SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be **University** *u/s 3 of UGC Act, 1956)*

MINI-PROJECT REPORT

ROAD ACCIDENT ANALYSIS

Submitted to the partial fulfillment of the requirement
for the ***18CSE355T-Data Mining and Analytics*** *course and*
for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENE AND ENGINEERING

Submitted by-
Mr. MOKSH VIJ
[Registration No.RA1811003030124]
Mr. NIMISH DEEP
[Registration No.RA1811003030129]
Group ID: 2020DDDD

Under the guidance of
Mr. PARTHA SARATHI CHAKRABORTY
(Assistant Professor, Department of Computer Science and Engineering)

DEPARTMENT OF COMPUTER SCIENE AND ENGINEERING
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
(Deemed to be University under Section 3 of the UGC Act, 1956)

NOVEMBER 2020

BONAFIDE CERTIFICATE

Certified that this project report titled "ROAD ACCIDENT ANALYSIS" is the bonafide work of Mr. MOKSH VIJ [Registration No. RA1811003030124] and Mr. NIMISH DEEP [Registration No. RA1811003030129], who carried out the project work under my supervision and submitted to the partial fulfillment of the requirement for the 18CSE355T Data Mining and Analytics course and for the award of the degree of Bachelor of Technology in Computer Science and Engineering of SRM Institute of Science and Technology.

Mr. Partha Sarathi Chakraborty
Supervisor,
Assistant Professor, Department of Computer Science and Engineering

ABSTARCT

At present, India is in a period of steady development of highways. At the same time, traffic safety issues are becoming increasingly serious. There are many inventories in automobile industries to design and build safety measures for automobiles, but traffic accidents are unavoidable. There is a huge number of accidents prevailing in all urban and rural areas. Patterns involved with different circumstances can be detected by developing an accurate prediction models which will be capable of automatic separation of various accidental scenarios. Data mining technology is an effective method for analyzing traffic accidents. In-depth information mining of traffic accident data is conducive to accident prevention and traffic safety management. This cluster will be useful to prevent accidents and develop safety measures. We believe to acquire maximum possibilities of accident reduction using low budget resources by using some scientific measures. Based on the data of traffic accidents from 2001 to 2014, this study selected factors including time at which the accident occurred, the state in which the accident occurred and the increase in the number of the accidents in different states in different years.

TABLE OF CONTENTS

INTRODUCTION

There is a huge impact on the society due to traffic accidents where there is a great cost of fatalities and injuries. In recent years, there is increase in the researches attention to determine the significantly affect the severity of the driver's injuries which is caused due to the road accidents. Accurate and comprehensive accident records are the basis of accident analysis. The effective use of accident records depends on some factors, like the accuracy of the data, record retention, and data analysis. There are many approaches applied to this scenario to study this problem. 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.

A recent study illustrated that the residential and shopping sites are more hazardous than village areas. As might have been predicted, the frequencies of the casualties were higher near the zones of residence possibly because of the higher exposure. A study revealed that the casualty rates among the residential areas are classified as relatively deprived and significantly higher than those from relatively affluent areas. Patterns involved in serious accidents can be able to classify various traffic accidents with developing accurate prediction models. Identifying the primary road traffic accident factors will help to provide an appropriate solution to minimize the adverse effect of severity on human and property loss. Data mining algorithms can be used to help in preventing these accidents by providing a systematic analysis of the accidents around the country which will eventually help in creating some major measures to prevent these accidents.

LITERATURE SURVEY

Data Mining is the process of discovering interesting patterns and knowledge from a large database and data classification is a form of data analysis that involves extracting models with important data classes.

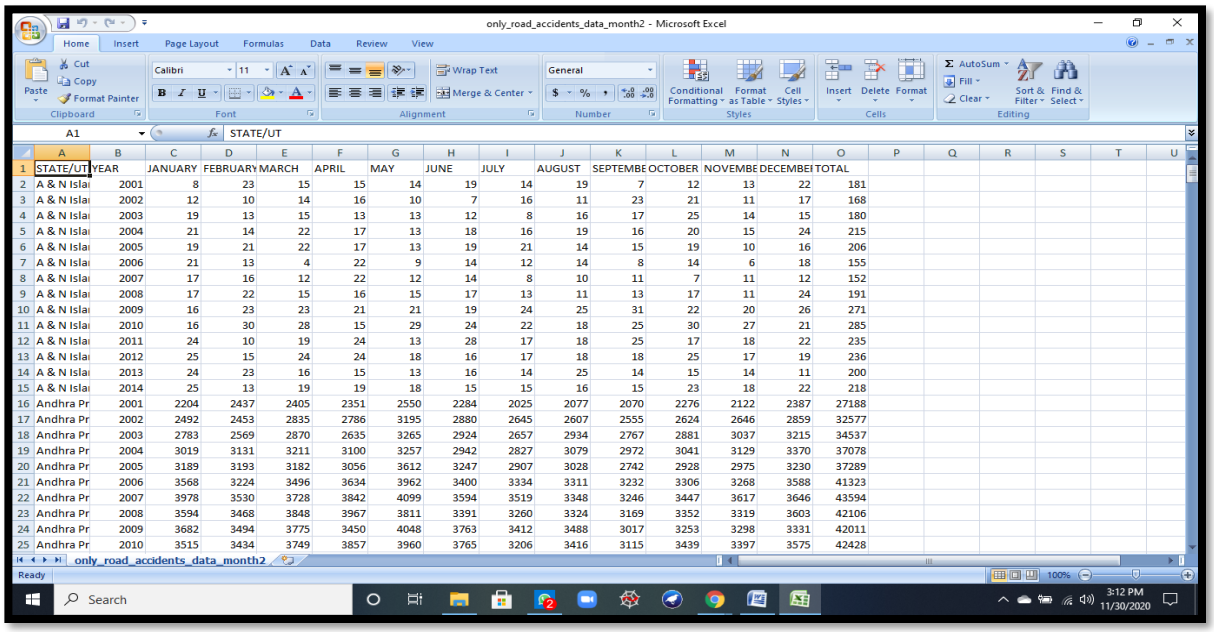
This literature survey provides a comprehensive review about the data mining techniques applied to achieve this analysis report.

