Data Analytics on Air Pollution and Traffic Accidents

A Project Report submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology

in

Computer Science and Engineering

by

Atharva Varade(111915021)
Yogesh Grandhi(111915037)
Adarsh Vardhan J(111915007)
Yeshwanth M(111915142)

Under the Supervision of: Dr. Jatin Majithia

Semester: VII

Name of Department: Department of Computer Science and Engineering

Indian Institute of Information And Technology, Pune

(An Institute of National Importance by an Act of Parliament)

DECEMBER 2022

BONAFIDE CERTIFICATE

This is to certify that the project report entitled **"Data Analytics on Air Pollution and traffic accidents"** submitted by **Atharva Varade** bearing the **MIS No: 111915021** , **Yogesh Grandhi** bearing the **MIS No: 111915037**, **Adarsh Vardhan J** bearing the **MIS No: 111915007**,**Yeshwanth Mootakoduru** bearing the **MIS No:111915142** in completion of his/her project work under the guidance of **Supervisor's Name** is accepted for the project report submission in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in the **Department of Computer Science and Engineering**, Indian Institute of Information Technology, Pune (IIIT Pune), during the academic year **2022-23**.

Dr. Jatin Majithia
Project Guide
Assistant Professor
Department of Applied Mathematics And Data Science
IIIT Pune

Dr. Sanjeev Sharma
Head of the Department
Assistant Professor
Department of CSE
IIIT Pune

Project Viva-voce held on 15-12-2022

Undertaking for Plagiarism

We **Atharva Varade, Yogesh Grandhi , Adarsh Vardhan J, Yeshwanth M** solemnly declare that research work presented in the **report** titled **“ Data Analytics on Air Pollution and Traffic Accidents”** is solely **my/our** research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete report/dissertation has been written by **us/me**. I understand the zero tolerance policy of **Indian Institute of Information Technology Pune** towards plagiarism. Therefore **we** declare that no portion of my **report/dissertation** has been plagiarized and any material used as reference is properly referred/cited. I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of the degree, the Institute reserves the right to withdraw/revoke my **B.Tech** degree.

Student’s Name and Signature with Date

Atharva Varade

Yogesh Grandhi

Adarsh Vardhan J

Yeshwanth M

Conflict of Interest

Manuscript title: DATA ANALYTICS ON AIR POLLUTION AND TRAFFIC ACCIDENTS

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Student's Name and Signature with Date

Atharva Varade

Yogesh Grandhi

Adarsh Vardhan J

Yeshwanth M

Problem Statement

The main goal of the project is to analyze the existing datasets and find patterns which identify the factors contributing to deaths due to Air pollution and Traffic accidents and propose ideas to achieve SDG-3 Goal.

Objective

The project aligns with the SDG 3 (by WHO) which aims to ensure healthy lives and promote well-being for all, at all ages. We have mainly focused on the following aspects of SDG:

- By 2030, halve the number of global deaths and injuries from road traffic accidents.
- By 2030, substantially reduce the number of deaths and illnesses from hazardous chemicals and air, water and soil pollution and contamination.

ACKNOWLEDGEMENT

This project would not have been possible without the help and cooperation of many. I would like to thank the people who helped me directly and indirectly in the completion of this project work.
First and foremost, I would like to express my gratitude to our honorable Director, **Prof. O.G. Kakde**, for providing his kind support in various aspects. I would like to express my gratitude to my project guide **Dr. Jatin Majithia**, **Department of CSE**, for providing excellent guidance, encouragement, inspiration, constant and timely support throughout this **B.Tech Project**. I would like to express my gratitude to Dr. **Sanjeev Sharma**, **Department of CSE**, for providing his kind support in various aspects. I would also like to thank all the faculty members in the **Department of CSE** and my classmates for their steadfast and strong support and engagement with this project.

Abstract

Main goal of the project is to analyze the existing datasets and find patterns which identify the factors contributing to deaths due to air pollution and traffic accidents and propose ideas to achieve SDG-3 goal, primary focus was on premature mortality. The project aligns with the SDG 3 (by WHO) which aims to ensure healthy lives and promote well-being for all, at all ages,substantially reduce the number of deaths and illnesses from hazardous chemicals and air, water and soil pollution and contamination.our analysis can be used by the different regulatory institutions to find the mortality based hotspots and pointers suggested by this project can be used to enforce methods and policies to reduce the overall premature mortalities.It is crucial to understand the collective impacts of multiple air pollutants on people's health in order to reduce the mortalities. Analyzing and understanding this huge amount of data can offer a better perspective on the long term effects of air pollution and what could be the most efficient and effective measures to prevent it. The mortality from accidents can be reduced with cautious driving and by following the traffic rules and regulations by all and at all times.

Keywords: LSTM, AQI,BDPA,SP,EP,SFFS,PM 2.5, PM 10

TABLE OF CONTENTS

List of Figures

Chapter 1

Introduction

1.1 Overview of Work

The recent pandemic has severely disrupted essential health services, shortened life expectancy and has worsen inequities in access to basic health services between countries and people, threatening to undo years of progress in healthcare..Big Data is widely recognised as one of the most powerful drivers to promote efficiency,productivity and innovation.Big Data has environmental and social impacts on Innovation and performance which can enhance sustainability.Big Data has the potential to drive and support Green Initiatives,such as Green Strategy for Air Pollution control and prevention.To capture value from big data to improve Green strategy by providing conceptual framework of green challenge.

1.2 Motivation of the Work

Mortality has increased a lot in recent years, due to Traffic accidents and Hazardous air pollutants causing air pollution.They are both artificial causes of death.So we wanted to reduce the casualties caused by air pollution and traffic accidents by identifying their respective hot-spots.

Air pollution contributes to a majority of diseases which may lead to death of a person as shown by many studies which use chronic lower respiratory diseases as an indicator for deaths due to air pollution,so we wanted to analyze the data sets related to road accidents and air pollution to identify factors which are the major causes of death.

1.3 Literature Review

J.Wu et al discovers the trend of big data era and that of the new generation green revolution, providing a comprehensive and panoramic literature survey of big data technologies classified according to the specific green impacts they want to have.

