Predictive Analysis on Road Accident Risks Based on Heterogeneous Sparse Data

Madhurima Chakraborty	Pragya Yadav	Debayan Das
19MCA0087	19MCA0125	19MCA0070

Project Guide: Prof. Deepa N.

School of Information Technology & Engineering (SITE)
Vellore Institute of Technology (VIT)
Vellore, Tamil Nadu, India

I. Abstract:

The population growth and the increase in the number of vehicles these days cause traffic congestion, hence leading to the increased count of road accidents. The risks of accidents have also increased relatively. This project is aimed to predict the accident hotspots based on various parameters. Various analytical data mining techniques have been taken into account to analyze the accidental heterogeneous sparse data of United States and make predictions of the risk zones and accident prone hotspots.

II. Introduction:

From past survey’s it is found that, due to urbanization, the standard of living of people has upgraded, and hence has boosted an increase in both population and vehicle. Every person owns one or more than one vehicle nowadays, which has increased. This has led to increased commotion on the roads. Transportation is a fundamental part of our lives, as every person needs vehicle at some point of the day. People going to schools, workplaces, recreational or shopping places need vehicles.

Not only the increase in vehicular traffic have given way to increased risk in road accidents, but also various natural conditions contribute to the same. A cloudy or humid day might have more chances of accidents as compared to a clear day. Road conditions are also a factor for accident risks, like potholes, bumps and sharp turnings, might cause unexpected vehicular collisions. Similarly, surveys have revealed that accidents are prone to occur at certain peak hours or rush hours, for example, the time at which the office hours end show much more traffic congestion than the rest of the hours. Thus, such congestions might lead to accidents.

Road accidents are a major concern nowadays. A road accident refers to any accident involving at least one road vehicle, occurring on a road open to public circulation, and in which at least one person is injured or killed. Accidents have always caused much loss, not only to life, but also economic.

In this project, we aim to predict such accident risks in USA on the basis of various features, like temperature, humidity, latitude, longitude, bumps, crossings, turning points, etc. The features cover all climatic, natural, and road related parameters, which might case an accident. The aim is to predict rush hours, at which the accidents might occur, locating the accident hotspots, and calculating the severity index of the accidents. We rely on neural network algorithm, and data mining techniques to predict the above.

III. Literature review:

In the paper [1], the authors have used Hadoop to analyze the large data based on various criteria to predict road accidents, in order to raise precaution alarms for the same. On comparing Hadoop with other methods, it has proved to be the most efficient method for big data analysis. The algorithms used in this paper are CCMF and TCAMP which analyze the dataset effectually to predict the risks of road accidents. The proposed algorithm tries to predict the road accident risks with the help of the enormous data about vehicle movement and circumstances that favor the accidents.

This paper [2] focuses on finding and predicting road accident patterns based on the severity, road type, accident type, climate, hour of accident etc. The method of finding interesting and useful patterns from spatial database is termed as spatial data mining. Spatiotemporal algorithms are able to locate hidden patterns more easily than traditional data mining techniques. Spatial data is the data which contains location on the earth surface. Two spatiotemporal clustering algorithms are used for finding patterns, which are DBSCAN or Density Based Spatial Clustering of Applications with Noise and Grid Based algorithms. While on one hand, the data for DBSCAN must have data points and limited threshold, the grid based clustering algorithms require a multi-dimensional data structure. It has been found that Grid based algorithms like STING and CLIQUE produce more accurate results along with faster processing time.

In [3] traffic accident is inferred form heterogeneous data. Huge amount of heterogeneous data containing accident data and GPS records have been collected, to check how vehicle mobility affect the road accidents. On analyzing this data a Stack de-noise auto encoder model is prepared which studies the features in human mobility to predict accident risks. The model on being prepared simulates real time accident risks, which can be used to warn people of the possible accident, for a safer route and journey.

Anupama Makkar et. Al has analysed the accident dataset of recent years to forecast road accidents in [4]. The proposed approach in this paper incorporates amalgamation of machine learning algorithms like Bayes Net, j48 graft and j48 decision tree in the data mining process to examine the performances of the algorithms in prediction of accidents. It has been thus noticesd that the combination of such algorithms render better results than a single algorithm used. The results obtained would support in forecasting road traffic accidents and hence prevention and control can be provided.

The crucial issue of predicting road traffic and the accidents caused due to it has been addressed in this paper [5]. The sparse and heterogeneous data have been worked upon by various algorithms to find interesting patterns. Through a case study, this paper explores various algorithms to predict the amount of accidents occurring every hour. The problem is framed as a binary classification problem. Big data inclusive of features like weather, accidents, road networks etc. have been map-matched. Algorithms like support vector machine, deep neural networks, random forest and decision tree have been evaluated along with Eigen analysis.

In this paper,[6] the features causing road traffic accidents were detected using three techniques of classifications; which are: Decision trees, SVM and ANN. Based on these algorithms a prediction model is built, testing all these algorithms on real time data set. The results determine that Random Forest algorithm gave the most accurate predictions, followed by ANN and then SVM. Hence a data mining model using decision trees is made for forecasting the accidents.

Accident Prediction Model is proposed by the authors in the paper [7], where accidents on horizontal curves are anticipated based on the patterns examined in the dataset. This prediction model takes measures to lessen the accidents at some extent. The factors which cause road accidents are analyzed to make predictions and take safety measures to decrease the accident rates. In this paper, the geometric features of the road are collected at different level length, height, stretch, etc. Regression modelling is used for the data mining procedure.

This paper [8] has used two predictive models for the analysis of previous accident data and current accident data to predict the number of accidents occurred in that year. Multiple Linear Regression as well as Artificial Neural Networks has been used for the predictive analysis. After conducting the analysis it is concluded that the predicted values from regression model had greater errors. While the predictions made from Artificial Neural Network analysis was more accurate having less errors. Hence ANN was proved to be a better methodology to make predictions for accidents.

The problems addresses in the paper [9] are, predicting the number of accidents occurring on intersection of roads or any road and finding roads prone to accident risks. Algorithms have been used on heterogeneous data to mine traffic risk. The algorithm incorporated for the framework is an advanced feature based non-negative matrix factorization (FNMF). This framework is successful in predicting the traffic risks at any road or intersection more accurately than the existing algorithms or methods. Two clusters were defined that segregate the risk locations, in which one cluster had larger roads with accident risks, and other cluster having higher risk of vehicle collision. Risky locations were ranked based on the results of the clusters.

In this paper [10], it states that reducing traffic accidents is a necessary concern for safety. Small datasets, depending on large data and not valid for real-time data are basic demerits of previous studies. The deep neural network model is used to provide a solution for real-time problems. It collects the data, integrates and obtains different attributes like traffic events, weather data and time. Using US-Accident data they perform a comparison between DAP (Deep Accident Prediction) and traditional model and resulting from that extensive model are more appropriate than the traditional approach.

In this paper [11], they surveyed on semi-automated vehicles factors that are affecting the crashes and accident. Using negative binomial regression model to decrease number of accident. As the ADT dataset are expensive and not available in whole road instead they taken sub sample dataset of same period of time. Then applying spatial; regression on the traffic accident dataset. So the result is that driver should drive more carefully to reduce the accident.

In this paper [12], stated that the spatial data is deal with the patchy data. For example, in population data the destruction is having hard edges which is patchy and have gaps in it. The comparative studies say that the approach used in spatially referenced data has ecologically utilize these attribute.