The given data is used to do a detailed analysis of road accidents across India using clustering which provide us with some unique patterns. These patterns help in classifying accidents across the country on the factors of accident-prone time, weather and place.

Identifying the primary road traffic accident factors will help us to reach an appropriate solution to minimize the adverse effects of accidents.

DATASET DESCRIPTION

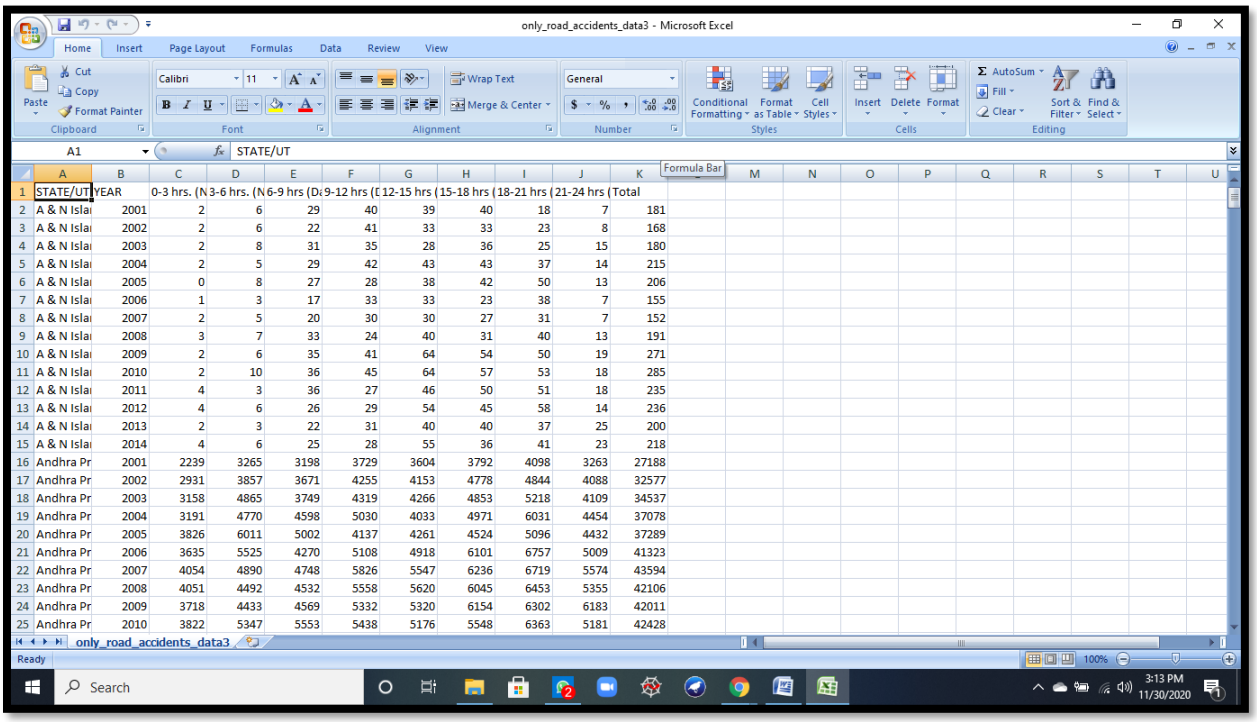
- The dataset used here contains a detailed information of the accidents happening all across the country based on the factors of time, month and seasons.
- The first table of the dataset contains the accident count of different states of India between the period of 2001-2014 during the different months in a year. The attributes are-
    1. State/UT - This column contains all the states and union territories of India.
    2. Year - This column contains the year count (2001-14).
    3. Column C to N contains the count of accident each month in different states.
    4. Total - This column contains the total count of accidents each year in each state.
- The attributes of this table is also used to group months according to seasons and give a accident pattern on the basis of seasons i.e. Summer, Winter, Autumn, Spring.
- Note - All attributes are being used of the table.



