



***Predicting Traffic Fatalities in Canada by using Machine Learning and Deep Learning***

BY

Tejas M Patel (0734912)

Soorya Suresh (0735168)

Sunny Singh (0733910)

# **INTRODUCTION**

Transportation plays an indispensable role in our society and their dependency is relying more on us. Automobilization has crusaded a serious threat to traffic and the environment. The disruption of transport flow has not been considered vital and more importance has been given to the infrastructure of roads and highways. Transportation is important as it enables communication, trade and exchange of goods and services across the globe. At the same time growth of transportation has very unfortunate impact on the society in terms of accidents. Accidents are increasing day by day. We will be trying to find possible insights from this data about the reasons as to why accidents occur, which area is more prone to accidents and what are the aftereffects of accidents on traffic flow. Since transportation industry is ripe for advancement, data science can bring about an evolution in this sector. Data analytics can provide an in-depth knowledge of methods for analyzing and implementing intelligent transportation systems. The benefits of big data and analytics helps transportation firms to precisely enhance the model capacity, demand, revenue, pricing, customer sentiments, cost. Some of the assets include implementing real time monitoring system for traffic management system to improve bus transportation and reduce traffic congestion. By developing this project, we can help society with implementing some laws needed for transportation which can prevent the accidents and helps them to know at where they must be safer on roads which are more prone to accidents.

**DATA DRIVEN APPROACH**

This dataset is all about accidents that occurred in Canada during the years 1999 to 2017 and accidents has major impact in every phase of the world. we will try to find enriching factors associated to accident fatalities and for that we will use machine learning which will help us to predict the fatalities.

Our hypothesis says that accident is causing a lot of fatalities because of weather, as we are aware Canada is known for its weather and due to snow many drivers are not able to drive properly and its difficult to drive in such situation as it requires a lot of attentiveness. Some lags in it and they end up being getting hit or damaging infrastructure and due to this there are long traffic jams on the road which increases the pollution and waste a lot of time of other people. We will try to find the vital effect behind accident and how we can reduce this by implementing safety for such reasons and try to predict the fatal injuries because of the accidents.

# RELATED WORK

* **An Improved Deep Learning Model for Traffic Crash Prediction**

This is the model which is based on the data from USA. In USA traffic crashes are the number one cause for death among people. This data is collected from the united states police and state highway-asset-management databases, the analysis of traffic safety estimate and predicate the likelihood of a traffic crash. The proposed model includes two modules, an unsupervised feature learning module to identify functional network between the explanatory variables and the feature representations and a supervised fine-tuning module to perform traffic crash prediction. The number of crashes occurring on a specific time period will be considered as the dependent variable and the and the factors affecting the likelihood of traffic crash are analyzed and examined. The proposed model that includes the MVNB (Multivariate negative binomial) regression layer in the supervised fine-tuning module can better account for differential distribution patterns in traffic crashes across injury severities and provides superior traffic crash predictions. The findings suggest that the proposed model is a superior alternative for traffic crash predictions and the average accuracy of the prediction that was measured by RMSD can be improved by 84.58% and 158.27% compared to the deep learning model without the regression layer and the SVM model, respectively.

* **Traffic accident analysis using Machine learning Paradigms**

This paper is talking about modeling the severity of injury that occurred during traffic accidents and the mechanisms they used for their research work is Artificial neural networks using hybrid learning, decision trees. These techniques can help to understand the characteristics of drivers’ behavior, roadway condition and weather condition that were causally connected with different injury severity. This can help decision makers to formulate better traffic safety control policies.

The analysis is focused on vehicle accidents that occurred at signalized intersections. The injury severity was divided into 3 classes no injury, possible injury and fatal injury. They compared the performance of Multi-layered Perceptron (MLP) which is a feed forward neural network with one or more hidden layers and Fuzzy ARTMAP (Adaptive Resonance Theory) and found that the MLP classification accuracy is higher than the Fuzzy ARTMAP. Levenberg-Marquardt algorithm was used for the MLP training and achieved 65.6 and 60.4 percent classification accuracy for the training and testing phases, respectively.

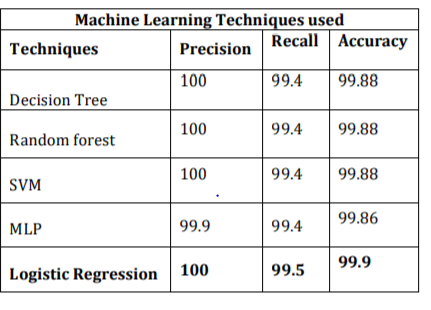
The Fuzzy ARTMAP achieved a classification accuracy of 56.1 percent. In Decision tree the performance for no injury, possible injury and fatal injury is 67.54, 64.40, and 89.46. In Hybrid DT-ANN approach the associated accuracies are 83.02% and 65.12%, 74.93% and 63.10%, 91.53% and 90.00%. The future work they are going to focus on to finding out the speed of the vehicles during crashes.