In this paper [13], it is a comparative study by seeing the condition of Bangladesh. So many accidents caused in Bangladesh due to heavy traffic applying following supervised machine learning technique they are decision tree, KNN, naive Bayes and ad boost. So it is observed that most of accident occur due to surface like if the surface is dry no accident occur and if the surface is wet they chance are more for the accident.

In this paper [14], stated that road transportation is the major form of transportation from one country to another there are few factors involve that are responsible for these accident. Using data mining approaches find the relation between these factor and perform which algorithm perform well. They basically use Decision tree, KNN, and Naïve Bayes classifier. It resulting as the KNN is the best among the other two algorithms using the confusion matrix they conclude the results.

In this paper [15], stated that the rate of accident occur in 1 hrs. in the major cities can be predicted using some machine learning algorithms. Every paper represent the factor involve in the accident and area affected but they time zoned it. They use balanced random forest and random forest algorithm. By comparing these two using various factors they concluded that random forest better than balanced random forest.

In this paper [16] the high-density areas of accident has been identified. For this GIS and KDE methodologies has been used to study the spatial patterns of injury related road accidents. And K-means clustering methodology is used for creating a classification of road accident hotspots in London and UK. Five groups and 15 clusters were created based on collision and attribute data. These clusters are discussed and evaluated according to their robustness and potential uses in road safety campaigning.

In this paper[17] they have surveyed the spanish road accident data and found the injury severity involving novice drivers in urban areas. The information root node variation (IRNV) method (based on decision trees) was used to get a rule set that provides useful information about the most probable causes of fatalities in accidents involving inexperienced drivers in urban areas. This method is based on the decision tree classifier. These rules provide useful knowledge in order to prevent these kinds of accidents.

In this article[18] the zones, roads and specific time in the CDMX in which the largest number of road traffic accidents are concentrated during 2016 has been identified. A database compiling information obtained from the social network known as Waze is built. The methodology Discovery of knowledge in the database (KDD) for the discovery of patterns in the accidents reports was used. The Maximization of Expectations (EM) algorithm was used to obtain the number ideal of clusters for the data and k-means was used for grouping method.

In this study[19], the classification of road accident on the basis of road user category has been performed. Self Organizing Map (SOM), K-modes clustering technique has been used to group the data into homogeneous segments and then Support vector machine (SVM), Naive Bayes (NB) and Decision tree models has been applied to classify the data. The classification has been performed on data with and without clustering.

In this paper[20] a learning model has been proposed to overcome the challenges that the decision makers face in order to encounter a huge number of resulting association rules that can make them unable to choose and decide rationally between these different extracted rules. The learning model is based on based on FP-growth algorithm using Apache Spark framework, in order to analyze data and extract interesting association rules by taking into account some quality measures.

IV. Summary:

S.No	Title	Models Used	Summary	Author
1	A Data Mining Framework to Analyze Road Accident Data using Map Reduce Methods CCMF and TCAMP Algorithms	CCMF and TCAMP	The authors have used Hadoop to analyze the large data based on various criteria to predict road accidents, in order to raise precaution alarms for the same. On comparing Hadoop with other methods, it has proved to be the most efficient method for big data analysis	S. Nagendra Babu, J. Jebamalar Tamilselvi
2	ACCIDENT PREDICTION BASED ON ACCIDENT TYPES USING SPATIOTEMPORAL CLUSTERING ALGORTIHMS	DBSCAN, STING and CLIQUE	This paper focuses on finding and predicting road accident patterns based on the severity, road type, accident type, climate, hour of accident etc. The method of finding interesting and useful patterns from spatial database is termed as spatial data mining. Spatiotemporal algorithms are able to locate hidden patterns more easily than traditional data mining techniques	Dara Anitha Kumari, Dr. A. Govardhan
3	Learning Deep Representation from Big and Heterogeneous Data for Traffic Accident Inference	Stack de-noise auto encoder model	Huge amount of heterogeneous data containing accident data and GPS records have been collected, to check how vehicle mobility affect the road accidents. On analyzing this data a Stack de-noise auto encoder model is prepared which studies the features in human mobility to predict accident risks	Quanjun Chen, Xuan Song, Harutoshi Yamada, Ryosuke Shibasaki
4	A Radical Approach to Forecast the Road Accident Using	Bayes Net, j48 graft and j48 decision tree	It has been thus noticesd that the combination of such algorithms render better results than a single algorithm used. The results obtained would	Anupama Makkar, Harpreet Singh Gill

	Data Mining Technique		support in forecasting road traffic accidents and hence prevention and control can be provided.	
5	Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study	support vector machine, deep neural networks, random forest and decision tree and Eigen Analysis	The crucial issue of predicting road traffic and the accidents caused due to it has been addressed in this paper [5]. The sparse and heterogeneous data have been worked upon by various algorithms to find interesting patterns. Through a case study, this paper explores various algorithms to predict the amount of accidents occurring every hour	Xun Zhou, Tianbao Yang, Zhuoning Yuan, James Tamerius, Ricardo Mantilla
6	Data Mining Methods for Traffic Accident Severity Prediction	Decision trees, SVM and ANN	In this paper, the features causing road traffic accidents were detected using three techniques of classifications; which are: Decision trees, SVM and ANN. Based on these algorithms a prediction model is built, testing all these algorithms on real time data set.	Qasem A. Al-Radaideh and Esraa J. Daoud
7	Development of Accident Prediction Model on Horizontal Curves	Regression analysis	Accident Prediction Model is proposed by the authors in the paper [7], where accidents on horizontal curves are anticipated based on the patterns examined in the dataset. This prediction model takes measures to lessen the accidents at some extent.	Jerry Soman, Jisha Akkara
8	Study of Road Accident Prediction Model at Accident Blackspot Area: A Case Study at Selangor	Multiple Linear Regression, Artificial Neural Networks	After conducting the analysis it is concluded that the predicted values from regression model had greater errors. While the predictions made from Artificial Neural Network analysis was more accurate having less errors. Hence ANN was proved to be a better methodology to make predictions for accidents.	Haikal Aiman Hartika, Mohd Zakwan Ramli, Muhamad Zaihafiz Zainal Abidin, Mohd Hafiz Zawawi
9	Traffic Risk Mining From Heterogeneous Road Statistics	advanced feature based non-negative matrix factorization (FNMF).	This framework is successful in predicting the traffic risks at any road or intersection more accurately than the existing algorithms or methods. Two clusters were defined that segregate the risk locations, in which one cluster had larger roads with accident risks, and other cluster having higher risk of vehicle collision. Risky locations were ranked based on the results of the	Koichi Moriya, Shin Matsushima, Kenji Yamanishi