STATE/UT	YEAR	JANUARY	FEBRUARY	MARCH	APRIL	MAY	JUNE	JULY	AUGUST	SEPTEMBE	OCTOBER	NOVEMBE	DECEMBE	TOTAL
A & N Islai	2001	8	23	15	15	14	19	14	19	7	12	13	22	181
A & N Islai	2002	12	10	14	16	10	7	16	11	23	21	11	17	168
A & N Islai	2003	19	13	15	13	13	12	8	16	17	25	14	15	180
A & N Islai	2004	21	14	22	17	13	18	16	19	16	20	15	24	215
A & N Islai	2005	19	21	22	17	13	19	21	14	15	19	10	16	206
A & N Islai	2006	21	13	4	22	9	14	12	14	8	14	6	18	155
A & N Islai	2007	17	16	12	22	12	14	8	10	11	7	11	12	152
A & N Islai	2008	17	22	15	16	15	17	13	11	13	17	11	24	191
A & N Islai	2009	16	23	23	21	21	19	24	25	31	22	20	26	271
A & N Islai	2010	16	30	28	15	29	24	22	18	25	30	27	21	285
A & N Islai	2011	24	10	19	24	13	28	17	18	25	17	18	22	235
A & N Islai	2012	25	15	24	24	18	16	17	18	18	25	17	19	236
A & N Islai	2013	24	23	16	15	13	16	14	25	14	15	14	11	200
A & N Islai	2014	25	13	19	19	18	15	15	16	15	23	18	22	218
Andhra Pr	2001	2204	2437	2405	2351	2550	2284	2025	2077	2070	2276	2122	2387	27188
Andhra Pr	2002	2492	2453	2835	2786	3195	2880	2645	2607	2555	2624	2646	2859	32577
Andhra Pr	2003	2783	2569	2870	2635	3265	2924	2657	2934	2767	2881	3037	3215	34537
Andhra Pr	2004	3019	3131	3211	3100	3257	2942	2827	3079	2972	3041	3129	3370	37078
Andhra Pr	2005	3189	3193	3182	3056	3612	3247	2907	3028	2742	2928	2975	3230	37289
Andhra Pr	2006	3568	3224	3496	3634	3962	3400	3334	3311	3232	3306	3268	3588	41323
Andhra Pr	2007	3978	3530	3728	3842	4099	3594	3519	3348	3246	3447	3617	3646	43594
Andhra Pr	2008	3594	3468	3848	3967	3811	3391	3260	3324	3169	3352	3319	3603	42106
Andhra Pr	2009	3682	3494	3775	3450	4048	3763	3412	3488	3017	3253	3298	3331	42011
Andhra Pr	2010	3515	3434	3749	3857	3960	3765	3206	3416	3115	3439	3397	3575	42428

Table 1

- The second table of the dataset contains the accident count of different states of India between the period of 2001-2014 during the different hours of the day in a period of 3 hours. The attributes are-
    1. State/UT - This column contains all the states and union territories of India.
    2. Year - This column contains the year count (2001-14).
    3. Column C to N contains the count of accident in every 3 hours during a single day in different states.
    4. Total - This column contains the total count of accidents each year in each state.
- Note - All attributes are being used of the table.



only_road_accidents_data3 - Microsoft Excel

STATE/UT	YEAR	0-3 hrs. (N	3-6 hrs. (N	6-9 hrs (D	9-12 hrs (	12-15 hrs (	15-18 hrs (	18-21 hrs (	21-24 hrs (	Total
A & N Isla	2001	2	6	29	40	39	40	18	7	181
A & N Isla	2002	2	6	22	41	33	33	23	8	168
A & N Isla	2003	2	8	31	35	28	36	25	15	180
A & N Isla	2004	2	5	29	42	43	43	37	14	215
A & N Isla	2005	0	8	27	28	38	42	50	13	206
A & N Isla	2006	1	3	17	33	33	23	38	7	155
A & N Isla	2007	2	5	20	30	30	27	31	7	152
A & N Isla	2008	3	7	33	24	40	31	40	13	191
A & N Isla	2009	2	6	35	41	64	54	50	19	271
A & N Isla	2010	2	10	36	45	64	57	53	18	285
A & N Isla	2011	4	3	36	27	46	50	51	18	235
A & N Isla	2012	4	6	26	29	54	45	58	14	236
A & N Isla	2013	2	3	22	31	40	40	37	25	200
A & N Isla	2014	4	6	25	28	55	36	41	23	218
Andhra Pr	2001	2239	3265	3198	3729	3604	3792	4098	3263	27188
Andhra Pr	2002	2931	3857	3671	4255	4153	4778	4844	4088	32577
Andhra Pr	2003	3158	4865	3749	4319	4266	4853	5218	4109	34537
Andhra Pr	2004	3191	4770	4598	5030	4033	4971	6031	4454	37078
Andhra Pr	2005	3826	6011	5002	4137	4261	4524	5096	4432	37289
Andhra Pr	2006	3635	5525	4270	5108	4918	6101	6757	5009	41323
Andhra Pr	2007	4054	4890	4748	5826	5547	6236	6719	5574	43594
Andhra Pr	2008	4051	4492	4532	5558	5620	6045	6453	5355	42106
Andhra Pr	2009	3718	4433	4569	5332	5320	6154	6302	6183	42011
Andhra Pr	2010	3822	5347	5553	5438	5176	5548	6363	5181	42428

Table 2

DATASET PREPROCESSING

1. Importing Libraries:

```
[49] import numpy as np
     import pandas as pd
     import seaborn as sns
     import matplotlib.pyplot as plt
     %matplotlib inline
     from IPython.display import display as display
     import plotly.graph_objs as go
```

