JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ENSEMBLE MACHINE LEARNING ALGORITHMS BASED ON ROAD TRAFFIC ACCIDENT DATA PREDICTION

[1]Gonda Ranjith, [2]Vislavath Praveen Kumar, [3]Sapavath Ramesh, [4]Dr. V Saravana Kumar

[1,2,3]IV Year Students, Dept. of IT, Sreenidhi Institute of Science and Technology(A), Hyderabad, India

[4] Associate Professor, Dept. of IT, Sreenidhi Institute of Science and Technology(A), Hyderabad,India

Abstract: Traffic is a major reason for road accidents. Due to road accidents occurred injuries and lives loss both. So safe driving and observe the road traffic to find information regarding road accidents. If you understand this situation, study road accidents and it helped us develop novel strategies to avoid road accidents. So many factors like road conditions, and traffic accidents impact accidents. To overcome this problem, make an accident prediction model. In our research, we use machine learning and ensemble learning. In our research study, compare all models and ensemble models with the road traffic accident dataset. We find the accuracy of all models. We observe support vector machines and decision trees predict a lower accuracy rate compared with other models. Ensemble models also do not give much accuracy compared to individual models. Finally, extra trees predict the highest accuracy rate.

Keywords: Ensemble learning, Road traffic accident, data prediction, Machine learning

I. INTRODUCTION

The issue of road accidents creates fear in common people because of the loss of their lives. Road accidents damage public life with multiple injuries [1]. So many factors affect such types of road accidents like environmental conditions, road designs, driver behavior, and vehicle conditions [4]. Major parameters associated with analysis of accidental data [2]. Different types of accidental data generate a job analysis through the framework. Accident data analysis interrupts the human life [3]. Using professional knowledge measure the heterogeneity data. Road accidents are divided into different clusters based on similarity. The data partition is useful for overcoming the dissimilarity of the accident data [1]. To provide safety rules for drivers, cautious road traffic statistics make it tough to find variables that are connected to road accidents [5]. In the past building data mining techniques to find high accidental places and recognize different factors that affect road accidents at dissimilar locations. Accident locations are divided into different clusters with the support of different clustering algorithms [6]. The research examines the responsibility of human, road, vehicle, and infrastructure correlation calculated by using data mining methods for road accident data [7].

In practical implementation of road accident records finalize based on accuracy, data analysis, and record retention [8]. These accidents affect on society in a huge number of families. Drivers' health is also caused by road accidents. Solving such types of problems using different types of techniques [9]. In a recent study locations of villages had less accidental rate. But in cities, the accident rate is higher than in villages. Residential zones probably higher accidental rate due to the high speed of vehicles with more public roads [10]. In undeveloped countries, the road accident rate is very high due to insufficient infrastructure and economy. Road accidents and safety are a major concern throughout the world, most researchers have been trying to solve this issue for a long time. Road traffic and uncontrolled driving occur in every part of the world [11]. Many pedestrians' are affected with no fault and they become victims due to road traffic accidents. Different factors affect most of road accidents like human faults, weather conditions, road conditions, and sharp curves [12].

The following paper continues with section 2 for the proposed architecture. Section 3 discusses with results and analysis. Section 4 describes the comparative study of machine learning algorithms. Section 5 concludes the paper.

II. PROPOSED ARCHITECTURE

The primary objectives of the Road Safety Policy in India are to reduce road traffic accidents, minimize fatalities and injuries resulting from road accidents, and enhance road infrastructure to make it safer and more efficient [13].