			clusters.	
10	Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights (2019)	Feature Vector Representation, Deep Accident Prediction (DAP) Model	it states that reducing traffic accidents is a necessary concern for safety. Small datasets, depending on large data and not valid for real-time data are basic demerits of previous studies. The deep neural network model is used to provide a solution for real-time problems. It collects the data, integrates and obtains different attributes like traffic events, weather data and time. Using US-Accident data they perform a comparison between DAP (Deep Accident Prediction) and traditional model and resulting from that extensive model are more appropriate than the traditional approach.	Sobhan Moosavi, Mohammad Hossein, Srinivasan Parthasarathy, Radu Teodorescu, Rajiv Ramnath
11	Spatial prediction of traffic accidents with critical driving events – Insights from a nationwide field study	Spatial regression	they surveyed on semi-automated vehicles factors that are affecting the crashes and accident. Using negative binomial regression model to decrease number of accident. As the ADT dataset are expensive and not available in whole road instead they taken sub sample dataset of same period of time. Then applying spatial; regression on the traffic accident dataset. So the result is that driver should drive more carefully to reduce the accident.	Benjamin Rydera,*, Andre Dahlingerb, Bernhard Gahrb, Peter Zundritscha, Felix Wortmannb, Elgar Fleischa
12	Twenty years and counting with ADIE: Spatial Analysis by Distance Indices software and review of its adoption and use	Ia, index of aggregation, Patch and gap cluster indices and Red blue Plot	It stated that the spatial data is deal with the patchy data. For example, in population data the destruction is having hard edges which is patchy and have gaps in it. The comparative studies say that the approach used in spatially referenced data has ecologically utilize these attribute.	Linton Winder1, Colin Alexander2, Georgianne Griffiths3, John Holland4, Chris Woolley5, Joe Perry6
13	Road Accident Analysis and Prediction of	Decision Tree, KNN, Naïve Bayes	it is a comparative study by seeing the condition of Bangladesh. So many	Md. Farhan Labib, Ahmed Sady Rifat, Md.

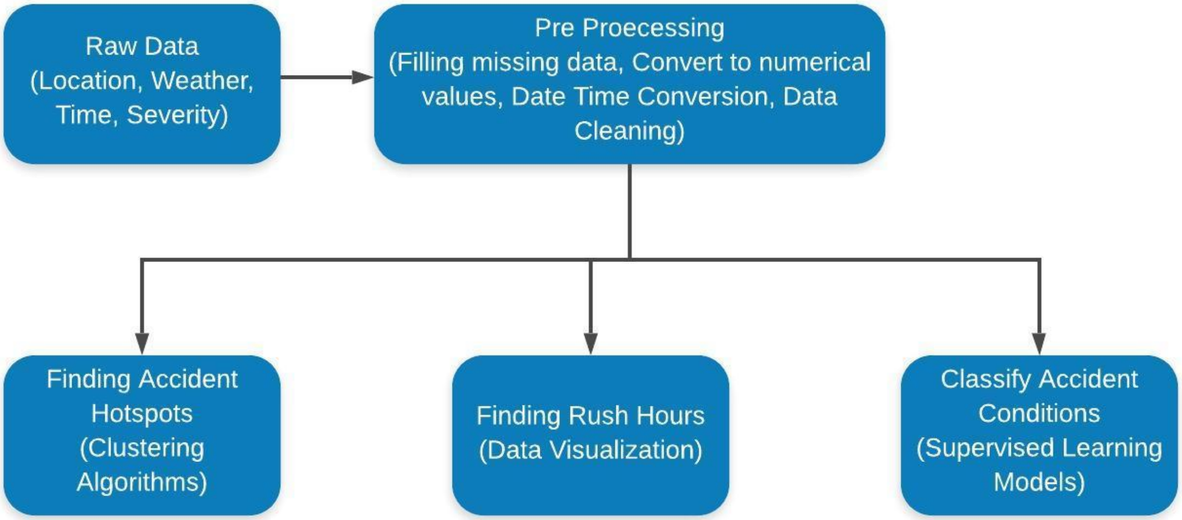
	Accident Severity by Using Machine Learning in Bangladesh	and Ada-Boost	accidents caused in Bangladesh due to heavy traffic applying following supervised machine learning technique they are decision tree, KNN, naive Bayes and ad boost. So it is observed that most of accident occur due to surface like if the surface is dry no accident occur and if the surface is wet they chance are more for the accident.	Mosabbir Hossain, Amit Kumar Das, Faria Nawrine
14	Analyzing Road Accident Criticality using Data mining	KNN, Decision Tree and Naïve Bayes	It stated that road transportation is the major form of transportation from one country to another there are few factors involve that are responsible for these accident. Using data mining approaches find the relation between these factor and perform which algorithm perform well. They basically use Decision tree, KNN, and Naïve Bayes classifier. It resulting as the KNN is the best among the other two algorithms using the confusion matrix they conclude the results.	Shahsitha Siddique V*, Nithin Ramakrishnan
15	High-Resolution Road Vehicle Collision Prediction for the City of Montreal	Balanced random forest, and random forest	It stated that the rate of accident occur in 1 hrs. in the major cities can be predicted using some machine learning algorithms. Every paper represent the factor involve in the accident and area affected but they time zoned it. They use balanced random forest and random forest algorithm. By comparing these two using various factors they concluded that random forest better than balanced random forest.	Antoine H´ebert_, Timoth´ee Gu´edon_, Tristan Glatard, Brigitte Jaumard
16	Kernel density estimation and K-means clustering to profile road accident hotspots	Geographical Information System, Kernel Density Estimation, K-means Clustering	In this paper the high-density areas of accident has been identified. For this GIS and KDE methodologies has been used to study the spatial patterns of injury related road accidents. And K-means clustering methodology is used for creating a classification of road accident hotspots in London and UK. Five groups and 15 clusters were created based on collision and attribute	Tessa K. Anderson

			data. These clusters are discussed and evaluated according to their robustness and potential uses in road safety campaigning.	
17	Decision Tree Ensemble Method for Analyzing Traffic Accidents of Novice Drivers in Urban Areas	Decision Tree	In this paper they have surveyed the spanish road accident data and found the injury severity involving novice drivers in urban areas. The information root node variation (IRNV) method (based on decision trees) was used to get a rule set that provides useful information about the most probable causes of fatalities in accidents involving inexperienced drivers in urban areas. This method is based on the decision tree classifier. These rules provide useful knowledge in order to prevent these kinds of accidents.	Serafín Moral-García , Javier G. Castellano , Carlos J. Mantas , Alfonso Montella and Joaquín Abellán
18	Road Traffic Accidents Analysis in Mexico City through Crowdsourcing Data and Data Mining Techniques	KDD, EM, K-Means	In this article the zones, roads and specific time in the CDMX in which the largest number of road traffic accidents are concentrated during 2016 has been identified. A database compiling information obtained from the social network known as Waze is built. The methodology Discovery of knowledge in the database (KDD) for the discovery of patterns in the accidents reports was used. The Maximization of Expectations (EM) algorithm was used to obtain the number ideal of clusters for the data and k-means was used for grouping method.	Gabriela V. Angeles Perez, Jose Castillejos Lopez, Araceli L. Reyes Cabello, Emilio Bravo Grajales, Adriana Perez Espinosa, Jose L. Quiroz Fabian
19	Road-User Specific Analysis of Traffic Accident Using Data Mining Techniques	SOM, K-Mode, SVM, Naive Bayes, Decision Tree	In this study, the classification of road accident on the basis of road user category has been performed. Self Organizing Map (SOM), K-modes clustering technique has been used to group the data into homogeneous segments and then Support vector machine (SVM), Naive Bayes (NB)	Prayag Tiwari, Sachin Kumar, and Denis Kalitin

			and Decision tree models has been applied to classify the data. The classification has been performed on data with and without clustering.	
20	Data mining for road accident analysis in a big data context	FP-growth algorithm, Apache Spark framework	In this paper a learning model has been proposed to overcome the challenges that the decision makers face in order to encounter a huge number of resulting association rules that can make them unable to choose and decide rationally between these different extracted rules. The learning model is based on based on FP-growth algorithm using Apache Spark framework, in order to analyze data and extract interesting association rules by taking into account some quality measures.	Fatima Zahra El Mazouri, Mohammed Chaouki Abounaima, Said Najah, Khalid Zenkouar

V. Proposed Work:

The following image describes the proposed framework that is being followed in the project,


Raw Data
(Location, Weather, Time, Severity)
Pre Proecessing
(Filling missing data, Convert to numerical values, Date Time Conversion, Data Cleaning)
Finding Accident Hotspots
(Clustering Algorithms)
Finding Rush Hours
(Data Visualization)
Classify Accident Conditions
(Supervised Learning Models)

In the work first we preprocessed the data then we find accident hotspot using clustering algorithm, rush hour from data visualization and finally classify the accident severity using supervised learning algorithms based on the weather conditions.