2. Reading files:

```
[50] # Read the data to the dataframe from the data files.

     #state_year_month_df contains data for each state, segragated into year and month
     state_year_month_df=pd.read_csv('only_road_accidents_data_month2.csv')

     #state_year_time_df contains data for each state, segragated into year and time of the day
     state_year_time_df=pd.read_csv('only_road_accidents_data3.csv')
```

```
state_year_month_df.head()
```

	STATE/UT	YEAR	JANUARY	FEBRUARY	MARCH	APRIL	MAY	JUNE	JULY	AUGUST	SEPTEMBER	OCTOBER	NOVEMBER	DECEMBER	TOTAL
0	A & N Islands	2001	8	23	15	15	14	19	14	19	7	12	13	22	181
1	A & N Islands	2002	12	10	14	16	10	7	16	11	23	21	11	17	168
2	A & N Islands	2003	19	13	15	13	13	12	8	16	17	25	14	15	180
3	A & N Islands	2004	21	14	22	17	13	18	16	19	16	20	15	24	215
4	A & N Islands	2005	19	21	22	17	13	19	21	14	15	19	10	16	206

```
[52] state_year_time_df.head()
```

	STATE/UT	YEAR	0-3 hrs. (Night)	3-6 hrs. (Night)	6-9 hrs (Day)	9-12 hrs (Day)	12-15 hrs (Day)	15-18 hrs (Day)	18-21 hrs (Night)	21-24 hrs (Night)	Total
0	A & N Islands	2001	2	6	29	40	39	40	18	7	181
1	A & N Islands	2002	2	6	22	41	33	33	23	8	168
2	A & N Islands	2003	2	8	31	35	28	36	25	15	180
3	A & N Islands	2004	2	5	29	42	43	43	37	14	215
4	A & N Islands	2005	0	8	27	28	38	42	50	13	206

3. Grouping of months into seasons and timeslots in the dataset in a more understandable way:

```
[53] #Get all the state names in an array..
     state_names=state_year_month_df['STATE/UT'].unique()
     print(state_names)

     ['A & N Islands' 'Andhra Pradesh' 'Arunachal Pradesh' 'Assam' 'Bihar'
      'Chandigarh' 'Chhattisgarh' 'D & N Haveli' 'D&N Haveli' 'Daman & Diu'
      'Delhi (Ut)' 'Delhi Ut' 'Goa' 'Gujarat' 'Haryana' 'Himachal Pradesh'
      'Jammu & Kashmir' 'Jharkhand' 'Karnataka' 'Kerala' 'Lakshadweep'
      'Madhya Pradesh' 'Maharashtra' 'Manipur' 'Meghalaya' 'Mizoram' 'Nagaland'
      'Odisha' 'Puducherry' 'Punjab' 'Rajasthan' 'Sikkim' 'Tamil Nadu'
      'Tripura' 'Uttar Pradesh' 'Uttarakhand' 'West Bengal']

[54] #state_year_month_df=state_year_month_df['STATE/UT']
     state_year_month_df['STATE/UT']=state_year_month_df['STATE/UT'].replace({'Delhi (Ut)': 'Delhi Ut', 'D & N Haveli':'D&N Haveli'})
     print(state_year_month_df['STATE/UT'].unique())

     ['A & N Islands' 'Andhra Pradesh' 'Arunachal Pradesh' 'Assam' 'Bihar'
      'Chandigarh' 'Chhattisgarh' 'D&N Haveli' 'Daman & Diu' 'Delhi Ut' 'Goa'
      'Gujarat' 'Haryana' 'Himachal Pradesh' 'Jammu & Kashmir' 'Jharkhand'
      'Karnataka' 'Kerala' 'Lakshadweep' 'Madhya Pradesh' 'Maharashtra'
      'Manipur' 'Meghalaya' 'Mizoram' 'Nagaland' 'Odisha' 'Puducherry' 'Punjab'
      'Rajasthan' 'Sikkim' 'Tamil Nadu' 'Tripura' 'Uttar Pradesh' 'Uttarakhand'
      'West Bengal']
```

```
[55] # Reassiging state names to variable..
     state_names=state_year_month_df['STATE/UT'].unique()

[56] #display(state_year_month_df.head())

     #Create season groups clubbing values from multiple month columns..
     state_year_month_df['SUMMER']=state_year_month_df[['JUNE','JULY','AUGUST']].sum(axis=1)
     state_year_month_df['AUTUMN']=state_year_month_df[['SEPTEMBER','OCTOBER','NOVEMBER']].sum(axis=1)
     state_year_month_df['WINTER']=state_year_month_df[['DECEMBER','JANUARY','FEBRUARY']].sum(axis=1)
     state_year_month_df['SPRING']=state_year_month_df[['MARCH','APRIL','MAY']].sum(axis=1)

     #Delete month columns..
     state_year_month_df=state_year_month_df.drop(['JANUARY','FEBRUARY','MARCH','APRIL','MAY','JUNE','JULY'
                                                   ,'AUGUST','SEPTEMBER','OCTOBER','NOVEMBER','DECEMBER'], axis=1)
     #Create groups of states, summing the values of accident number for each year..
     state_grouped=state_year_month_df.groupby(['STATE/UT']).sum()

     #Create % columns for noting the % of accidents happening in each state for each season..
     state_grouped['%_SUMMER']=state_grouped['SUMMER']/state_grouped['TOTAL']
     state_grouped['%_AUTUMN']=state_grouped['AUTUMN']/state_grouped['TOTAL']
     state_grouped['%_WINTER']=state_grouped['WINTER']/state_grouped['TOTAL']
     state_grouped['%_SPRING']=state_grouped['SPRING']/state_grouped['TOTAL']

     display(state_grouped.iloc[:,1:].head())
```