R.Dubey et al investigates the effects of BDPA on social performance (SP)and environmental performance (EP)using variance based structural equation modeling.

Francesco Calza et al explore how to capture value from big data to improve green engagement by providing a conceptual model and potentiality of Big Data to embrace green challenge.

1.4 Research Gap

Previous Studies showed that big data has environmental & social impacts on innovation and performance of the supply chain using different theoretical perspectives, focusing on when and how these data can enhance sustainability in the supply chain by reducing risks and uncertainties.

- J Wu et al failed to provide a correlation framework between big data complexity and green strategy complexity
- R Dubey et al failed to provide performance effects in particular contextual conditions.
- Francesco calza et al instead of developing necessary architecture to exploit big data refer to external partners.

To Improve the accuracy and reliability of analysis, we are going to use IoT devices in our system which will collect real time data and will provide up-to-date results.

Chapter 2
Problem Statement

The main goal of the project is to analyze the existing datasets and find patterns which identify the factors contributing to deaths due to Air pollution and Traffic accidents and propose ideas to achieve SDG-3 Goal.

2.1 Air Pollution Analysis and Prediction Methodology

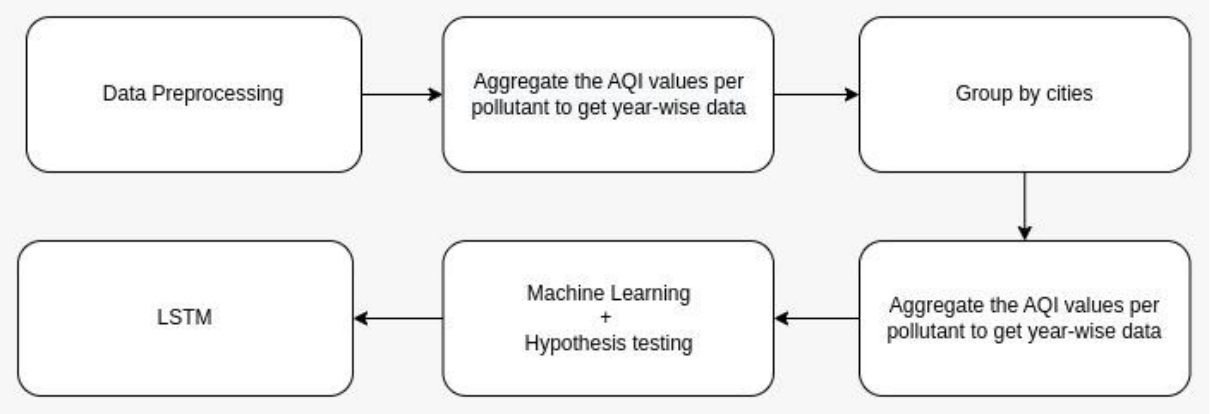


Fig. 2.1 Methodology for Air Pollution Analysis and Prediction

```
In [20]: Good= len(ind_data[(ind_data['VALUE']>0) & (ind_data['VALUE']<=30)].count(axis=1))
         Satisfactory=len(ind_data[(ind_data['VALUE']>31)&(ind_data['VALUE']<=60)].count(axis=1))
         Moderate=len(ind_data[(ind_data['VALUE']>61)&(ind_data['VALUE']<=90)].count(axis=1))
         Poor=len(ind_data[(ind_data['VALUE']>91)&(ind_data['VALUE']<=120)].count(axis=1))
         VeryPoor=len(ind_data[(ind_data['VALUE']>121)&(ind_data['VALUE']<=250)].count(axis=1))
         Severe=len(ind_data[(ind_data['VALUE']>=250)].count(axis=1))

         colors=['orangered','darkred','red','lightgreen','mediumgreen','yellow']
         AgiGrp=['Good','Satisfactory','Moderate','Poor','VeryPoor','Severe']
         AgiGrpVal=[Good,Satisfactory,Moderate,Poor,VeryPoor,Severe]

         Fgr=go.Figure(data=[go.Pie(labels=AgiGrp,values=AgiGrpVal,sort=False,
                 title="'AQI,Health BreakPoints and Pollutants from Jan 2019 to May 2021 for Indian Cities'",
                 marker=dict(colors=colors),textfont_size=12)])

         Fgr.update_layout(margin=dict(t=0, b=0, l=0, r=0))
         Fgr.show()
```

```
In [55]: list1 = []
         list2 = []
         window = 20
         X = avg_data.values
         hist = [X[i] for i in range(window)]
         test = [X[i] for i in range(window, len(X))]
         pred = list()


         counter = 0
         print("First 10 values:\n")


         for t in range(len(test)):
             length = len(hist)
             y = mean([hist[i] for i in range(length-window,length)])
             obs = test[t]
             pred.append(y)
             hist.append(obs)

             ## Predicted and Expected value
             if(counter<10):
                 print('> Predicted = %.2f,  Expected = %.2f ' % (y, obs))
             list1.append(y)
             list2.append(obs)
             counter+=1


         error = mean_squared_error(test, pred)
         print('\n-> Test MSE : %.2f' % error)
```

```
First 10 values:

> Predicted = 229.85,  Expected = 222.00
> Predicted = 225.25,  Expected = 116.00
> Predicted = 224.30,  Expected = 157.00
> Predicted = 224.15,  Expected = 199.00
> Predicted = 225.95,  Expected = 122.00
> Predicted = 223.30,  Expected = 153.00
> Predicted = 222.90,  Expected = 138.00
```

The methodology applied for conducting the Air Pollution analysis study is purely based on data taken from Government websites.

Firstly,
Data Preprocessing, In this stage we used pandas, numpy, geopandas, pandas and numpy are a necessity for storing and doing computations on data. Geopandas help us in dealing with geospatial data i.e. data which has an association with a location relative to earth.
using these libraries we read the data, checked for any NULL and NaN values in the data if so we removed those values

Exploratory Data Analysis,
With the preprocessed data, We aggregated the AQI values (Air Quality Index) per pollutant to get year-wise data, we checked the distribution of AQI for India, removed any values from the data which has AQI values over 800, plotted the graphs using seaborn library, and divided the data into categories based on their AQI values
like,
0-50 -> Good
51-100 -> Satisfactory
101-200 -> Moderately polluted
201-300 -> Poor
301-400 -> Very Poor
401-500 -> Severe

For all the gasses like NO2,SO2.
and we measured the air quality which is between PM 2.5 and PM 10 -> PM - Particles in micrograms per cubic metre. We did some visualizations to show the in-depth relation between AQI values and the PM 2.5 value.