A. Data preprocessing:

In the data preprocessing stage, the dataset is analyzed to find the missing values and handling them as such, by either filling up with suitable values or 0. Python’s ‘numpy’ class is used to handle the missing data. Few columns are discarded in order to refine the dataset for better predictions, as unnecessary columns might lead to deviation from proper predictions. The required features are hence chosen for carrying out the prediction functionalities. Categorical data is also encoded into numerical data for the sake of calculations and for fitting purpose in the machine learning models. LabelEncoder class of scikit learn is used for the process.

Code:

```
import pandas as pd
import numpy as np
import sklearn
df=pd.read_csv("data.csv")
df.head()
df.columns
df.isnull().sum()
df1=df
df1['Start_Time'] = pd.to_datetime(df1['Start_Time'], format = '%Y/%m/%d
%H:%M:%S')
df1['Start_Time'] = pd.to_datetime(df1['Start_Time'],errors='coerce')
column_1=df1.ix[:,4]
db=pd.DataFrame({"year": column_1.dt.year,
              "month": column_1.dt.month,
              "day": column_1.dt.day,
              "hour": column_1.dt.hour,
             })
df1=df1.drop('Start_Time',axis=1)
df1=pd.concat([db,df1],axis=1)
df1=df1.drop(['TMC', 'End_Time', 'End_Lat', 'End_Lng', 'Distance(mi)',
'Description', 'Number', 'Timezone', 'Airport_Code', 'Weather_Timestamp'],
axis=1)
df1.columns
df1.isnull().sum()
df1.shape
df1=df1.dropna(how='any', subset=['City', 'Zipcode', 'Weather_Condition',
'Sunrise_Sunset', 'Civil_Twilight', 'Wind_Direction', 'Nautical_Twilight',
'Astronomical_Twilight'])
df1.shape
df1.isnull().sum()
from sklearn.linear_model import LinearRegression
model=LinearRegression()
df1.year=df1.year.astype('float64')
df1.month=df1.month.astype('float64')
df1.day=df1.day.astype('float64')
df1.hour=df1.hour.astype('float64')
d=df1
def regression_fill(df1, y):
  print("doing for ", y)
  df2=df1[df1[y].isnull()]
  df1=df1.dropna(subset=[y])
  model.fit(df1[['year', 'month', 'day', 'hour', 'Start_Lat',
'Start_Lng']], df1[y])
  Y=model.predict(df2[['year', 'month', 'day', 'hour', 'Start_Lat',
'Start_Lng']])
  df2=df2.drop([y], axis=1)
  df2[y]=Y
  df1=pd.concat([df1, df2])
  return df1
targets=['Temperature(F)', 'Wind_Chill(F)', 'Humidity(%)', 'Pressure(in)',
'Visibility(mi)', 'Wind_Speed(mph)']
for t in targets:
  df1=regression_fill(df1, t)
df1['Precipitation(in)'].fillna(0, inplace=True)
df1.isnull().sum()
df1.dtypes
from sklearn.preprocessing import LabelEncoder
encoder=LabelEncoder()
for c in df1.columns:
  if df1[c].dtype=='object' or df1[c].dtype=='bool':
    df1[c]=encoder.fit_transform(df1[c]).astype('int32')
df1.head()
df1.dtypes
df1=df1.sort_values(by=['ID']).reset_index(drop=True)
df1.shape
df1.head()
df1.to_csv("preprocessed_data.csv")
```

B. Finding Rush Hours:

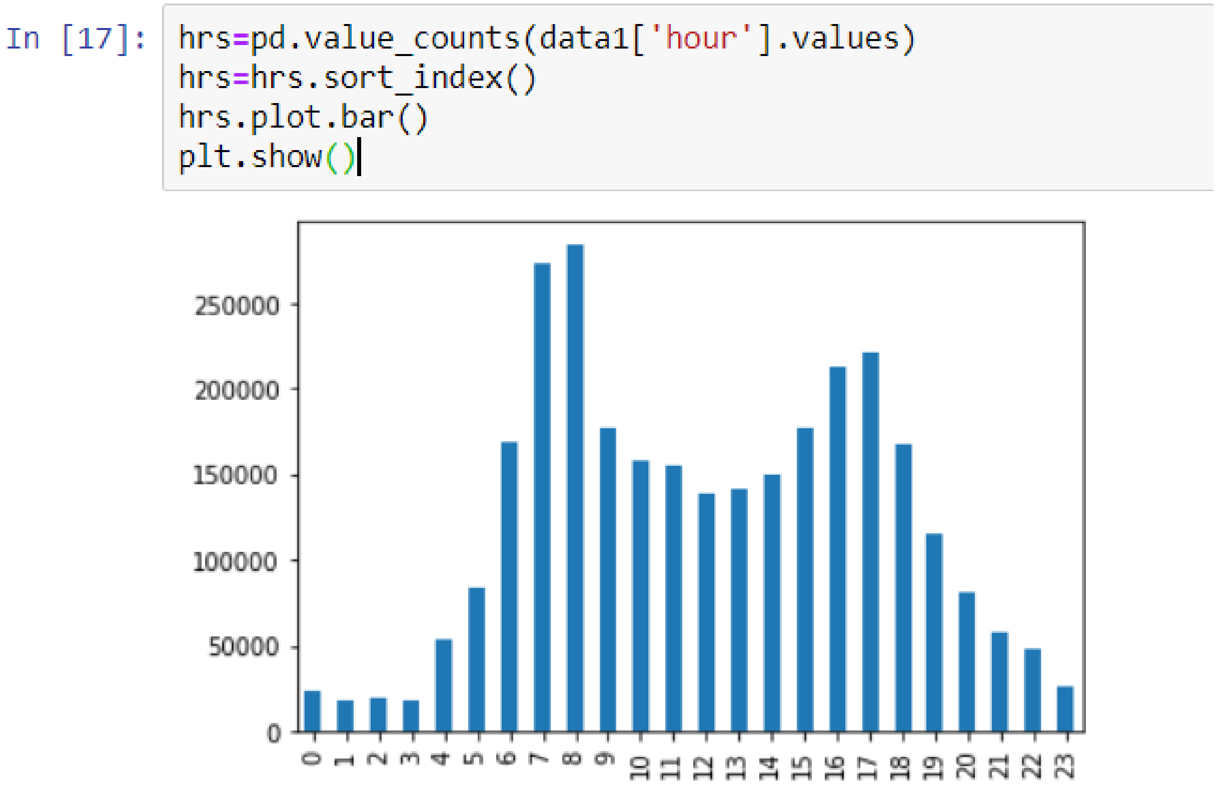
Rush hours are the times that have potential of accidents. We have located the rush hours using visualization methods. We have used **seaborn** and **matplotlib** libraries to visualize the data and locate the rush hours. We have used **matplotlib's barplot** to visualize and find the rush hours. The bar plot have used the number of accidents occurred, and the time in terms of 24 hours to find the rush hour. The results thus obtained are satisfactory and accurate.