[56]

STATE/UT	TOTAL	SUMMER	AUTUMN	WINTER	SPRING	%_SUMMER	%_AUTUMN	%_WINTER	%_SPRING
A & N Islands	2893	689	710	779	715	0.238161	0.245420	0.269271	0.247148
Andhra Pradesh	546821	132891	129230	138633	146067	0.243025	0.236330	0.253525	0.267120
Arunachal Pradesh	3389	748	875	920	846	0.220714	0.258188	0.271467	0.249631
Assam	61718	14696	15382	15671	15969	0.238115	0.249230	0.253913	0.258741
Bihar	92648	23506	20866	21994	26282	0.253713	0.225218	0.237393	0.283676

[57]
```
#Working on the over the day data...
state_year_time_df.rename(columns={'0-3 hrs. (Night)':'0-3',
                                   '3-6 hrs. (Night)':'3-6',
                                     '6-9 hrs (Day)':'6-9', '9-12 hrs (Day)':'9-12','12-15 hrs (Day)':'12-15','15-18 hrs (Day)':'15-18',
                                       '18-21 hrs (Night)':'18-21','21-24 hrs (Night)':'21-24'}, inplace=True)
state_time_grouped=state_year_time_df.groupby(['STATE/UT']).sum()

state_time_grouped['%_MORNING']=(state_time_grouped['6-9']+state_time_grouped['9-12'])/state_time_grouped['Total']
state_time_grouped['%_AFTERNOON']=(state_time_grouped['12-15']+state_time_grouped['15-18'])/state_time_grouped['Total']
state_time_grouped['%_EVENING']=(state_time_grouped['18-21']+state_time_grouped['21-24'])/state_time_grouped['Total']
state_time_grouped['%_NIGHT']=(state_time_grouped['0-3']+state_time_grouped['3-6'])/state_time_grouped['Total']

state_time_grouped=state_time_grouped.drop(state_time_grouped.columns[0:9], axis=1)
display(state_time_grouped.head())
```

STATE/UT	Total	%_MORNING	%_AFTERNOON	%_EVENING	%_NIGHT
A & N Islands	2893	0.297961	0.402351	0.260283	0.039405
Andhra Pradesh	546821	0.243550	0.266599	0.281997	0.207854
Arunachal Pradesh	3389	0.291531	0.383889	0.213927	0.110652
Assam	61718	0.370556	0.363022	0.172510	0.093911
Bihar	92648	0.304853	0.299855	0.208423	0.186869

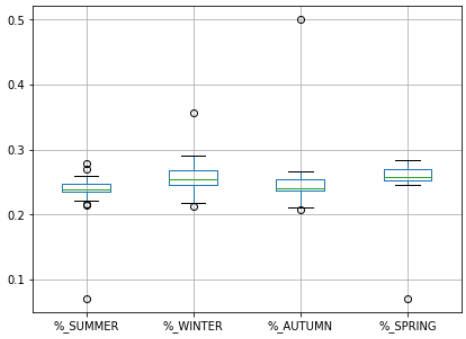
DATA ANALYSIS

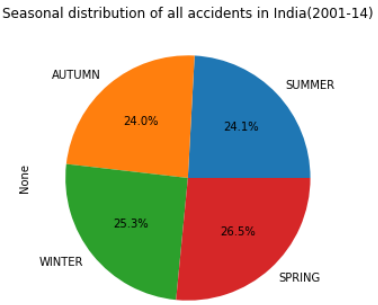
1. Seasonal Distribution of Accident Count:

- The pie chart and boxplot given below contains the data of the accidents occurring in the different seasons in a year across the country which will help us in predicting the most accident-prone season.

```
[58] plt.figure(figsize=(15,5))
     ax=plt.subplot(1,2,1)
     boxplot=state_grouped.boxplot(ax=ax,column=['%_SUMMER','%_WINTER','%_AUTUMN','%_SPRING'])

     ax=plt.subplot(1,2,2)
     state_grouped.loc[:,'SUMMER':'SPRING'].sum(axis=0).plot.pie(title='Seasonal distribution of all accidents in India(2001-14)',autopct='%1.1f%%')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6c2d9438d0>
```