* **Using machine learning to predict car accident risk**

This paper is about Predicting accident risk per road segment per hour within the specified area using ML and the mechanisms they used are supervised machine learning and gradient boosting There are numerous possible applications, including the following that we have considered for applications: Safe route planning, Emergency vehicle allocation, Roadway design, Where to place additional signage (e.g. to warn for curves). The future work they are going to do is to use real time traffic information to improve the model significantly.

* **Machine learning based traffic congestion prediction in a IoT based smart city**

This paper problem statement is Traffic congestion prediction using machine learning and mechanism they are using decision tree, Random forest algorithm, SVP, MLP, Logistic Regression in this model they are using various sensors to analyze the traffic flow and free flowing of traffic. with all the sensors deployed at the crucial junctions. The data is being gathered from different junction points through different sensors. The data is assumed to be stream data which is time dependent. Our goal is to predict the congestion on any specific path which is about to occur in the due time. They applied five different machine learning techniques with the help of WEKA tool to identify the best method which can predict accurately the traffic congestion. Among this Logistic Regression has outperformed all the other machine learning techniques. The reason for this is that, since the data is time dependent and regression methods are good at predicting for the time dependent data. The metrics used to measure the prediction results are Precision, Accuracy and Recall. Precision is defined as True Positive (true positive+false positive). Recall is defined as True Positive (true positive + false negative). Accuracy is defined as all True values by summation of all True and False values. The future work they are going to do is to do prediction of traffic congestion using hybrid techniques which can give high accurate results. Their prediction result table is below.



* **Vehicle crashes and machine learning**.

This includes predicting crash fatalities with machine learning. The mechanisms used are Regression, Random forest. The Crash Analysis System (CAS) data is available in different formats and APIs. It is simple to grab them through API interfaces, instead of downloading to your local machine. This is beneficial, as we will access the latest updated data every time, we run the Jupyter notebook. This will help to understand the causes and factors that affect the car crash severity. Road accidents constitute a major problem in our societies around the world. For example. In the year 2016, the USA alone had recorded 37, 461 motor vehicle crash-related deaths, averaging around 102 people per day. Roads, speed limit, weather and other related attributes also indicate crash severity as well as fatality level. The modeling part of this problem is approached in different ways. Regression problem will help us predict the number of fatalities based on the attributes of the crash dataset. It can also be approached as a classification problem and predict the severity of the crash based on the crash dataset. Random forest algorithm is also used since it performs well on many datasets. In future this project can be improved by considering more models and evaluating accuracy.

* **Live Prediction of traffic accidents risk using machine learning and Google Maps**

The dataset is based on UK where the government has published detailed records of traffic accidents in the country dating back to 2002. With the help of this a machine learning model is being made that predicts with high accuracy as to when and where accidents are likely to occur in Greater London. The main aim of this project is to create an interactive traffic accident predictor that would be easily accessible by anyone. The steps undertaken to achieve this objective are – create an interactive model that can identify the accident spots along a driving route with London taking into account the local weather condition during the travel, then Data collection, Data Processing, Supervised Learning and finally Deployment. The algorithms used were Logistic Regression, SVM and Random Forest. Among the models tested, the best performing model was Random Forest trained on all numerical features. This was the model then selected for Deployment. The results of the modelling steps are summarized in the table below:

A screenshot of a cell phone

Description automatically generated

* **High-Resolution Road Vehicle Collision Prediction for the City of Montreal**

Road accidents are an important issue of our modern societies, responsible for millions of deaths and injuries every year in the world. In Quebec only, in 2018, road accidents are responsible for 359 deaths and 33 thousand of injuries. In this paper, we show how one can leverage open datasets of a city like Montreal, Canada, to create high-resolution accident prediction models, using big data analytics. Compared to other studies in road accident prediction, we have a much higher prediction resolution, i.e., our models predict the occurrence of an accident within an hour, on road segments defined by intersections. Such models could be used in the context of road accident prevention, but also to identify key factors that can lead to a road accident, and consequently, help elaborate new policies.

We tested various machine learning methods to deal with the severe class imbalance inherent to accident prediction problems. In particular, we implemented the Balanced Random Forest algorithm, a variant of the Random Forest machine learning algorithm in Apache Spark. Interestingly, we found that in our case, Balanced Random Forest does not perform significantly better than Random Forest.