Code:

```
df['Start_Time'] = pd.to_datetime(df['Start_Time'], format = '%Y-%m-%d
%H:%M:%S')
column_1=df.ix[:,4]
column_1.head()
db=pd.DataFrame({"year": column_1.dt.year,
              "month": column_1.dt.month,
              "day": column_1.dt.day,
              "hour": column_1.dt.hour,
                })
dataset=df
dataset1=dataset.drop('Start_Time',axis=1)
data1=pd.concat([db,dataset1],axis=1)
data1.head()
hrs=pd.value_counts(data1['hour'].values)
```

```
hrs=hrs.sort_index()
hrs.plot.bar()
plt.show()
```

Output:



```
In [17]: hrs=pd.value_counts(data1['hour'].values)
         hrs=hrs.sort_index()
         hrs.plot.bar()
         plt.show()
```

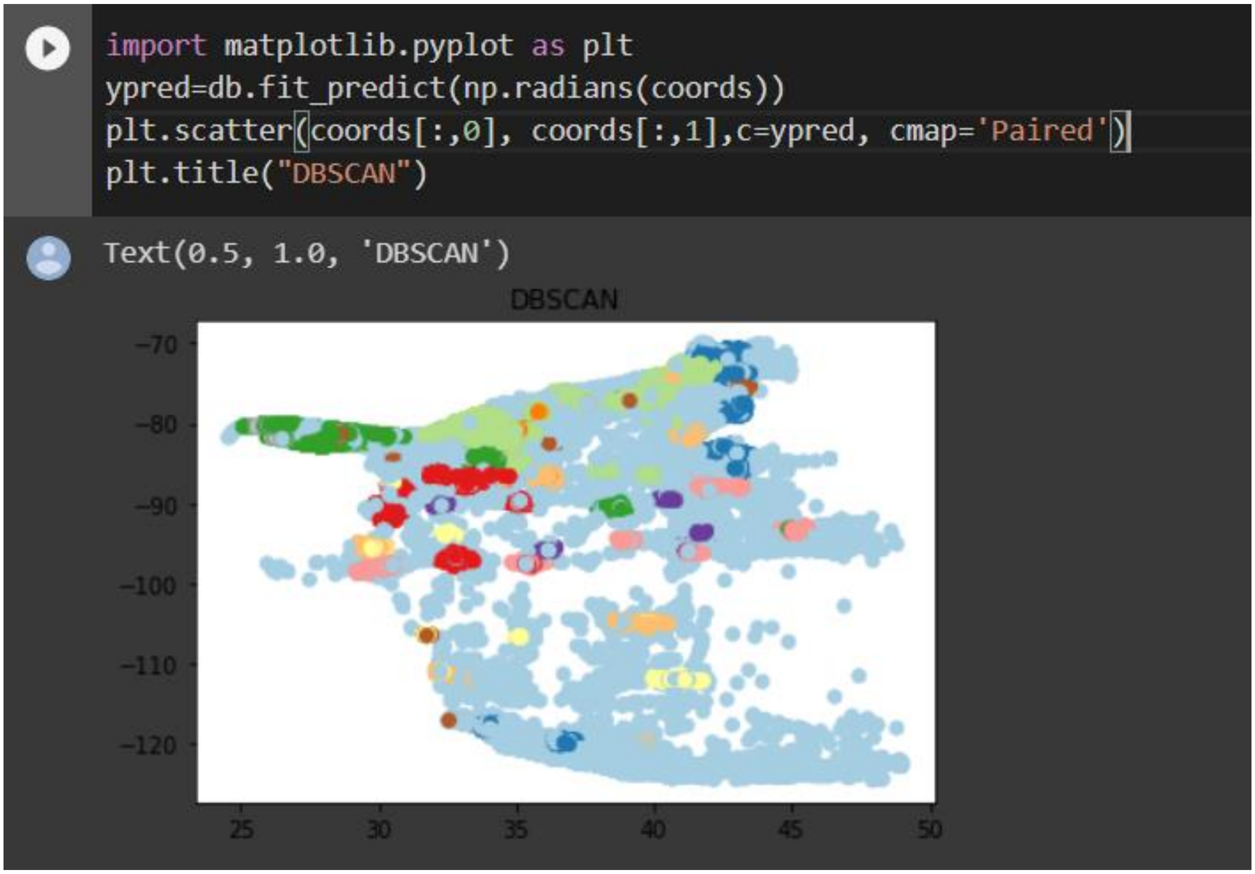
C. Finding Accident hotspots:

To find the accident hotspot, we have used the DBSCAN clustering algorithm. **DBSCAN** is a **density based clustering algorithm**, it does a great job of seeking areas in the data that have a high density of observations, versus areas of the data that are not very dense with observations. It mainly uses two parameters: ‘eps’ or the epsilon value that defines the neighborhood around a data point .If the distance between two points is lower or equal to ‘eps’ then they are considered as neighbors. We have taken **36.6 km area as per earth’s radius** to determine the epsilon value; the other parameter is min points which define the minimum number of data points in a neighborhood. We have used the “Euclidean distance formula” to calculate the distance between two points in order to determine neighborhood. Thus the clusters formed show high densities of accidents which might occur on the basis of turning point, bumps, crossings, etc.

```
from sklearn.cluster import DBSCAN
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import normalize
from sklearn.decomposition import PCA
from geopy.distance import great_circle
from shapely.geometry import MultiPoint
```

```
coords=df1[['Start_Lat',
'Start_Lng','Amenity','Bump','Crossing','Give_Way','Junction','No_Exit','R
ailway','Roundabout','Station','Stop','Traffic_Calming','Traffic_Signal','
Turning_Loop']].to_numpy()

coords=coords[0:297000]
print(coords)
kms_per_radian = 6371
epsilon = 36.6 / kms_per_radian
db = DBSCAN(eps=epsilon, min_samples=250, algorithm='ball_tree',
metric='euclidean').fit(np.radians(coords))
cluster_labels = db.labels_
num_clusters = len(set(cluster_labels))
clusters = pd.Series([coords[cluster_labels == n] for n in
range(num_clusters)])
print('Number of clusters: {}'.format(num_clusters))
import matplotlib.pyplot as plt
ypred=db.fit_predict(np.radians(coords))
plt.scatter(coords[:,0], coords[:,1],c=ypred, cmap='Paired')
plt.title("DBSCAN")
```

Output:



```
import matplotlib.pyplot as plt
ypred=db.fit_predict(np.radians(coords))
plt.scatter(coords[:,0], coords[:,1],c=ypred, cmap='Paired')
plt.title("DBSCAN")
```

```
Text(0.5, 1.0, 'DBSCAN')
```