Group by Cities,
Then we Group the data by Cities,
check their values during covid and after covid, then we analyzed the top 10 maximum pollutant cities of India during lockdown and after lockdown.

Machine Learning, Hypothesis testing and LSTM, We performed Visualizations on the data using geopandas and seaborn. Then performed Time Series Analysis since the data we had was time series data i.e. series of data points indexed in time order. We used the LSTM algorithm (Long Short Term Memory) to calculate the prediction based on the data, matplotlib to plot the Predicted values vs Actual values, did the hypothesis testing and thus trained the data.

2.2 Road Accident Analysis Methodology

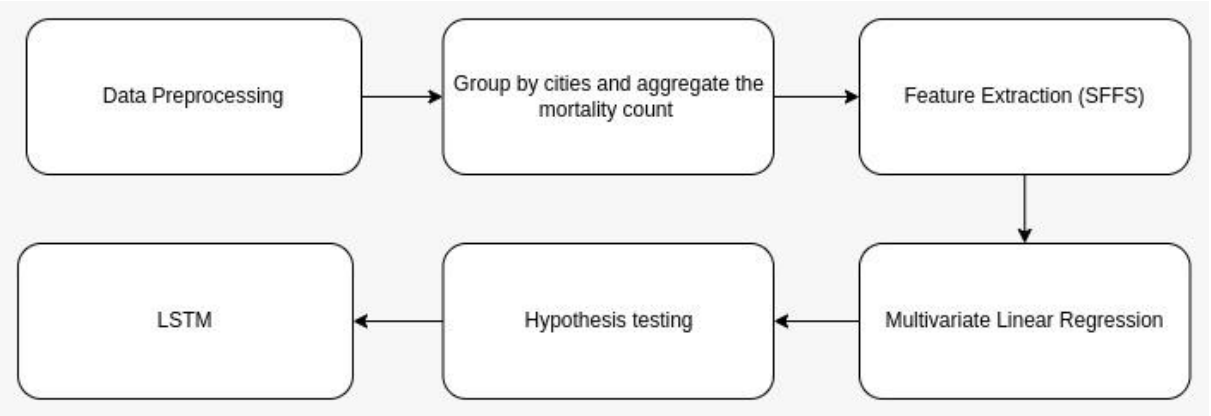


Fig. 2.2 Methodology for Road Accident Analysis

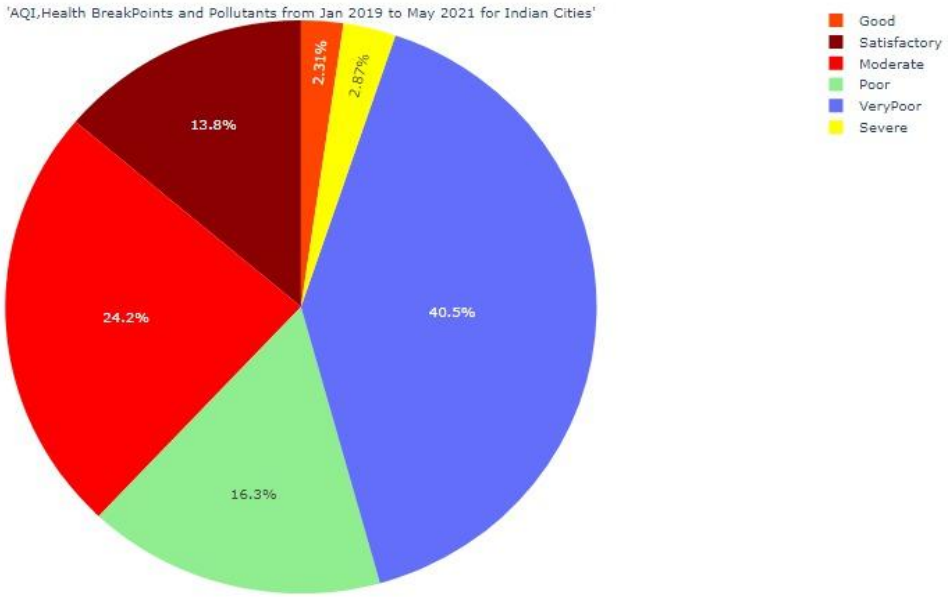
```
In [62]: simple_lstm = Sequential()
         simple_lstm.add(LSTM(64, input_shape=(buffer, 1)))
         simple_lstm.add(Dense(1))
         simple_lstm.compile(loss='mae', optimizer=RMSprop())
         start=time.time()
         checkpointer = ModelCheckpoint(filepath='./simple_lstm_weights.hdf5'
                                        , save_best_only=True,verbose=1)
         earlystopper = EarlyStopping(monitor='val_loss'
                                      , patience=10)
         with open("./simple_lstm.json", "w") as m:
             m.write(simple_lstm.to_json())

         simple_lstm_history = simple_lstm.fit_generator(train_data_gen
                                                         , epochs=1000
                                                         , validation_data=valid_data_gen
                                                         , callbacks=[checkpointer, earlystopper],verbose=1)
         plot_loss(simple_lstm_history, 'Simple LSTM - Train & Validation Loss')
         simple_lstm.save("./custom_lstm")
         end=time.time()
         tt=end-start
```

The methodology applied for conducting the Road Accidents analysis is similar to air pollution analysis, We had state/UTs-year-accidents data that we preprocessed to perform some visualizations, No.of accidents/Year, mean of state/UT-wise total no. of Road accidents vs years, Total number of Accidents in each State, Count of Accidents per each state according to time of day, daytime and night time accidents year-wise, Offender who died according to gender, victims who died according to gender, check for driving license permanent license, learners license, without license, weather and road conditions in the time of accident, Accidents involving non wearing of safety gear i.e helmet or seat-belt, grouped data on season wise accidents, type of vehicles, then grouped them by total number of accidents for each vehicle of that state