Experimental results show that 85% of road vehicle collisions are detected by our model with a false positive rate of 13%. The examples identified as positive are likely to correspond to high risk situations. In addition, we identify the most important predictors of vehicle collisions for the area of Montreal: the count of accidents on the same road segment during previous years, the temperature, the day of the year, the hour and the visibility.

* **Road Accident Prediction using Machine Learning Algorithm**

In India, more than 150,000 people are killed each year in traffic accidents. That’s about 400 fatalities a day and far higher than developed auto markets like the US, which in 2016 logged about 40,000. Every year over 1 million vehicles are added to traffic averagely. 1.2 million People have died and over 50 million people have been injured in road accidents in the world every year. Studies on traffic have executed that road accidents and death- laceration ratio will increase.

The traffic has been transformed into the difficult structure in points of designing and managing by the reason of increasing number of vehicles. This situation has discovered road accidents problem, influenced public health and country economy and done the studies on solution of the problem. Arising the need of accession to information from this large calibrated data obtained the corner stone of the data mining. In this study, assignment of the most compatible machine learning classification techniques for road accidents estimation by data mining has been intended.

After the data is been cleaned and transformed it’s ready to process further. After the data has been cleaned and we have taken the required constraints. We divide the whole dataset int o the two parts that can be either 70-30 or 80-20. The larger portion of the data is for the processing. The algorithm is applied on that part of data. Which helps the algorithm to learn on its own and make prediction for the future data or the unknown data. The algorithm is executed in which we take only the required constraints from the cleaned data. The output of the algorithm is in ‘yes’ and ‘no’. It gives the error rate and the success rate.

In that classification algorithm we will use Logistic Regression Algorithm The logistic algorithm will make the prediction in terms of percentage, to find accuracy level in percentage and Error percentages. This Algorithm is only for the yes and no type of result or successful and unsuccessful. The equation for combinations of all 15 input variables

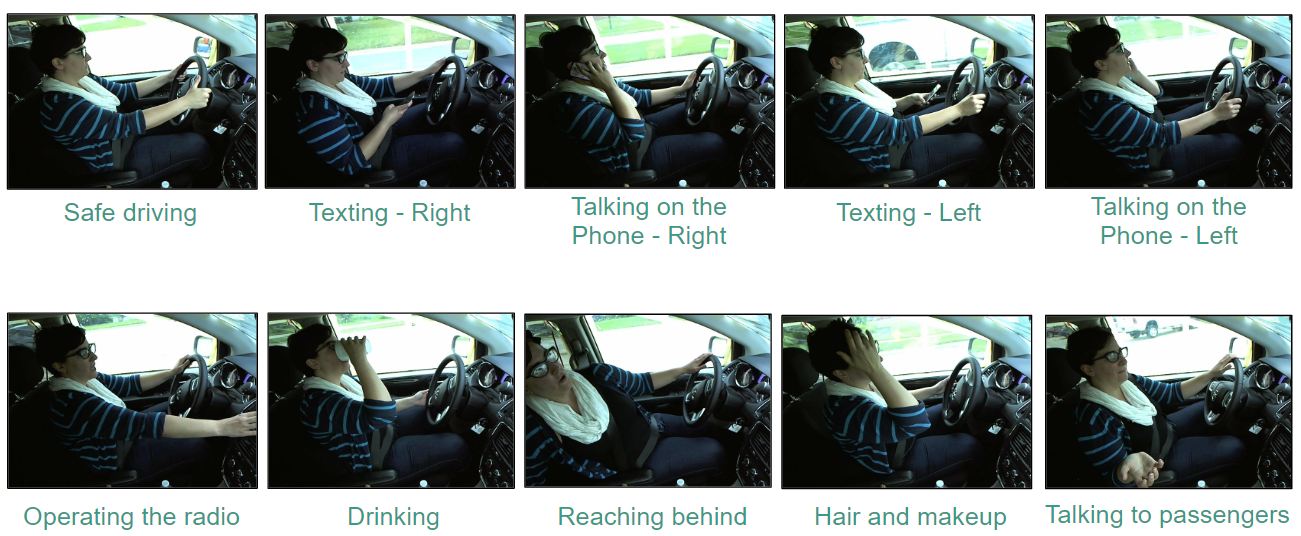
# Distracted Driver Detection using Deep Learning

Driving a car is a complex task, and it requires complete attention. [Distracted driving](https://www.unece.org/fileadmin/DAM/trans/doc/2016/wp1/ECE-TRANS-WP1-2016-1e.pdf) is any activity that takes away the driver’s attention from the road. Several studies have identified three main types of distraction: visual distractions (driver’s eyes off the road), manual distractions (driver’s hands off the wheel) and cognitive distractions (driver’s mind off the driving task).