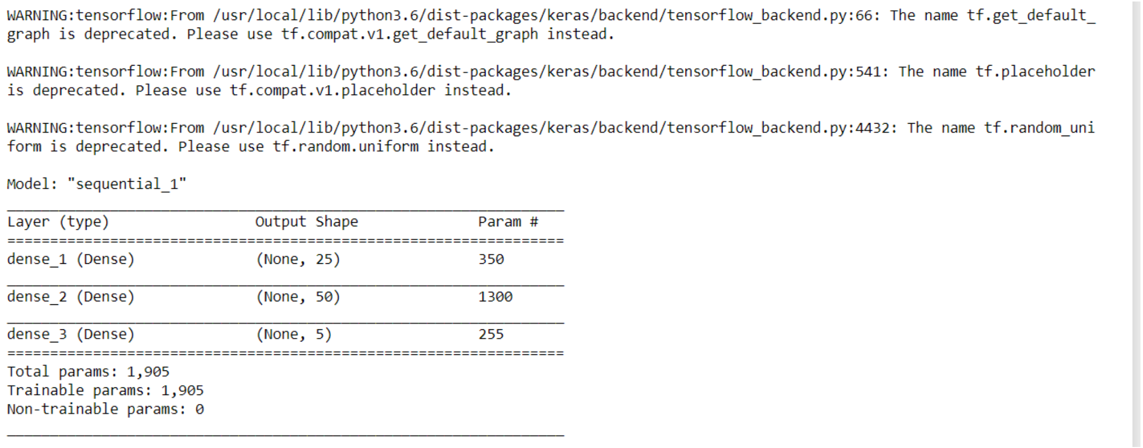
D. Classify the accident severity based on weather conditions:

For the classification purpose we have used supervised learning algorithm multilayer perceptron neural networks. The network is made of 3 fully connected layer with activation function “sigmoid”. For the first layer i.e. the input layer 25 neurons are used; in the second i.e. in the hidden layer 50 neurons are used and in the final layer i.e. the output layer is consisting of only 5 layers. For optimizing we have used SGD (Stochastic Gradient Descent) Optimizer. The model is trained for 10 epochs with batch size of 32.

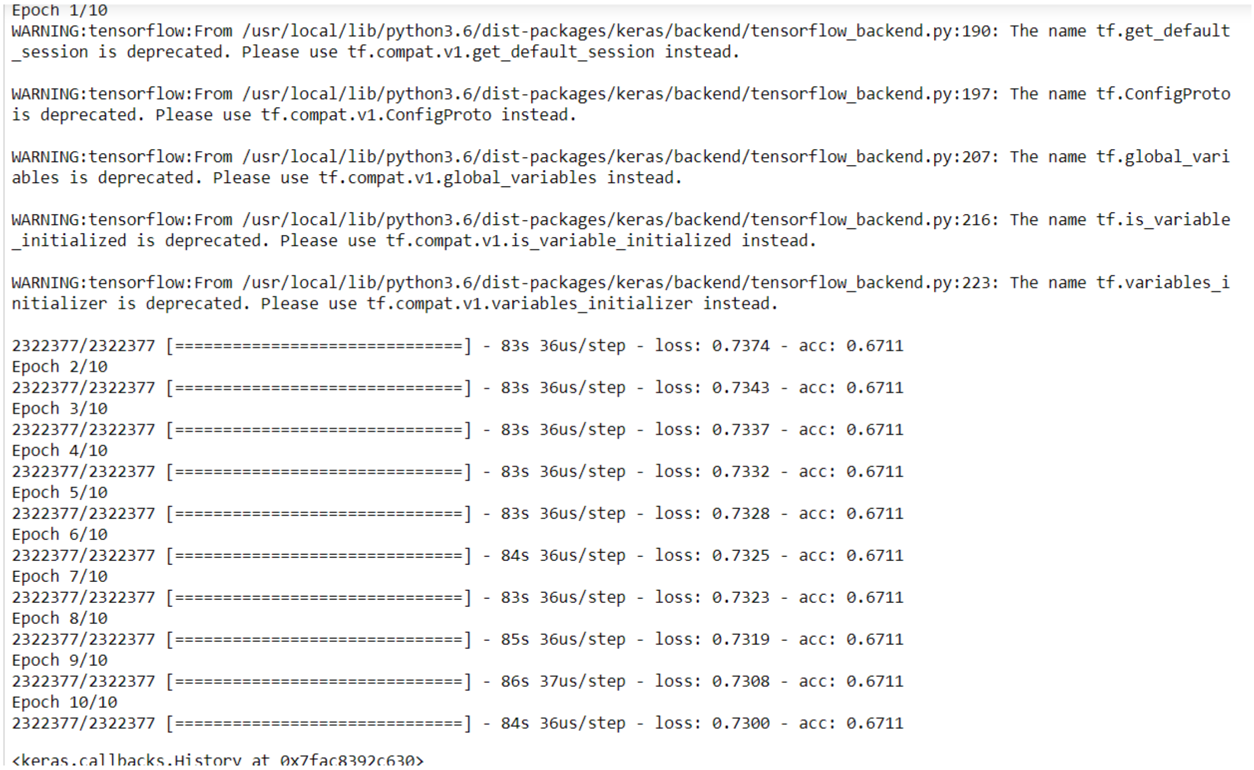
Code:

```
import pandas as pd
import numpy as np
import sklearn
from keras.optimizers import SGD, RMSprop, Adadelta, Adagrad, Adam, Adamax
from keras.models import Sequential
from keras.layers.core import Dense, Activation, Flatten
import keras
from sklearn.model_selection import train_test_split
from keras.utils import to_categorical
df=pd.read_csv('preprocessed_data.csv')
df.head()
df.columns
data=df[['Astronomical_Twilight', 'Civil_Twilight', 'Humidity(%)',
'Nautical_Twilight', 'Precipitation(in)', 'Pressure(in)',
'Sunrise_Sunset', 'Temperature(F)', 'Visibility(mi)', 'Weather_Condition',
'Wind_Chill(F)', 'Wind_Direction', 'Wind_Speed(mph)', 'Severity']].values
x=data[:, 0: 13]
y=data[:, 13]
xtrain, xtest, ytrain, ytest=train_test_split(x, y, test_size=.2)
ytrain = to_categorical(ytrain.astype('float32'))
model = Sequential()
model.add(Dense(25, input_dim=13, activation= "sigmoid"))
model.add(Dense(50, activation= "sigmoid"))
model.add(Dense(5, activation="sigmoid"))
model.summary()
opt = SGD()
model.compile(loss='categorical_crossentropy', optimizer=opt,
metrics=['accuracy'])
model.fit(xtrain, ytrain, nb_epoch=10, batch_size=32)
model.save("train2.model")
ytest = to_categorical(ytest.astype('float32'))
score, acc = model.evaluate(xtest, ytest)
print(score)
print(acc*100)
d=df[['Astronomical_Twilight', 'Civil_Twilight', 'Humidity(%)',
'Nautical_Twilight', 'Precipitation(in)', 'Pressure(in)',
'Sunrise_Sunset', 'Temperature(F)', 'Visibility(mi)', 'Weather_Condition',
'Wind_Chill(F)', 'Wind_Direction', 'Wind_Speed(mph)', 'Severity']]
d.head()
predx=np.array([[1, 0, 56.0, 0, 0.03, 30.09, 0, 90.78, 7.0, 7, 23.90, 7,
8.9]])
ypred=model.predict(predx).reshape(-1)
print("calss : ", np.argmax(ypred))
```