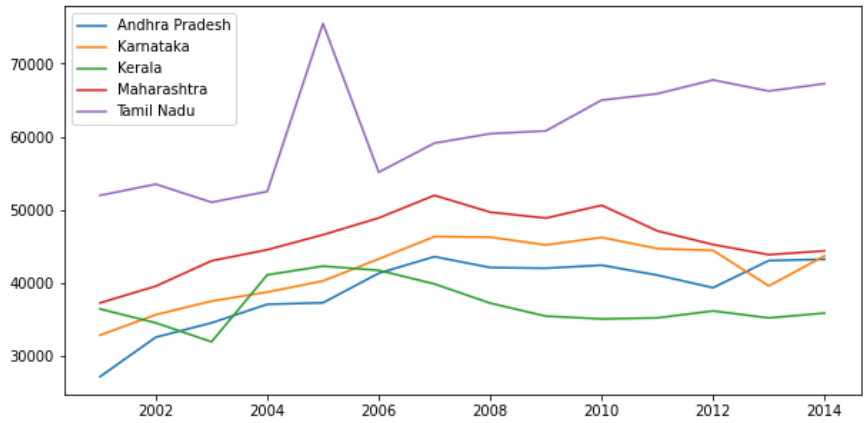




2. States\UT Wise Distribution of Accident Count:

- The graph given below contains the data of the total number of accidents occurring in the different states\UT in a year across the country which will help us in predicting the most accident-prone state\UT.

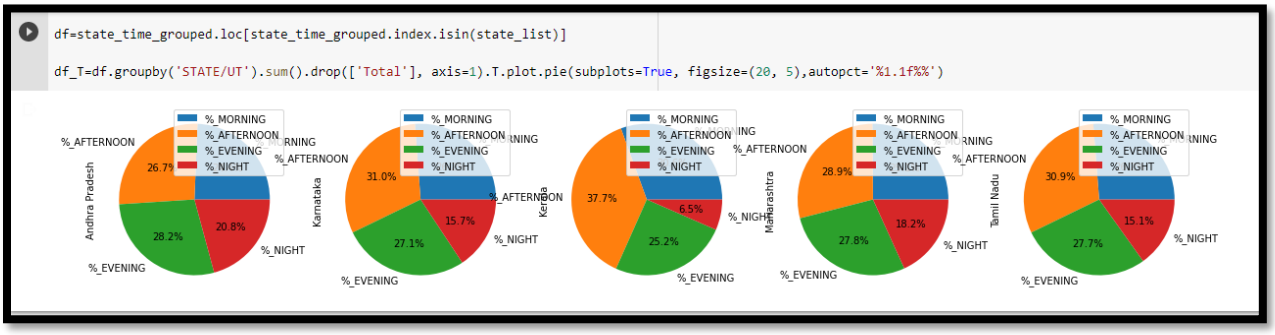
```
highest_accident_states=state_grouped.sort_values('TOTAL', ascending=False)
high_states=list(highest_accident_states.head().index)
df4=state_year_month_df.loc[state_year_month_df['STATE/UT'].isin(high_states),['STATE/UT','YEAR','TOTAL']]

plt.figure(figsize=(10,5))
ax=plt.subplot(111)
for key, grp in df4.groupby(['STATE/UT']):
    ax = grp.plot(ax=ax, kind='line', x='YEAR', y='TOTAL', label=key)

plt.show()
```



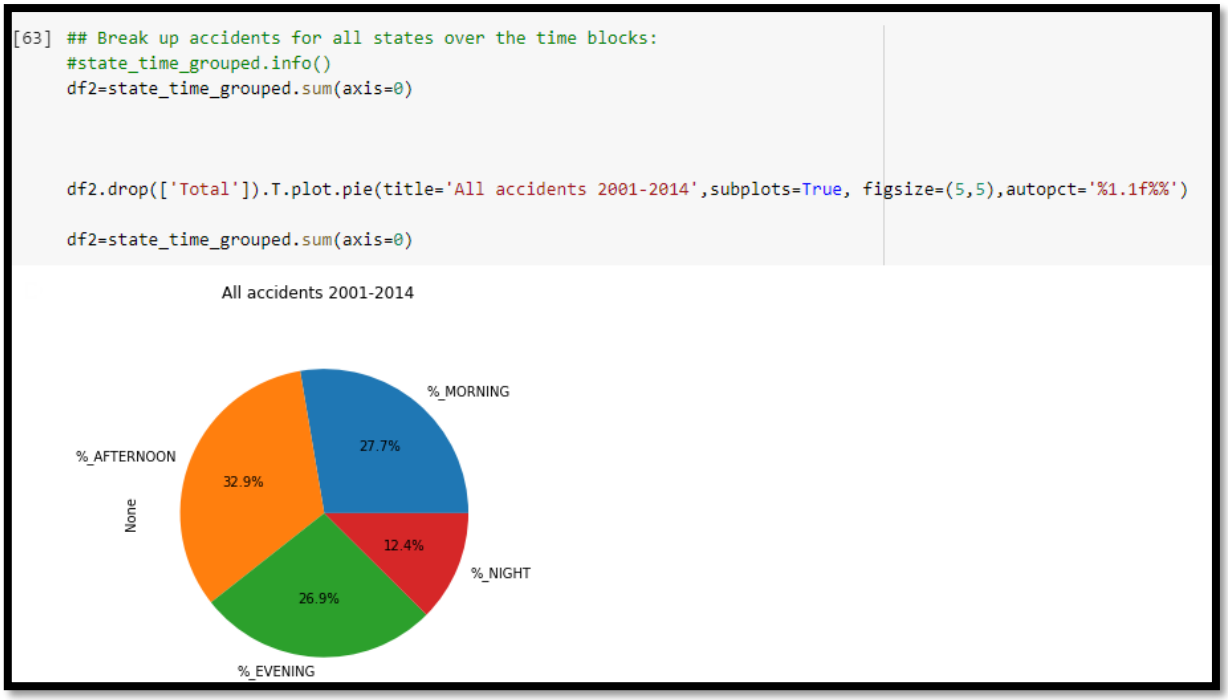

```
[61] highest_accident_states=state_grouped.sort_values('TOTAL', ascending=False)
     state_list=list(highest_accident_states.head().index)
     print(state_list)

     ['Tamil Nadu', 'Maharashtra', 'Karnataka', 'Andhra Pradesh', 'Kerala']
```