The National Highway Traffic Safety Administration (NHTSA) reported that 36,750 people died in motor vehicle crashes in 2018, and 12% of it was due to distracted driving. Texting is the most alarming distraction. Sending or reading a text takes your eyes off the road for 5 seconds. At 55 mph, that’s like driving the length of an entire football field with your eyes closed.

We took the [State Farm](https://www.kaggle.com/c/state-farm-distracted-driver-detection) dataset which contained snapshots from a video captured by a camera mounted in the car. The training set has **~22.4 K** labeled samples with equal distribution among the classes and **79.7 K** unlabeled test samples. There are 10 classes of images:





* **Vehicle Accidents prediction using machine learning in New Zealand**

Road accidents constitute a major problem in our societies around the world. The World Health Organization (WHO) estimated that 1.25 million deaths were related to road traffic injuries in the year 2010. For the year 2016, the USA alone had recorded 37, 461 motor vehicle crash-related deaths, averaging around 102 people per day.**Can machine learning help us understand the causes and the factors that affect car crash severity. In** New Zealand, the total fatalities in crash accidents since the year 2000, up to 2018 is 6922. While the total number of serious injuries and minor injuries in car accidents reach 45044, 205895 respectively. Although this dataset records all crashes reported to NZ Police, we have to consider that all crashes are not reported to NZ police especially non-fatal crashes. Most of the crashes are non-injury crashes while fatal crashes are the least. In terms of fatality counts, most of the crashes have 0 fatality rate. **In** this article, we will do a complete machine learning pipeline from getting data through APIs, performing exploratory data analysis and formulating a real-world problem into a machine learning model.

We can approach the modeling part of this problem in different ways. We could take it as a regression problem and predict the number of fatalities based on the attributes of the crash dataset. We can also approach it as a classification problem and predict the severity of the crash based on the crash dataset. In this example, I will approach it as a regression problem. Feel free to build a classification model if you want to give it a try. It will basically be the same approach. I will not do any feature engineering in this case, I think the attributes we have are enough to build a baseline, and we can always revisit this and do feature engineering later to boost our model accuracy

# METHODS

We will be working on the dataset which is having in total of **6772563 (6.77 Million) rows and 23 features** and in this data we need to do data wrangling to make it useful for analysis as in this data we have columns that are not that useful for model to do analysis so we have to remove that and also there will be some unknown values in the data which has to be removed. And this data is in categorical format, so we must modify the data according to our need and create a useful dataset.

We will be building lot of models for doing the analysis on the data like we will use classification models like **Logistic, Random Forest, XGBoost, and Deep learning**. we have data which is categorical. We will also build graph and do visualization of the data and find some valuable insights from the graphs.

**Step 1**

**About the data**

The data that we have is about accidents in Canada from year 1999-2017 downloaded from opensource Canada.org the data consist of 6772563 rows and 23 features.

**Step 2**

**Importing the libraries necessary building model**

**import** **numpy** **as** **np**

**import** **pandas** **as** **pd**

**import** **matplotlib.pyplot** **as** **plt**

**import** **os**

**import** **datetime**

**import** **seaborn**

**import** **warnings**

**from** **sklearn.preprocessing** **import** MinMaxScaler

**from** **sklearn.model\_selection** **import** train\_test\_split, cross\_val\_score

**from** **sklearn** **import** metrics

**import** **xgboost** **as** **xgb**

**from** **xgboost** **import** plot\_importance

**Libraries in brief:**

* NumPy provides objects for multi dimensional arrays for high level mathematical functions and scientific computations
* Pandas provides objects for Data Frames which are useful for analysing data
* Matplotlib is a plotting library which help in visualisation
* OS allows to interface with the underlying operating system
* Datetime is a module which contains a type that is also called datetime. since we wanted to make date as index column
* seaborn is a data visualization library in python based on matplotlib. It helps to create more attractive and informative statistical matplotlib graphics
* warnings messages are issues in situations where it is useful to alert the user of some condition in a program
* Min Max scaler helps in transforming features by scaling each feature to a given range which will help in better visualization
* Train-Test split for splitting data into train and test.
* Metrics was used to evaluate machine learning algorithms in python. Choice of metrics influence how the performance of machine learning algorithm is measured and compared
* import xgboost as xgb is used for extreme gradient boosting. It is used to increase speed and increase the efficiency of the model
* from xgboost import plot importance was used to visualize xgboost model feature importance in python

**Step 3**

**DATA CLEANING**

* At first the data was from year 1999-2017 which was huge data and this huge data is not required for analysis because there is no benefit , as it is only going to slow the analysis process so we filtered the data from 2012 and now we have 1849249 rows in the data.
* Then we checked for the null values in the data, there are no null values in our dataset.
* Then we dropped all the unknown values from the dataset UU, XX.
* Now the rows are reduced to 1051219
* Then we dropped the unwanted columns which are of no help in predicting the model. Now the cleaned data has 1051219 rows and 19 columns instead of 23.