Chapter 3
Results and Discussion

Air Pollution Analysis and Prediction:



> We can see that almost half i.e 40.5% of Indian Cities fall in Very Poor Category

Selecting PM 2.5 (24 hr) Column Division for Segregating Data's

AQI Category, Pollutants and Health Breakpoints

AQI Category (Range)	PM_{10} (24hr)	$PM_{2.5}$ (24hr)	NO_2 (24hr)	O_3 (8hr)	CO (8hr)	SO_2 (24hr)	NH_3 (24hr)	Pb (24hr)
Good (0–50)	0–50	0–30	0–40	0–50	0–1.0	0–40	0–200	0–0.5
Satisfactory (51–100)	51–100	31–60	41–80	51–100	1.1–2.0	41–80	201–400	0.5–1.0
Moderately polluted (101–200)	101–250	61–90	81–180	101–168	2.1–10	81–380	401–800	1.1–2.0
Poor (201–300)	251–350	91–120	181–280	169–208	10–17	381–800	801–1200	2.1–3.0
Very poor (301–400)	351–430	121–250	281–400	209–748	17–34	801–1600	1200–1800	3.1–3.5
Severe (401–500)	430+	250+	400+	748+	34+	1600+	1800+	3.5+

Fig 3.1 AQI Index from Jan 2019 to May 2021 for Indian cities & AQI category breakpoints

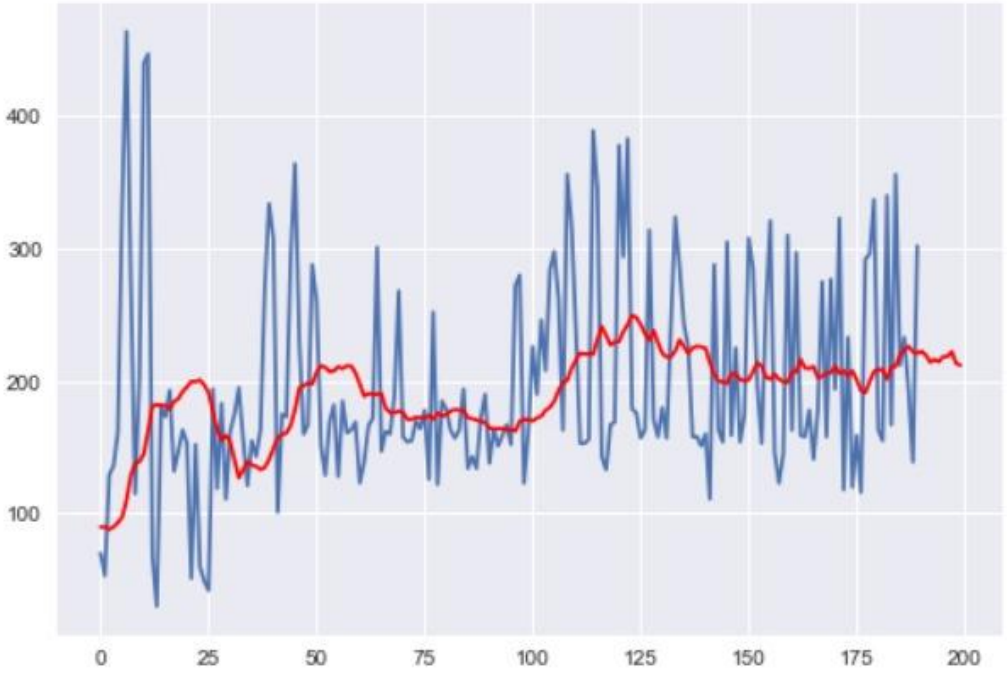


Fig 3.2 Top 400 Expected and Predicted Values in Plot

```
You should buy house around the station code where no2 is minumum
                no2
stn_code
700.0     8.280612
```