3. Time Wise Distribution of Accident Count:

- The pie charts given below contains the data of the total number of accidents occurring in the different states\UT during different time period of a day which will help us in predicting the most accident-prone time period of the day in different states\UT.



- The pie chart given below contains the data of the total number of accidents occurring during different time period of a day in the country which will help us in predicting the most accident-prone time period of the day in the country.

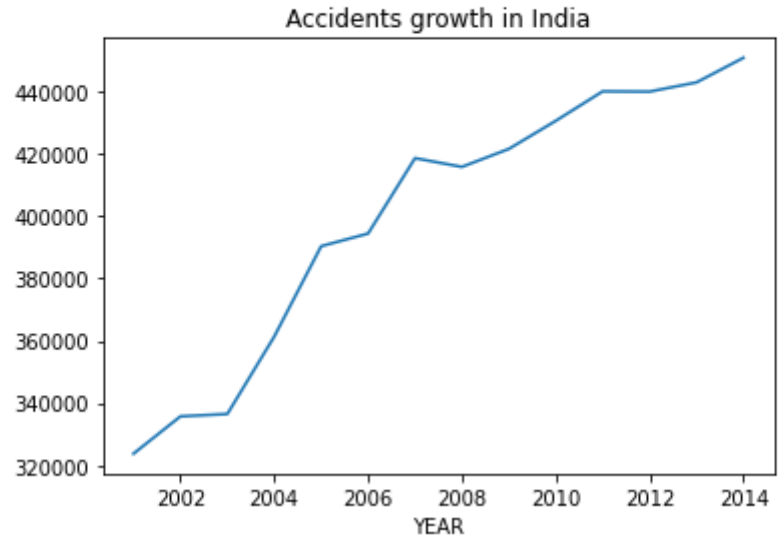


4. Accident Growth in India:

- The graph given below shows the increase in the accidents across the country by each passing year.

```
[64] df3=state_year_time_df.groupby(['YEAR']).sum()
     df3.loc[:,'Total'].plot(title='Accidents growth in India')
```