**Step 4**

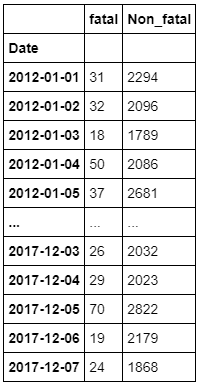
**Data Transformation**

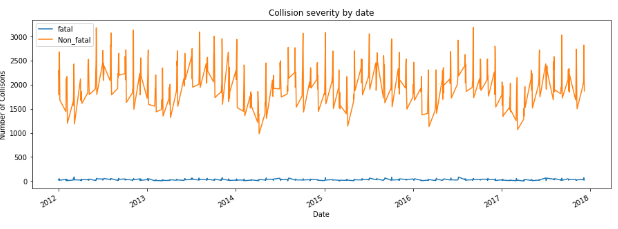
* We used the date time function and made the date as index.
* Then we performed correlation, there are some rows that are positively correlated and negatively correlated. The most positively correlated attribute is C\_RCFG (collision road configuration) and the most negatively correlated attribute is P\_ISEV (passenger medical treatment required)
* Since the P\_SEX column was categorical, and it has characters in it. So, to convert it to integer we made a new column PSR\_SEX. using **np. where** which will help in assigning the associated value 1 for male and 0 for female.
* We applied one hot encoding to the column P\_USER (road user class) for creating more features to improve the model performance and generalize the data better. This can only be applied to categorical columns.
* This will help us find which user is more prone to accidents and will help in giving better result for column P\_ISEV (medical treatment required) The column is represented in codes where
* code 1 – motor vehicle driver
* code 2 – motor vehicle passenger
* code 3 – pedestrian
* code 4 –Bicyclist
* code 5 – Motorcyclist
* There is no column for P\_USER\_3 because no pedestrian was involved in accidents in the data
* Then we MinMaxScaler to perform scaling in the data to bring the data in one range to perform a good analysis.

**Step** **5**

**Exploratory Analysis**

* To find the pattern from the data and to test our hypothesis by using visualization charts
* Then we checked for fatal and non-fatal columns and Plotted a chart on collision severity by date
* From the chart we can see that it illustrates seasonal behavior within the data got the result as follows





* Then we checked for collision severity by weather
* The graph shows severity of collision due to weather.
* From the graph we can understand that number of fatal deaths due to weather is more than 10k and is higher during clear and sunny day and lowest during freezing rain and strong wind days.

1 - Clear and sunny

2 - Overcast, cloudy but no precipitation

3 - Raining

4 - Snowing, not including drifting snow

5 - Freezing rain, sleet, hail

6 - Visibility limitation eg: drifting snow, fog, smoke, mist, dust

7 - Strong wind

A screenshot of a social media post

Description automatically generated

* Then we checked for collision severity by road configuration
* From the graph we can understand that fatalities due to road configuration is more than 8k which is at non-intersection and at an intersection of at least two public roadways. Other reasons are having no much impact to collisions.

1- Non-intersection

2 - At an intersection of at least two public roadways

3 - Intersection with parking lot entrance/exit, private driveway or laneway

4 - Railroad level crossing

5 - Bridge, overpass, viaduct

6 - Tunnel or underpass

7 - Passing or climbing lane

8 - Ramp

9 - Traffic circle

10 - Express lane of a freeway system

11 - Collector lane of a freeway system

12 - Transfer lane of a freeway system

A screenshot of a social media post

Description automatically generated

# **RESULTS**

**Insights from data by doing Analysis**

* From the column PSR\_SEX we found the count of number of males 554095 and that for females 497124
* From the column P\_ISEV (Medical treatment required) which is represented in codes 1,2,3. We got the count for:

Code 1- No injury – 456145

Code 2- injury – 589357

Code 3- Fatality- 5717

* From the column C\_SEV (collision severity) which is also represented in codes we found the count for:

Code 1- collision producing at last one fatality- 14590

Code 2- collision producing nonfatal injury - 1036629

A screenshot of a cell phone

Description automatically generated

* **Model building**

## **We split the data into train and test with 75% and 25% respectively**

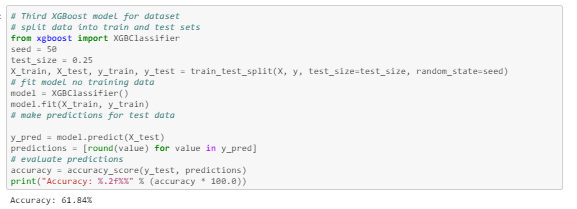
* **Model No 1: Logistic Regression**
* Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist.
* We imported necessary libraries and Min Max scaler for feature transformation and split the data into train and test 75 and 25 percent respectively.
* We tried 3 Logistic Regression models taking parameters 0.1, 10 and 1000 and got 59% accuracy for both training and test.
* Model screenshots are shown below.



* **Model 2: RANDOM FOREST CLASSIFIER**
* The random forest is a classification algorithm consisting of many decision trees.
* Using Random Forest, we got 63% accuracy for training and 61% accuracy for test.
* We tried n\_estimators = 50,100,200 and got this accuracy. After that there was no improvement. So, we took this accuracy
* Random forest model is not much good for dataset as there is no improvement in the accuracy more than 60%



* **Model 3: XG BOOST CLASSIFIER**
* XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.
* In XG Boost classifier, we used the seed value of 50 and we were able to get an accuracy of 61.84



## **We splitted the data into train and test with 60% and 40% respectively**

* We used all same models to evaluate with different train and test split
* **Model No 1: Logistic Regression**
* We tried 3 Logistic Regression models taking parameters 0.1, 10 and 1000 and again got same accuracy 59% accuracy for both training and test.
* Model screenshots are shown below.



* **Model 2: Random Forest model**
* Using Random Forest, we got same accuracy again 63% accuracy for training and 61% accuracy for test.
* We tried n= 50,100,200 and got this accuracy. After that there was no improvement. So, we took this accuracy.



* **Model 3: XG BOOST CLASSIFIER**
* XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.
* In XG Boost classifier, we used the seed value of 20, 50 and 20000 and we were able to get an accuracy of 61.87 for seed value of 20000%



* **Model 4:**  **XG BOOST CLASSIFIER WITH KFOLD**
* In XG Boost classifier using cross validation, we took kfold of 7 and the accuracy was low to 61.83%.
* We tried Kfold of 5 and 10 also but the best accuracy we got was with kfold 7.
* By this accuracy score we got to know that our average cross validation score is not that high, there can be splits for which the classifier performed significantly less.
* The XGBoost is one of the best models and we used it by using various hyperparameter.



* **Model 5 : NEURAL NETWORK**
* The model type we used is **Sequential.**
* Sequential is the easiest way to build a model in Keras. It allows you to build a model layer by layer. Each layer has weights that correspond to the layer the follows it.
* ‘Dense’ is the layer type. Dense is a standard layer type that works for most cases.
* Early stopping will stop the model from training before the number of epochs is reached if the model stops improving. We will set our early stopping monitor to 3.

A screenshot of a cell phone

Description automatically generated

* We use the ‘add ()’ function to add layers to our model. Activation’ is the activation function for the layer. An activation function allows models to consider nonlinear relationships. The activation function we will be using is ReLU or Rectified Linear Activation and Softmax. ReLU is two linear pieces, it has been proven to work well in neural networks. Softmax makes the output sum up to 1 so the output can be interpreted as probabilities.