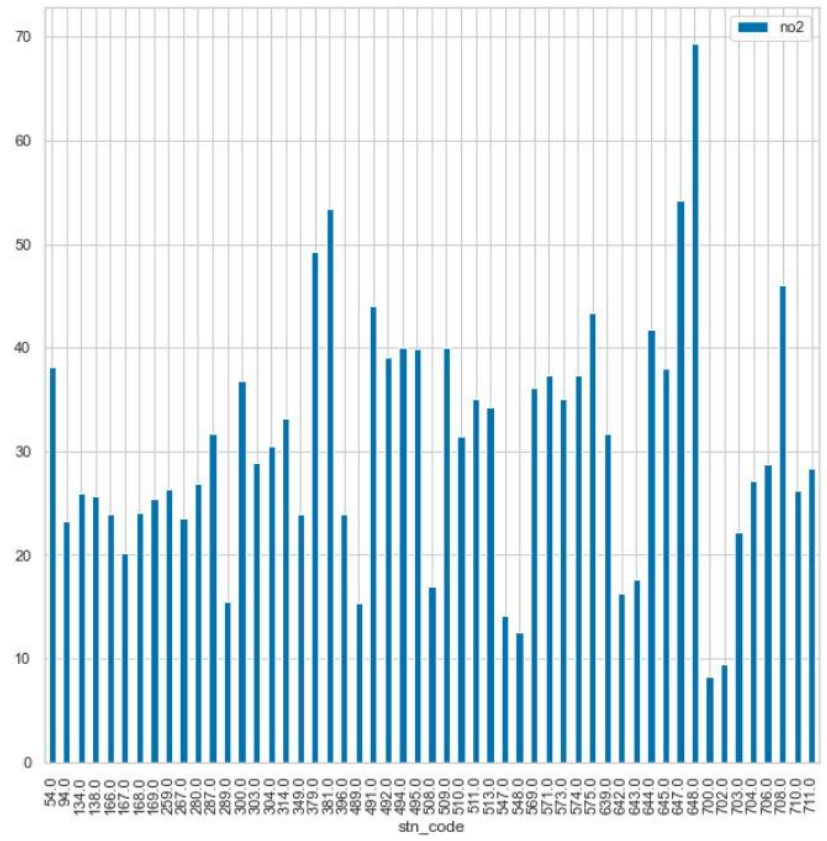


Fig 3.3 NO_2 values based on std codes

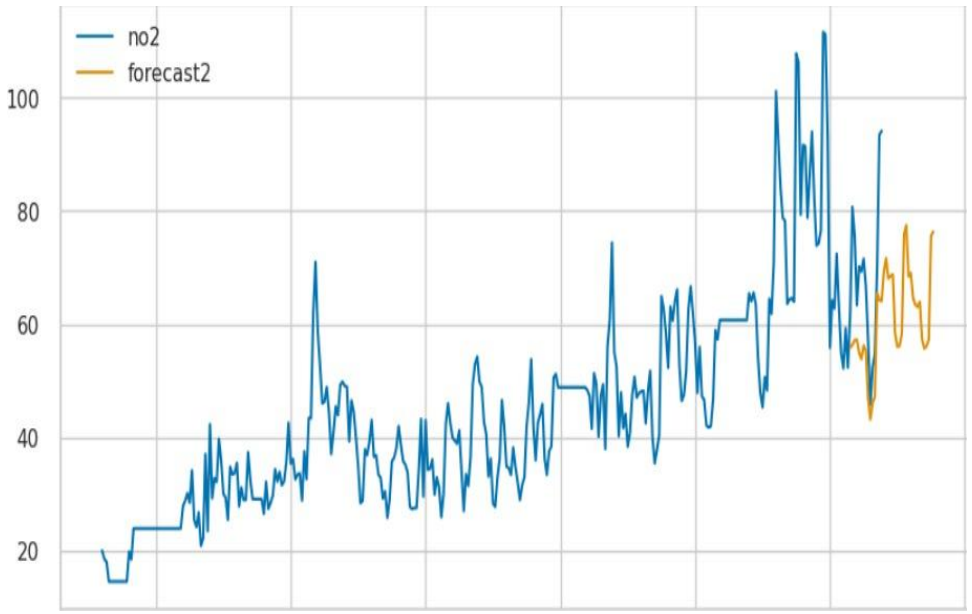


Fig 3.4 NO_2 values and the forecast future NO_2 values

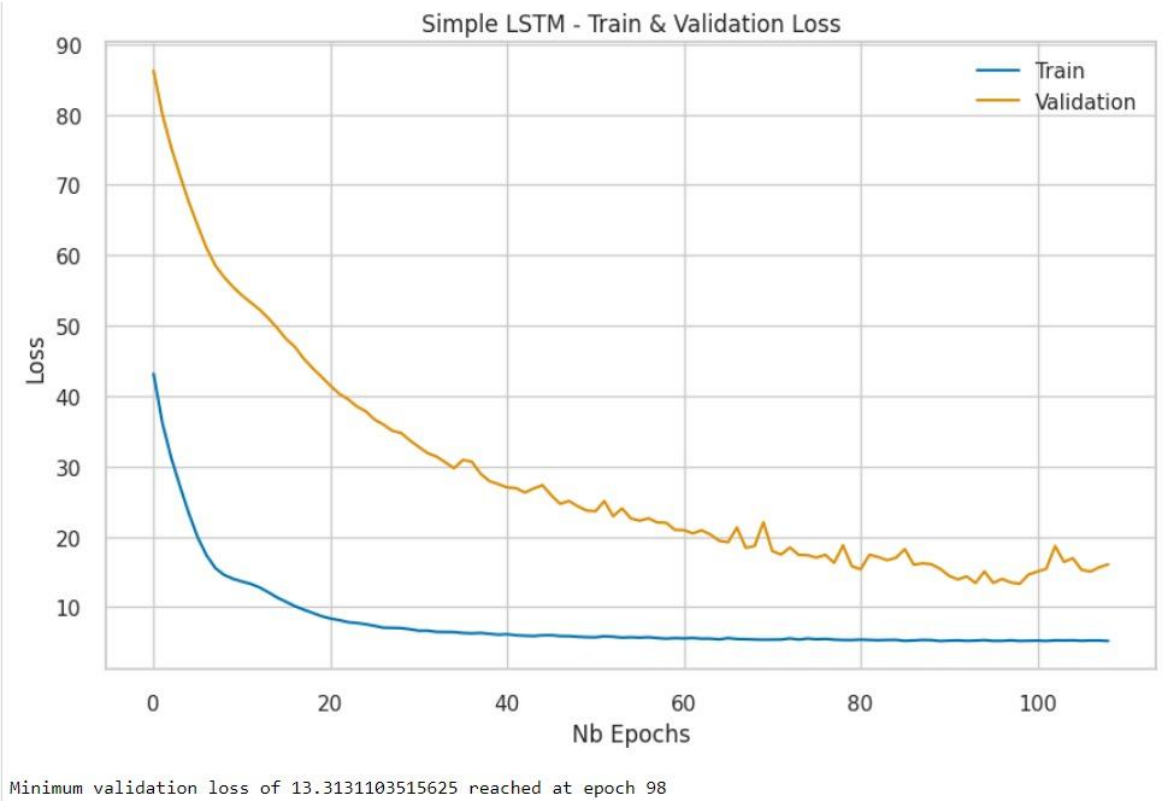


Fig 3.5 Simple LSTM - Training Loss & Validation Loss

Road Accidents Analysis:

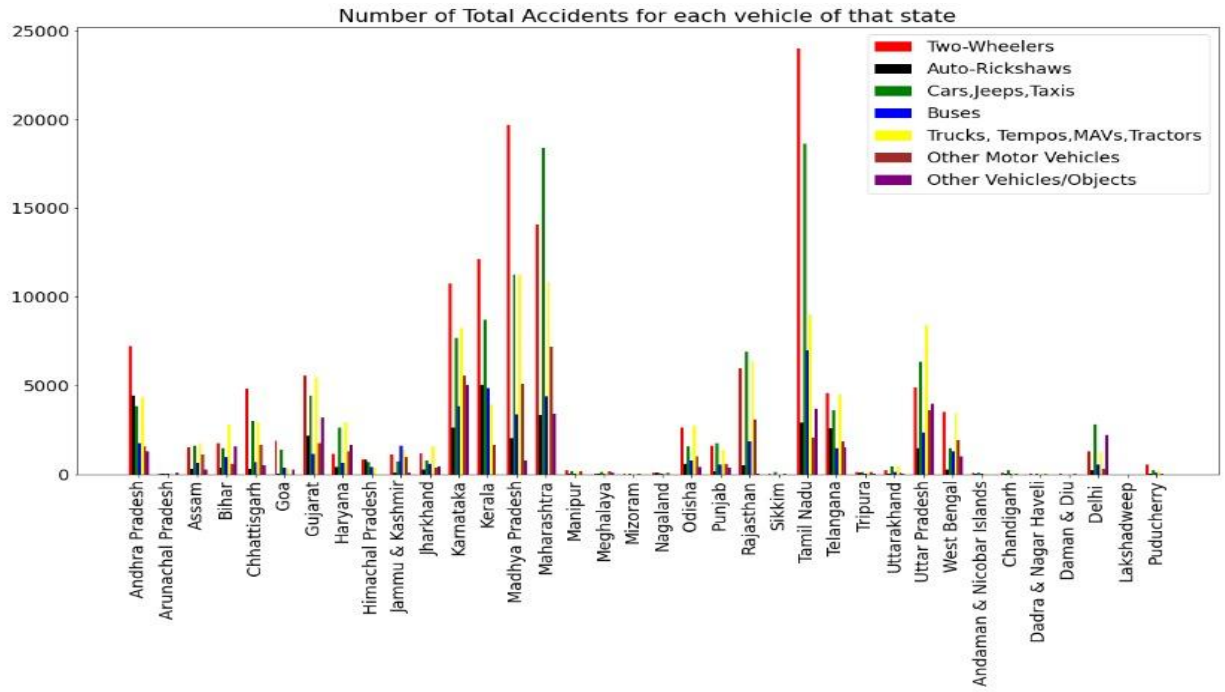


Fig 3.6 Number of total accidents for each vehicle of that state

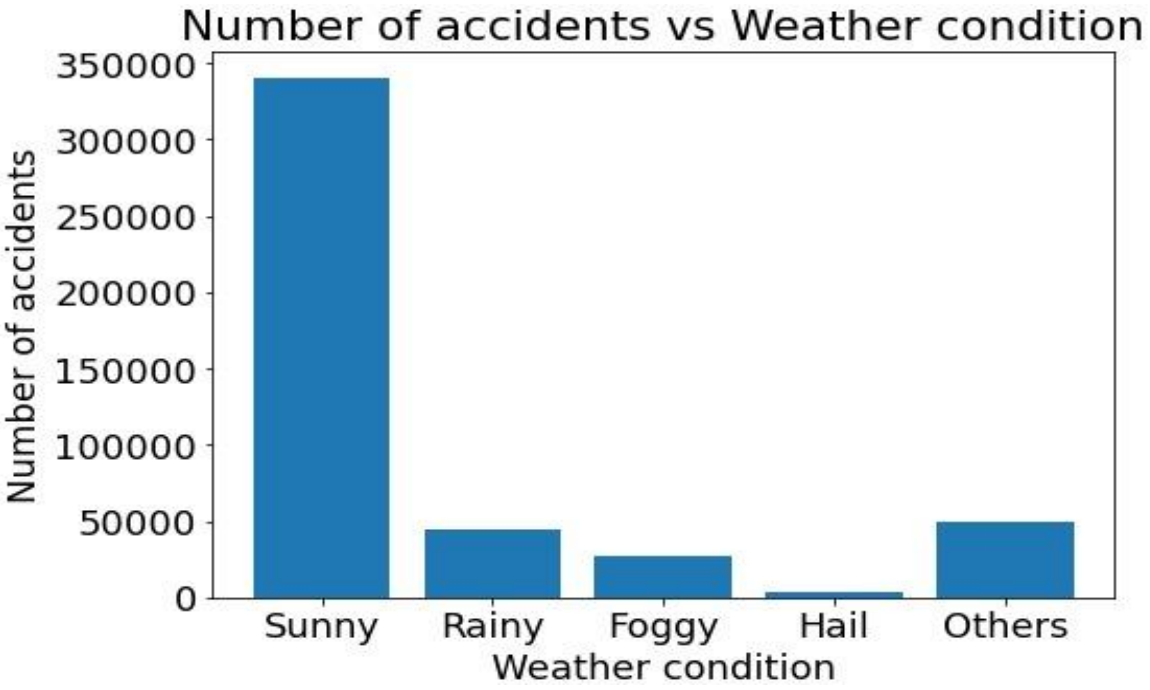


Fig 3.7 Number of accidents vs weather condition

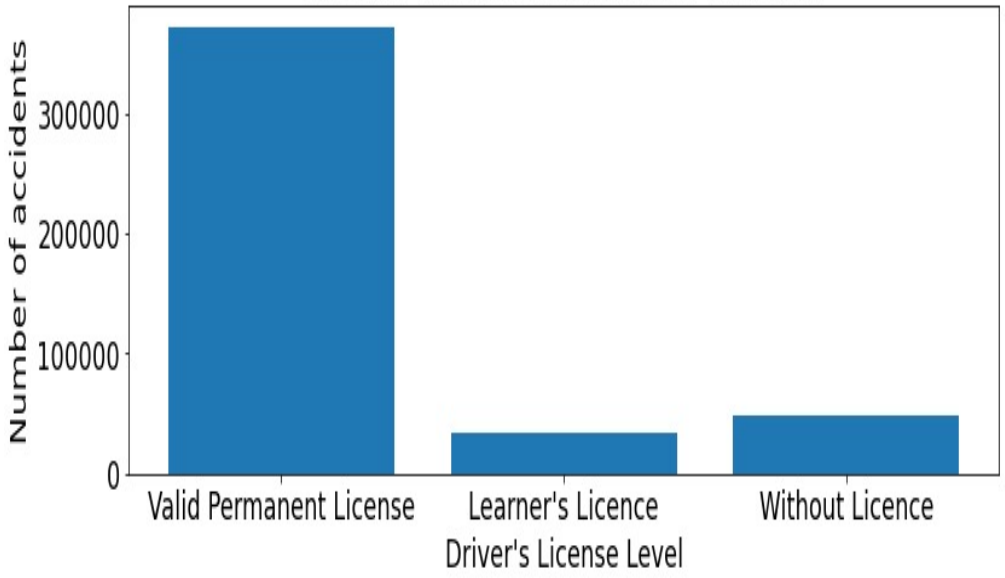


Fig 3.8 Number of accidents vs drivers license

- Expectedly, most accidents are seen by Valid permanent license holders because of their majority. However, it is interesting to see that there are more drivers who met with an accident without a license than with a learner's license