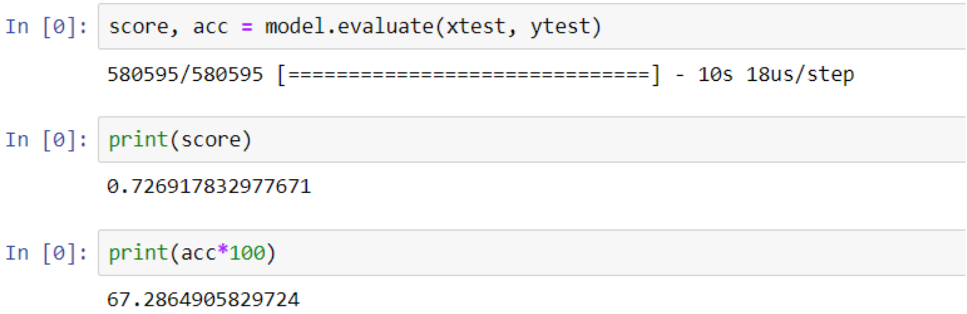
Output:



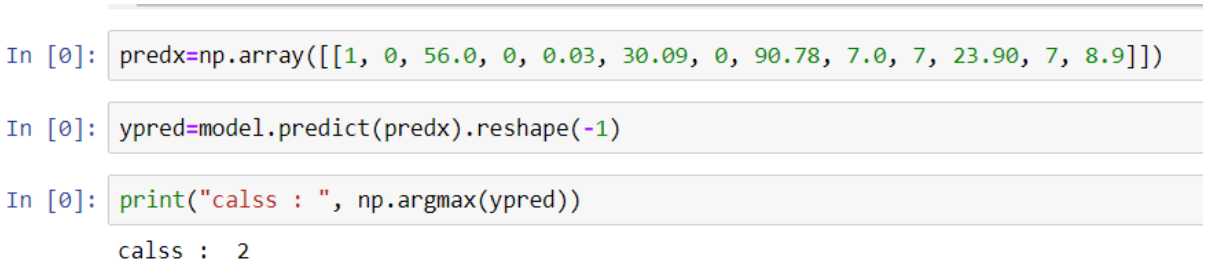
```
WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/keras/backend/tensorflow_backend.py:66: The name tf.get_default_
graph is deprecated. Please use tf.compat.v1.get_default_graph instead.

WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/keras/backend/tensorflow_backend.py:541: The name tf.placeholder
is deprecated. Please use tf.compat.v1.placeholder instead.

WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/keras/backend/tensorflow_backend.py:4432: The name tf.random_uni
form is deprecated. Please use tf.random.uniform instead.

Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
dense_1 (Dense)              (None, 25)                350       
_________________________________________________________________
dense_2 (Dense)              (None, 50)                1300      
_________________________________________________________________
dense_3 (Dense)              (None, 5)                 255       
=================================================================
Total params: 1,905
Trainable params: 1,905
Non-trainable params: 0
_________________________________________________________________
```



```
Epoch 1/10
WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/keras/backend/tensorflow_backend.py:190: The name tf.get_default
_session is deprecated. Please use tf.compat.v1.get_default_session instead.

WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/keras/backend/tensorflow_backend.py:197: The name tf.ConfigProto
is deprecated. Please use tf.compat.v1.ConfigProto instead.

WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/keras/backend/tensorflow_backend.py:207: The name tf.global_vari
ables is deprecated. Please use tf.compat.v1.global_variables instead.

WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/keras/backend/tensorflow_backend.py:216: The name tf.is_variable
_initialized is deprecated. Please use tf.compat.v1.is_variable_initialized instead.

WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/keras/backend/tensorflow_backend.py:223: The name tf.variables_i
nitializer is deprecated. Please use tf.compat.v1.variables_initializer instead.

2322377/2322377 [==============================] - 83s 36us/step - loss: 0.7374 - acc: 0.6711
Epoch 2/10
2322377/2322377 [==============================] - 83s 36us/step - loss: 0.7343 - acc: 0.6711
Epoch 3/10
2322377/2322377 [==============================] - 83s 36us/step - loss: 0.7337 - acc: 0.6711
Epoch 4/10
2322377/2322377 [==============================] - 83s 36us/step - loss: 0.7332 - acc: 0.6711
Epoch 5/10
2322377/2322377 [==============================] - 83s 36us/step - loss: 0.7328 - acc: 0.6711
Epoch 6/10
2322377/2322377 [==============================] - 84s 36us/step - loss: 0.7325 - acc: 0.6711
Epoch 7/10
2322377/2322377 [==============================] - 83s 36us/step - loss: 0.7323 - acc: 0.6711
Epoch 8/10
2322377/2322377 [==============================] - 85s 36us/step - loss: 0.7319 - acc: 0.6711
Epoch 9/10
2322377/2322377 [==============================] - 86s 37us/step - loss: 0.7308 - acc: 0.6711
Epoch 10/10
2322377/2322377 [==============================] - 84s 36us/step - loss: 0.7300 - acc: 0.6711

<keras.callbacks.History at 0x7fac8392c630>
```



```
In [0]: score, acc = model.evaluate(xtest, ytest)

        580595/580595 [==============================] - 10s 18us/step

In [0]: print(score)

        0.726917832977671

In [0]: print(acc*100)

        67.2864905829724
```



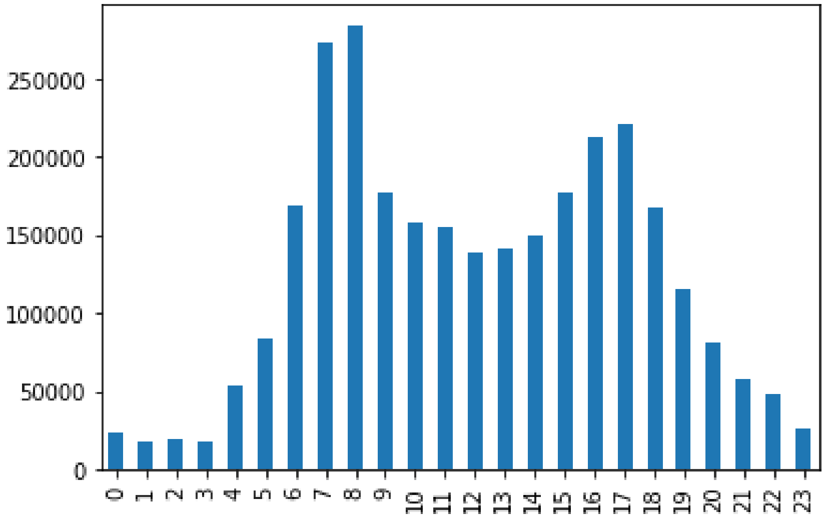
```
In [0]: predx=np.array([[1, 0, 56.0, 0, 0.03, 30.09, 0, 90.78, 7.0, 7, 23.90, 7, 8.9]])

In [0]: ypred=model.predict(predx).reshape(-1)

In [0]: print("calss : ", np.argmax(ypred))

        calss :  2
```