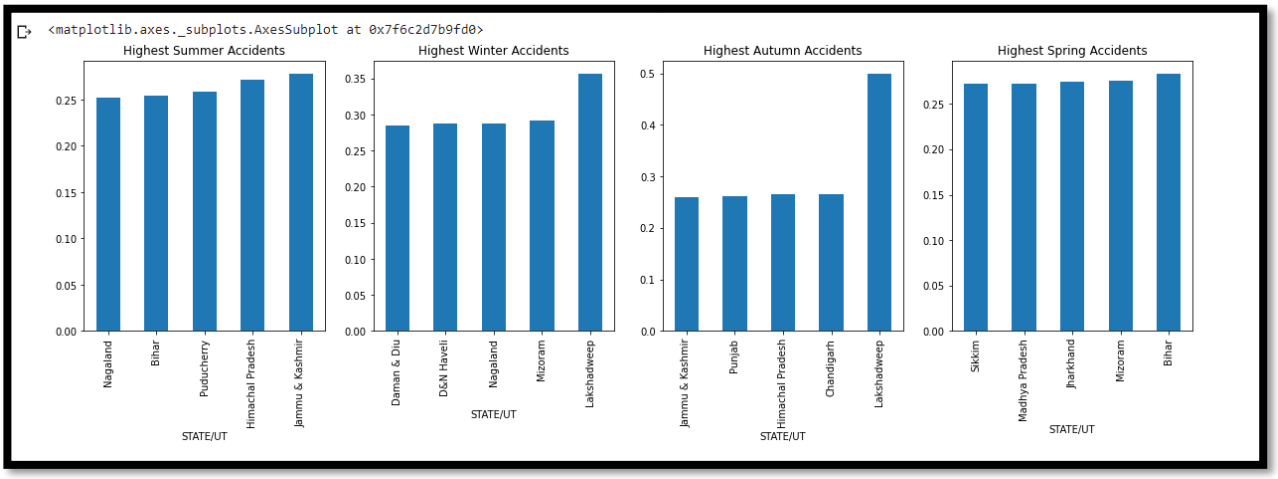
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6c2dfea8d0>
```



5. States Wise Distribution of Accident Count in Different Seasons:

- The bar graphs given below contains the data of the total number of accidents occurring in the different states\UT during different seasons of a year across the country which will help us in predicting the most accident-prone seasons in different states\UT of India.

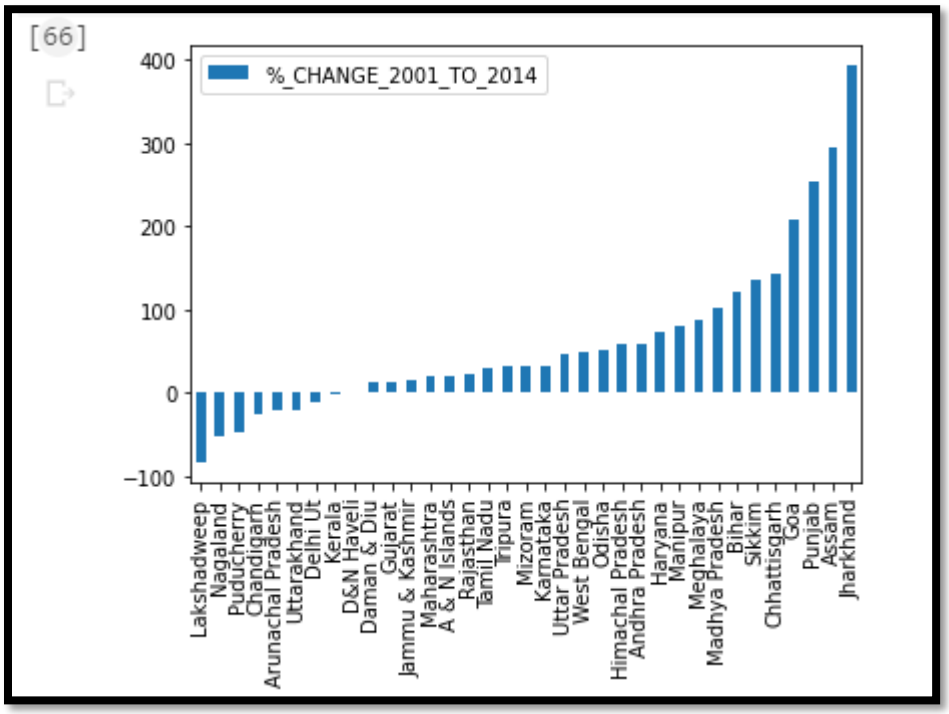
```
plt.figure(figsize=(20,5))
plt.subplot(141)
summer_sorted=state_grouped.sort_values('%_SUMMER')
summer_sorted['%_SUMMER'].tail(5).plot.bar(title='Highest Summer Accidents')
plt.subplot(142)
winter_sorted=state_grouped.sort_values('%_WINTER')
winter_sorted['%_WINTER'].tail(5).plot.bar(title='Highest Winter Accidents')
plt.subplot(143)
autumn_sorted=state_grouped.sort_values('%_AUTUMN')
autumn_sorted['%_AUTUMN'].tail(5).plot.bar(title='Highest Autumn Accidents')
plt.subplot(144)
spring_sorted=state_grouped.sort_values('%_SPRING')
spring_sorted['%_SPRING'].tail(5).plot.bar(title='Highest Spring Accidents')
```



6. Checking Performance of States\UT from 2001 to 2014:

- The bar graphs given below tracks the performance of each state\UT of India during the period of 2001-2014. This will help us in tracing the state with highest increase in accident count between different states\UT of India.

```
[65] #Create a new dataframe - period_performance.
     period_performance=pd.DataFrame(columns=['STATE/UT','%_CHANGE_2001_TO_2014'])

     #Take one state name at a time,
     for state in state_names:
         #print(state)
         total_2001=state_year_month_df.loc[(state_year_month_df['STATE/UT']==state) & (state_year_month_df['YEAR']==2001), 'TOTAL']
         total_2014=state_year_month_df.loc[(state_year_month_df['STATE/UT']==state) & (state_year_month_df['YEAR']==2014), 'TOTAL']
         value_2001=total_2001.iloc[0]
         value_2014=total_2014.iloc[0]
         change_in_percent= (value_2014-value_2001)*100/value_2001

         new_data=pd.Series({'STATE/UT':state, '%_CHANGE_2001_TO_2014':change_in_percent})
         period_performance=period_performance.append(new_data, ignore_index=True)

[66] best_performing=period_performance.sort_values('%_CHANGE_2001_TO_2014')
     #print(best_performing.head())
     ax=best_performing.plot(kind='bar').set_xticklabels(best_performing['STATE/UT'])
```



CONCLUSION

In this project we did a detailed analysis on the road accidents across the country using python's machine learning algorithms on the basis of various different attributes. The analysis helped us in predicting accident-prone states\UT, time period in a day, seasons in a year and months across the country and in individual states. The project also comments on the overall growth of the accident count in India during the time period of 2001-2014 and helps us in tracking the performance of individual states over the same time period. All the above gained information will help the central government as well as state government in preparing required action plans to tackle this problem by studying the unique patterns.

REFERENCES

- Dataset - https://www.kaggle.com/manugupta/road-accidents-in-india?select=only_road_accidents_data_month2.csv
- https://morth.nic.in/road-accident-in-india
- https://www.who.int/entity/healthinfo/global_burden_disease/GHE_Deaths_2012_country.xls?ua=1
- G. GURURAJ (2008), Road traffic deaths, injuries and disabilities in India: Current scenario, The National Medical Journal of India, volume 21, no 1, page 116
- https://www.who.int/violence_injury_prevention/road_safety_status/country_profiles/india.pdf