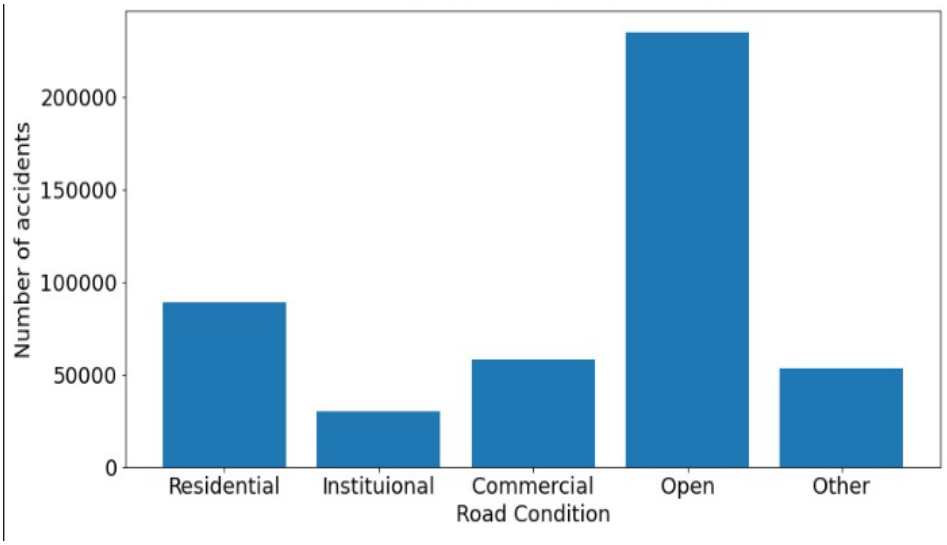


Fig 3.9 Number of accidents vs Road Conditions

- We observe that open roads witness more accidents than road conditions.

