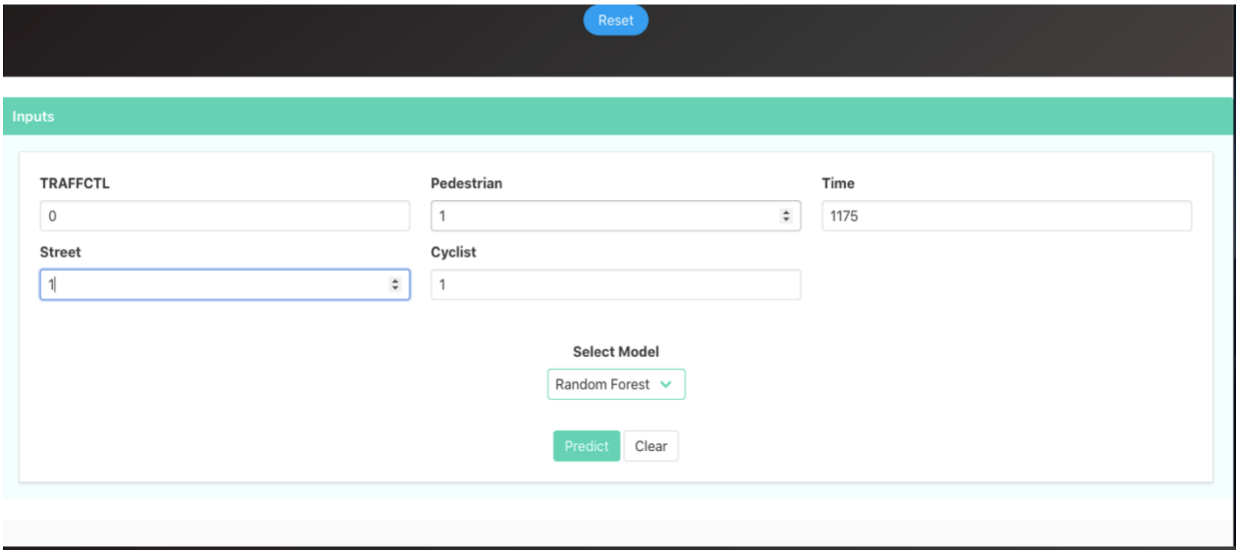
20. Gunn, S. R. (1998). Support vector machines for classification and regression. ISIS technical report, 14(1), 5-16. http://users.ecs.soton.ac.uk/srg/publications/pdf/SVM.pdf

21. When do support vector machines trump other classification methods by Bala Deshpande http://www.simafore.com/blog/bid/112816/When-do-support-vector-machines-trump-otherclassification-methods

APPENDIX



Select Model
Random Forest



Inputs

TRAFFCTL: 0

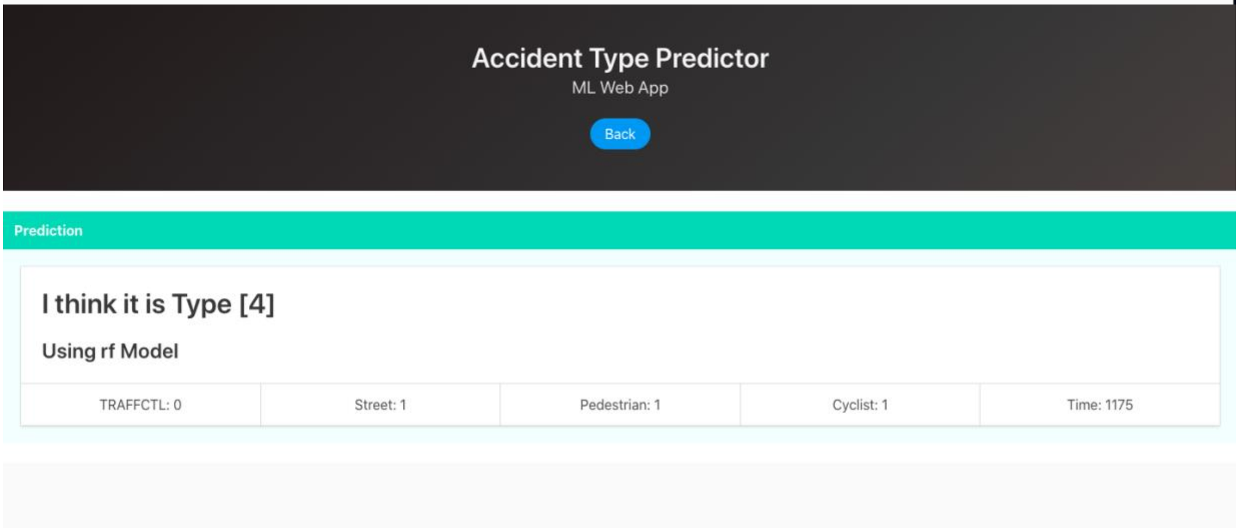
Pedestrian: 1

Time: 1175

Street: 1

Cyclist: 1

Select Model: Random Forest



Accident Type Predictor

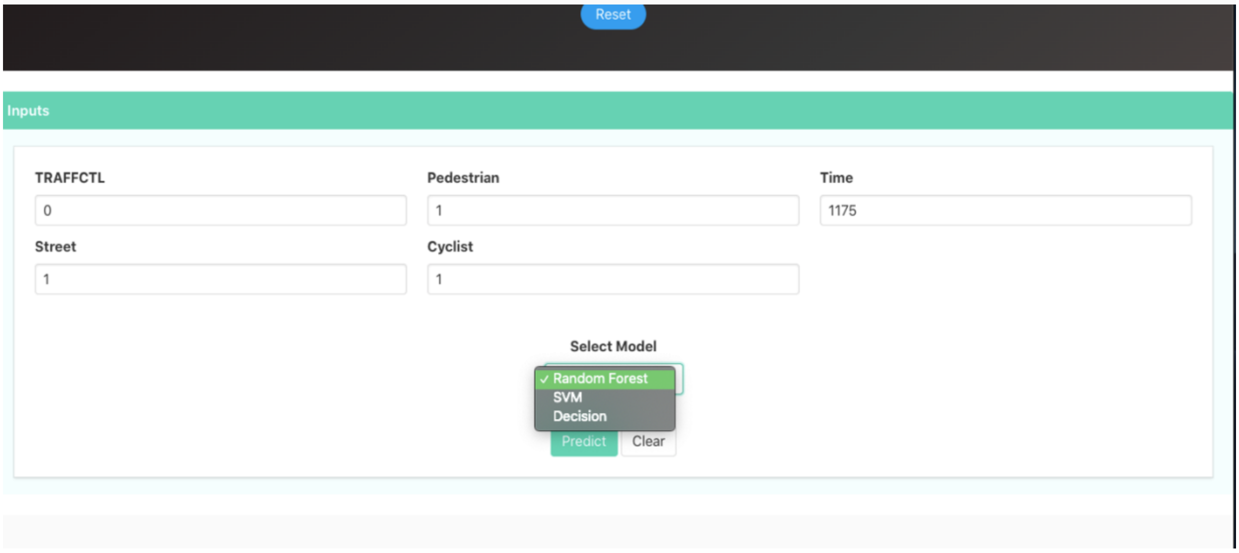
ML Web App

Prediction

I think it is Type [4]

Using rf Model

TRAFFCTL: 0	Street: 1	Pedestrian: 1	Cyclist: 1	Time: 1175



Inputs

TRAFFCTL: 0

Pedestrian: 1

Time: 1175

Street: 1

Cyclist: 1

Select Model