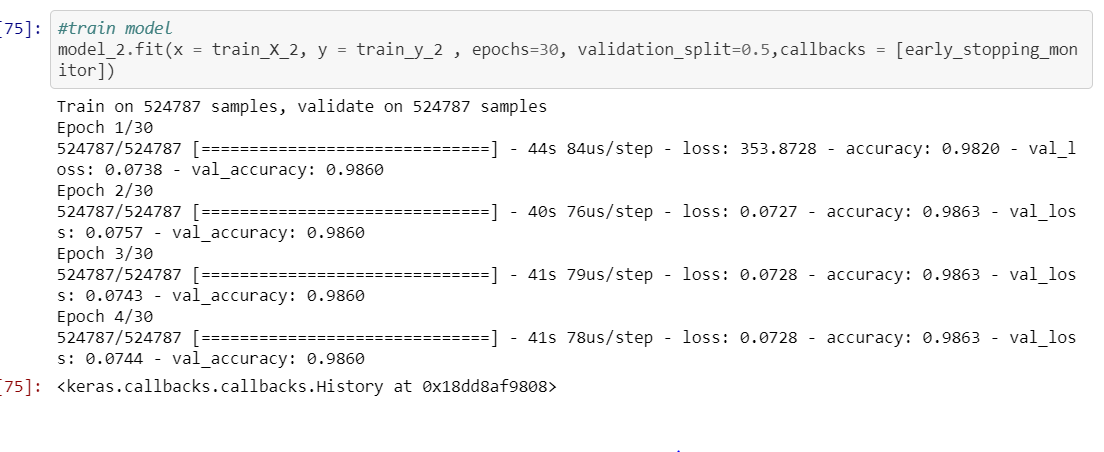
A screenshot of a cell phone

Description automatically generated

* The optimizer controls the learning rate. We will be using ‘Adam’ as our optimizer. In our dataset as our column C\_SEV has 1 and 2 so when we will do it's one hot encoding then it will convert in binary code so it will take 3 values for it that's why it will show 3 value.
* When separating the target column, we need to call the ‘to\_categorical ()’ function so that column will be ‘one-hot encoded’
* In our dataset as our column C\_SEV has 1 and 2 so when we will do it's one hot encoding then it will convert in binary code so it will take 3 values for it that's why it will show 3 value
* The activation is ‘Softmax’. Softmax makes the output sum up to 1 so the output can be interpreted as probabilities. The model will then make its prediction based on which option has a higher probability
* We will use categorical\_crossentropy’ for our loss function. This is the most common choice for classification. A lower score indicates that the model is performing better
* To make things even easier to interpret, we will use the ‘accuracy’ metric to see the accuracy score on the validation set at the end of each epoch.

**Train model**

* Validation set is all about use to predict for test value and it is one of the best features of Keras library
* The validation split [0.1]. Keras proportionally split your training set by the value of the variable. The first set is used for training and the 2nd set for validation after each epoch.
* In this dataset I have used total two split of validation one is 0.2 and 0.5 respectively. So it will use the data for training set according to it as we have seen the dataset in which for first split it used around 8 lakh and 5 lakhs for training respectively.
* Keras proportionally split your training set by the value of the variable. The validation split variable in Keras is a value between [0..1]. We got the highest model accuracy for neural network that is 98.6%.



A screenshot of a social media post

Description automatically generated

# DISCUSSION

* By applying analytics in this data we got to know that in 5 years period from 2012 to 2017 in which the death were 5717 and we thought the majority of accident would have occurred during winter climate but by our visualization we got to know that in summer there was the most.
* We got the accuracy of 98.6% so we have achieved what we desired to achieve and also, we were able to find more insights from the data as I said above.
* The results are in line with the others as there are not many people worked on this dataset the candidates who worked on dataset they have worked in different way and they have only focused on visualization and analysis and we have did prediction as well as applied deep learning to our model.
* The data was very imbalanced so it was very difficult to choose model to work pretty well on the data and the dataset is also very large because of that model selection was very tough so I used KNN model but that was not able to run because computing distance for such a big test set is very difficult so I was not able to run this model properly this was the biggest failure I experienced in this work.
* The challenges which we have faced while working on this project was that to deal with unknow value and selecting the models which will work best on the dataset and there are many columns in data which are not that important and the data is very imbalanced and after doing this challenges I got the best model

# Conclusion

* In this study we conducted an analysis on traffic fatalities and factors responsible for the increasing number of deaths in Canada using open data provided by Opengov.data
* By applying analytics in this data, we got to know that in 5 years period from 2012 to 2017 in which the death was 5717
* It is obvious as more people travel, and more transportation occurs so to reduce the accident we need to implement strict rules.
* In future we can improve this model by more accuracy if we get to know speed and location of accident where the accident occurs.
* Using the dataset, we build traffic fatality prediction models using Supervised Learning Algorithms.
* Our best model can predict 98.6% of traffic fatalities and predict the number of collision and factors responsible for the deaths in Canada.
* Our analysis can be used to identify the major reason for death due to road configuration and extreme climatic condition in order to take action to reduce the risk of accidents.
* we believe that our work can easily be reproduced for other countries and cities under the condition that similar datasets are available
* Finally, our study shows that open data initiatives are useful to society because they make it possible to study critical issues like road accidents.