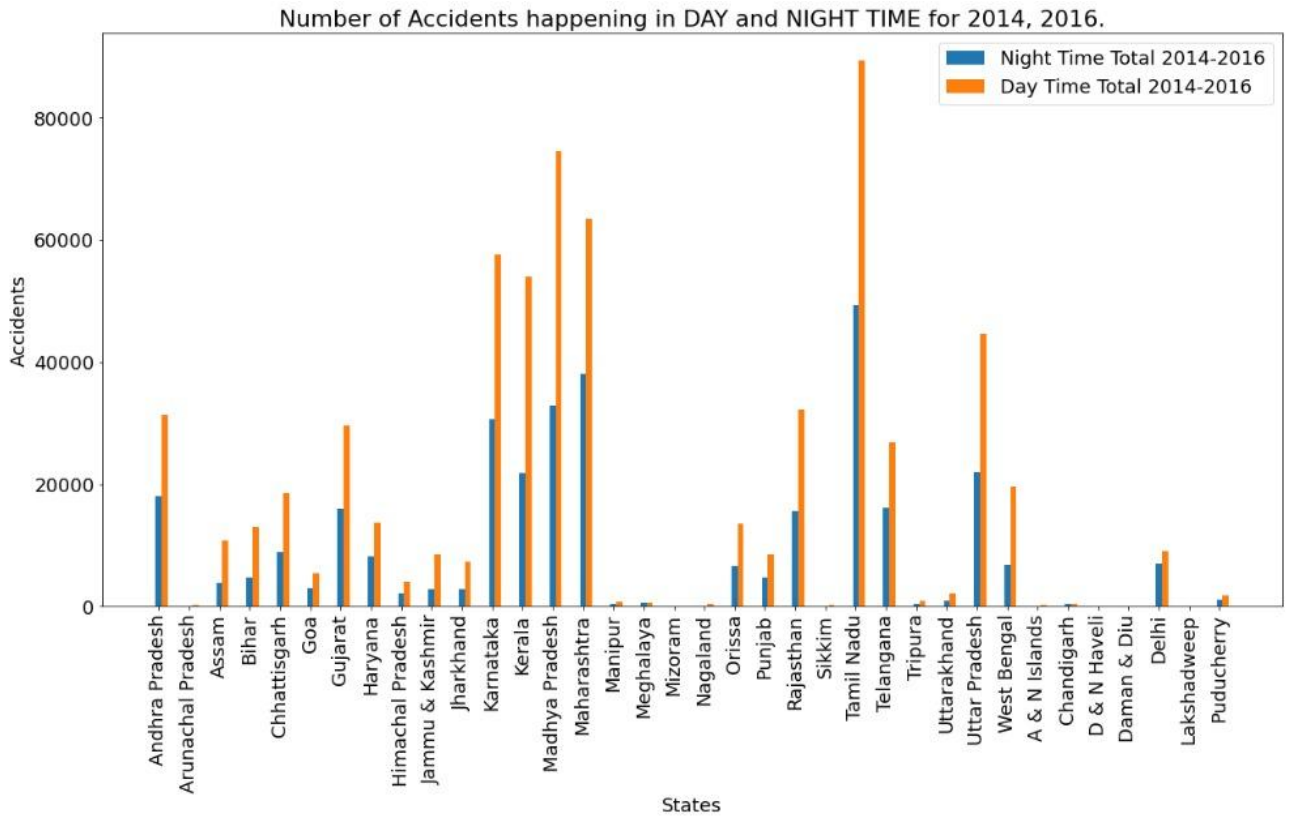


Fig 3.10 Number of accidents happening in day and night for 2014,2016

Fig 3.11 Drivers without safety equipments

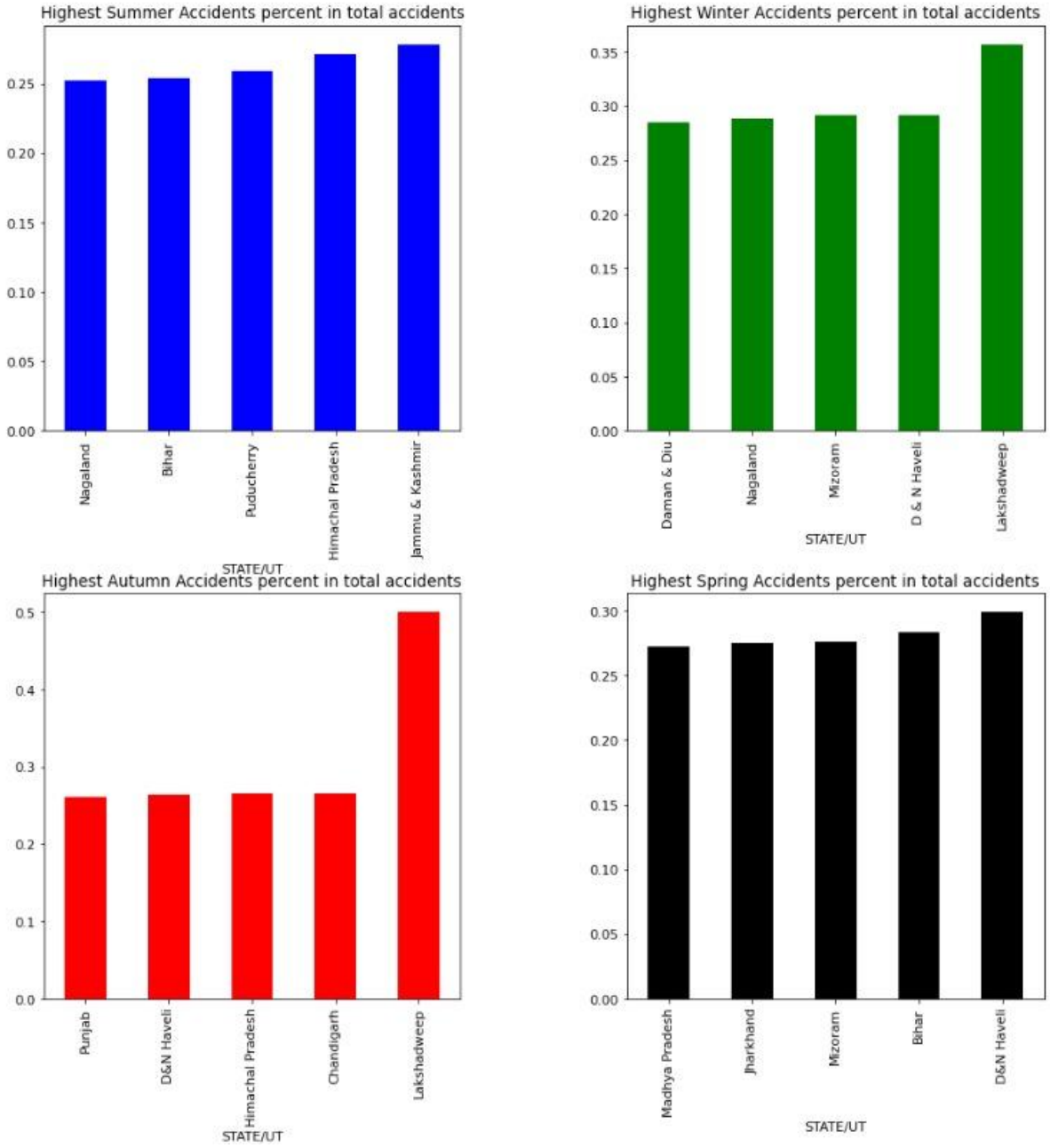


Fig 3.12 Highest accidents as per the season

Chapter 4
Conclusion and Future Scope

It is crucial to understand the collective impacts of multiple air pollutants on people's health in order to reduce the mortalities. Analyzing and understanding this huge amount of data can offer a better perspective on the long term effects of air pollution and what could be the most efficient and effective measures to prevent it. The mortality from accidents can be reduced with cautious driving and by following the traffic rules and regulations by all and at all times.

Our analysis can be used by the different regulatory institutions to find the mortality based hotspots and pointers suggested by this project can be used to enforce methods and policies to reduce the overall premature mortalities

- We Analyzed the air pollution and traffic fatalities data across the years based on that :
  - We identified the cities with the most mortality and found the factors which influence mortality the most.
  - Gave Pointers to reduce the mortality.
  - Strengthen the prevention measures of air pollution and traffic accidents
  - Identified Hot-spots of air pollution and traffic accidents
  - The most common measurement used to measure air quality is PM 2.5 and PM 10. It measures the particles in micrograms per cubic meter.
  - PM 2.5 refers to the concentration of microscopic particles less than 2.5 microns in diameter and PM 10 refers to the concentration of particles less than 10 microns in diameter.
  - The above graph shows in-depth relation AQI values and PM 2.5 values based on the predicted and estimated condition.
  - The top city to invest in is Ahmedabad, followed by Delhi and Patna.
  - The main gasses or pollutants responsible for high AQI count are NO, Particulate matter, and CO.
  - Also to reduce pollution measure should be taken to restrict the modes causing increase in the above pollutants.

References

- Wu, J., Guo, S., Li, J., Zeng, D. (2016). Big data meets green challenges: Big data toward green applications. *IEEE Systems Journal*, *10*(3), 888-900.

- Dubey, R., Gunasekaran, A., Childe, S. J., Papadopoulos, T., Luo, Z., Wamba, S. F., & Roubaud, D. (2019). Can big data and predictive analytics improve social and environmental sustainability?. *Technological Forecasting and Social Change*, *144*, 534-545.

- Calza, F., Parmentola, A., & Tutore, I. (2020). Big data and natural environment. How does different data support different green strategies?. *Sustainable Futures*, *2*, 100029.

- https://cpcb.nic.in/namp-data/

- https://morth.nic.in/road-accident-in-india