VI. Experimentation and Results:

1. Dataset Description:

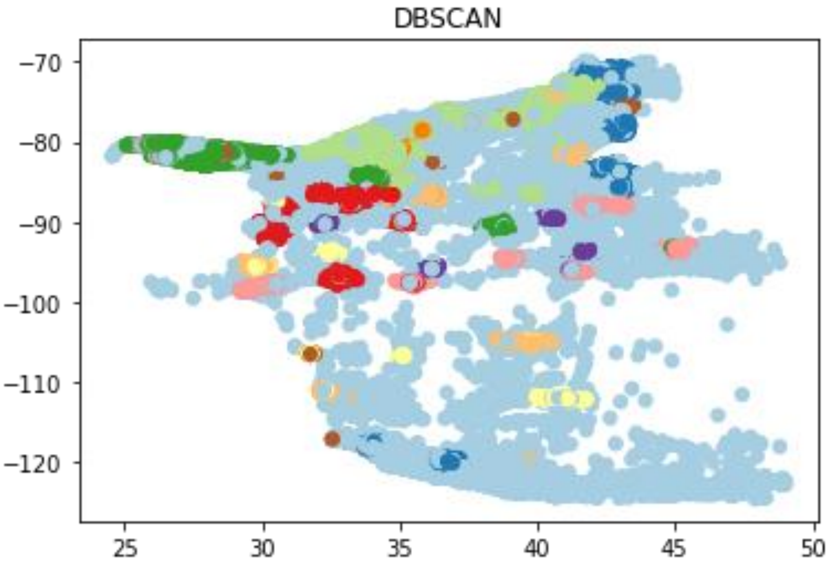
The dataset obtained from Kaggle is a countrywide traffic accident dataset, which covers 49 states of the United States. The data is collected from February 2016 to December 2019, using several data providers, including two APIs which provide streaming traffic event data. These APIs broadcast traffic events captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about 3.0 million accident records in this dataset. The dataset has 49 features as; ID , Source, TMC, Severity, Start_Time , End_Time , Start_Lat, Start_Lng, End_Lat, End_Lng, Distance(mi), Description, Number, Street, Side, City, County, State, Zipcode,Country, Timezone, Airport_Code, Weather_Timestamp, Temperature(F), Wind_Chill(F), Humidity(%),Pressure(in), Visibility(mi), Wind_Direction, Wind_Speed(mph), Precipitation(in), Weather_Condition, Amenity, Bump, Crossing, Give_Way, Junction, No_Exit, Railway, Roundabout, Station, Stop, Traffic_Calming, Traffic_Signal, Turning_Loop, Sunrise_Sunset, Civil_Twilight,Nautical_Twilight, Astronomical_Twilight.

2. Finding Rush Hours:



This image is obtained from the data visualizations. So from here we can see that most of the accidents occur in between 7 am to 9 am in the morning and 4pm to 6pm in the evening. So we can conclude that the morning 7-9 hours and the evening 16-18 hours are the rush hours for accidents.

3. Finding Accident Hotspots:



This accident hotspots are generated in the picture based on the road conditions. The various colors show different clusters having same densities based on road conditions, areas, like public sector or nearby to airports and stations and latitude and longitude, on which the accidents might occur. The blue marked areas cluster on similar conditions where accident can occur; the dark green ones have different densities as such.

4. **Classify the accident severity based on weather conditions:**

   In this we are getting almost 68% accuracy on predicting the severity of the accident based on the various weather conditions. So the model is giving good prediction.

VII. Conclusion:

So, we can conclude that the rush hours are the office times when the traffics on the road are maximum. The Hotspots are the places where the public service sectors like railway station and airports are nearby. Also, the severity classification deep learning model is working good with a good accuracy. So, by our work we can predict whether there will be an accident or not and if the accident occurs how much sever it may be. So necessary measures should be taken to prevent these accidents.

VIII. References:

[1] S. Nagendra Babu, J. Jebamalar Tamilselvi," A Data Mining Framework to Analyze Road Accident Data using Map Reduce Methods CCMF and TCAMP Algorithms" , IJSSST, 2013

[2] Dara Anitha Kumari, Dr. A. Govardhan," ACCIDENT PREDICTION BASED ON ACCIDENT TYPES USING SPATIOTEMPORAL CLUSTERING ALGORTIHMS", International Journal of Pure and Applied Mathematics Volume 120 No. 6 , 2018

[3] Quanjun Chen, Xuan Song, Harutoshi Yamada, Ryosuke Shibasaki," Learning Deep Representation from Big and Heterogeneous Data for Traffic Accident Inference", Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence

[4] Anupama Makkar, Harpreet Singh Gill," A Radical Approach to Forecast the Road Accident Using Data Mining Technique", International Journal of Innovative Science and Research Technology ISSN No: - 2456 – 2165 Volume 2, Issue 8,2017

[5] Xun Zhou, Tianbao Yang, Zhuoning Yuan, James Tamerius, Ricardo Mantilla," Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study", Proceedings of 6th International Workshop on Urban Computing, Halifax, Nova Scotia, Canada,2017

[6] Qasem A. Al-Radaideh and Esraa J. Daoud, " Data Mining Methods for Traffic Accident

Severity Prediction”, INTERNATIONAL JOURNAL OF NEURAL NETWORKS and ADVANCED APPLICATIONS,2018

[7] Jerry Soman, Jisha Akkara, ”Development of Accident Prediction Model on Horizontal Curves”, International Research Journal of Engineering and Technology (IRJET) Volume: 06 Issue: 03,2019

[8] Haikal Aiman Hartika, Mohd Zakwan Ramli, Muhamad Zaihafiz Zainal Abidin, Mohd Hafiz Zawawi,” Study of Road Accident Prediction Model at Accident Blackspot Area: A Case Study at Selangor”, International Journal of Scientific Research in Science, Engineering and Technology, 2017

[9] Koichi Moriya, Shin Matsushima, Kenji Yamanishi,” Traffic Risk Mining From Heterogeneous Road Statistics”, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 19, NO,2018

[10] “Sobhan Moosavi, Mohammad Hossein, Srinivasan Parthasarathy, Radu Teodorescu, Rajiv Ramnath” Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights (2019)

[11] “Benjamin Rydera,∗, Andre Dahlingerb, Bernhard Gahrb, Peter Zundritscha, Felix Wortmannb, Elgar Fleischa” Spatial prediction of traffic accidents with critical driving events – Insights from a nationwide field study, 2018

[12] “Linton Winder1, Colin Alexander2, Georgianne Griffiths3, John Holland4, Chris Woolley5, Joe Perry6” Twenty years and counting with ADIE: Spatial Analysis by Distance Indices software and review of its adoption and use,2019

[13] “Md. Farhan Labib, Ahmed Sady Rifat, Md. Mosabbir Hossain, Amit Kumar Das, Faria Nawrine” Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh,2019

[14] “Shahsitha Siddique V*, Nithin Ramakrishnan”, Analyzing Road Accident Criticality using Data mining, 2019

[15] “Antoine H´ebert_, Timoth´ee Gu´edon_, Tristan Glatard, Brigitte Jaumard” High-Resolution Road Vehicle Collision Prediction for the City of Montreal, 2019

[16] Tessa K. Anderson, “Kernel density estimation and K-means clustering to profile road accident hotspots”, 2008

[17] Serafín Moral-García , Javier G. Castellano , Carlos J. Mantas , Alfonso Montella and Joaquín Abellán , “Decision Tree Ensemble Method for Analyzing Traffic Accidents of Novice Drivers in Urban Areas”, 2019

[18] Gabriela V. Angeles Perez, Jose Castillejos Lopez, Araceli L. Reyes Cabello, Emilio Bravo Grajales, Adriana Perez Espinosa, Jose L. Quiroz Fabian, “Road Traffic Accidents Analysis in Mexico City through Crowdsourcing Data and Data Mining Techniques”, 2018

[19] Prayag Tiwari, Sachin Kumar, and Denis Kalitin, “Road-User Specific Analysis of Traffic Accident Using Data Mining Techniques”, 2017

[20] Fatima Zahra El Mazouri, Mohammed Chaouki Abounaima, Said Najah, Khalid Zenkouar, “Data mining for road accident analysis in a big data context”, 2019