# Contribution

* **All assessments** – Prepared by Soorya, Sunny and Tejas
* **All weekly meeting minutes**- Prepared by Soorya, sunny, Tejas
* **Python File**- Prepared by Tejas
* **Presentation**- Prepared by Soorya
* **Report**- Prepared by Soorya, sunny, Tejas

# References

* **Dataset Reference**

Open.canada.ca. (2020). *National Collision Database - Open Government Portal*. [online] Available at: https://open.canada.ca/data/en/dataset/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a [Accessed 5 Mar. 2020].

* **Related Work References**

Medium. (2020). *Predicting Vehicle Accidents with Machine Learning*. [online] Available at: https://towardsdatascience.com/predicting-vehicle-accidents-with-machine-learning-ce956467fa74 [Accessed 5 Mar. 2020].

Medium. (2020). *Live Prediction of Traffic Accident Risks Using Machine Learning and Google Maps*. [online] Available at: https://towardsdatascience.com/live-prediction-of-traffic-accident-risks-using-machine-learning-and-google-maps-d2eeffb9389e [Accessed 5 Mar. 2020].

Devi, S. and Neetha, T. (2020). *Machine Learning based traffic congestion prediction in a IoT based Smart City*. [online] Semanticscholar.org. Available at: https://www.semanticscholar.org/paper/Machine-Learning-based-traffic-congestion-in-a-IoT-Devi-Neetha/5df39be1e0397a164a4ba6d198481ee233301047 [Accessed 5 Mar. 2020].

Arxiv.org. (2020). [online] Available at: https://arxiv.org/pdf/1905.08770.pdf [Accessed 5 Mar. 2020].

Medium. (2020). *Distracted Driver Detection using Deep Learning*. [online] Available at: https://towardsdatascience.com/distracted-driver-detection-using-deep-learning-e893715e02a4 [Accessed 5 Mar. 2020].

Medium. (2020). *Predicting Vehicle Accidents with Machine Learning*. [online] Available at: https://towardsdatascience.com/predicting-vehicle-accidents-with-machine-learning-ce956467fa74 [Accessed 5 Mar. 2020].

Anon, (2020). [online] Available at: https://www.researchgate.net/publication/220166391\_Traffic\_Accident\_Analysis\_Using\_Machine\_Learning\_Paradigms [Accessed 5 Mar. 2020].

Pdfs.semanticscholar.org. (2020). [online] Available at: https://pdfs.semanticscholar.org/5df3/9be1e0397a164a4ba6d198481ee233301047.pdf [Accessed 5 Mar. 2020].

Irjet.net. (2020). [online] Available at: https://www.irjet.net/archives/V4/i5/IRJET-V4I5834.pdf [Accessed 5 Mar. 2020].

Medium. (2020). *Predicting Crash Severity for NZ Road Accidents*. [online] Available at: https://towardsdatascience.com/predicting-crash-severity-for-nz-road-accidents-6214117e73fb [Accessed 5 Mar. 2020].

* Github Repository link: <https://github.com/sooryasuresh-star/capstone-project>

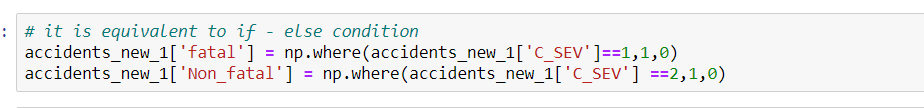
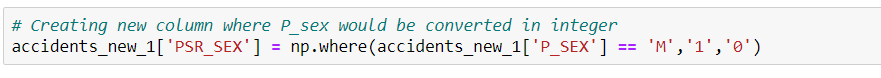
# Appendices

We have included few codes into our appendices for easy understanding

A picture containing screenshot

Description automatically generated

There were few columns in or dataset which were having unknown values like UU and XX. So, we wanted to remove these from or data for better understanding. We selected all columns and dropped those unwanted columns and replaced those errors with Nan using the python function coerce.



The column P\_SEX was in categorical format. So, we wanted to convert it to integer for better model performance. So, we made a new column PSR\_SEX and using np. where function assigned the value 1 for male and 0 for female.

The attribute C\_SEV has two codes which defines code 1 for fatality and code 2 for nonfatal. So,

for better understanding, using the function np. where we assigned the values 1 for fatal and 0 for non-fatal. Fatal in first column and nonfatal in second column.

A screenshot of a cell phone

Description automatically generated

In Artificial Neural Network, in order for the model to work we first convert it into categorical value using one hot encoding Since the target variable is C\_SEV, when we convert into categorical it takes three input as to represent one we need “0, 1,0” and for 2 we need “0,0,1” if it would be 0 and 1 then they would have reported with two column instead of three format for better model performance. Hence, using the function to\_categorical converted a class vector to binary class matrix and that’s why in final layer we have three nodes.