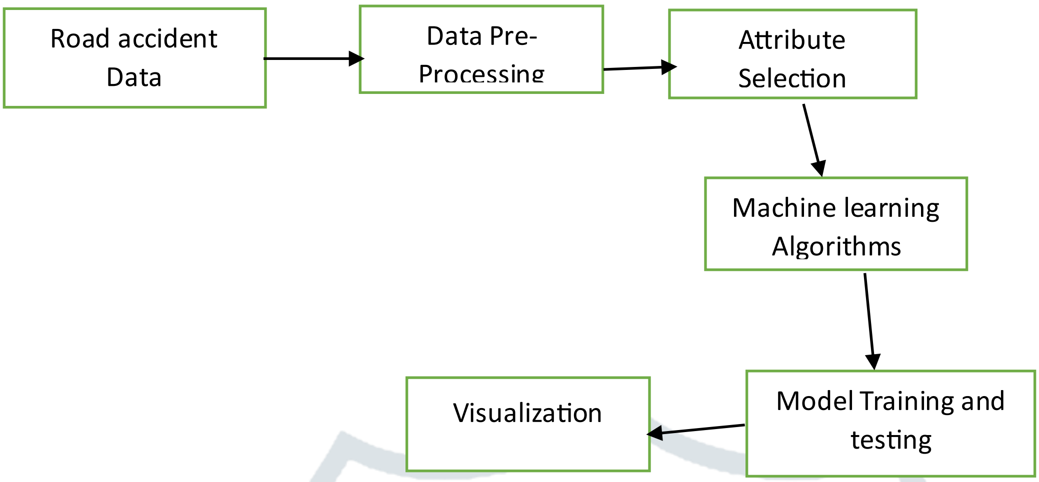

Figure 1: Proposed system architecture

The following Figure 1 provides information on the different phases of our proposed architecture. The following phases are 1. Road accident data (input), 2. Data Preprocessing (remove abnormal data), 3. Attribute selection (apply redundancy algorithms), 4. Selection of Machine Learning algorithms (suitable algorithm selection), 5. Build model and training and 6. Predict the result (visualization) [15].

III. RESULTS AND ANALYSIS

3.1 Dataset Description

The following data was collected from Addis Ababa Sub-city police departments for research work. Upload the data into the system for execution.

Table 1: Sample Dataset(part-1)

	Time	Day_of_week	Age_band_of_driver	Sex_of_driver	Educational_level	Vehicle_driver_relation	Driving_experience	Type_of_vehicle	Owner_of_vehicle	Service_year_of_vehicle	Defect_of_vehicle
0	17:02:00	Monday	18-30	Male	Above high school	Employee	1-2yr	Automobile	Owner	Above 10yr	No defect
1	17:02:00	Monday	31-50	Male	Junior high school	Employee	Above 10yr	Public (> 45 seats)	Owner	5-10yrs	No defect
2	17:02:00	Monday	18-30	Male	Junior high school	Employee	1-2yr	Lorry (41?100Q)	Owner	NaN	No defect
3	1:06:00	Sunday	18-30	Male	Junior high school	Employee	5-10yr	Public (> 45 seats)	Governmental	NaN	No defect
4	1:06:00	Sunday	18-30	Male	Junior high school	Employee	2-5yr	NaN	Owner	5-10yrs	No defect

3.2 Data Preprocessing

Data preprocessing is the crucial procedure for the removal of abnormal values. For this purpose, use different techniques based on requirements.

3.3 Data Visualization

Data visualization is a critical stage for displaying the data in a certain format. It may be represented in different types of graphs.

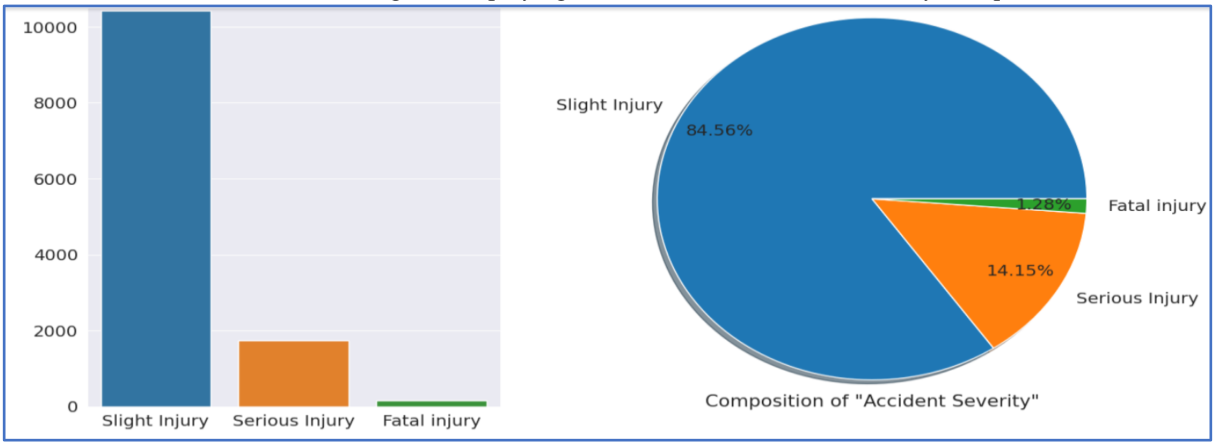

Figure 2: statistical percentage of injuries

The following table 2 checks the numerical statistics of our data.

Table 2: numerical statistics of our data

	count	mean	std	min	25%	50%	75%	max
Number_of_vehicles_involved	12316.0	2.040679	0.688790	1.0	2.0	2.0	2.0	7.0
Number_of_casualties	12316.0	1.548149	1.007179	1.0	1.0	1.0	2.0	8.0

The following table 3 shows the list of issues for road traffic accident data.

Table 3: types of different issues road accident data

```
No distancing                            2263
Changing lane to the right               1808
Changing lane to the left                1473
Driving carelessly                       1402
No priority to vehicle                   1207
Moving Backward                          1137
No priority to pedestrian                 721
Other                                     456
Overtaking                                430
Driving under the influence of drugs      340
Driving to the left                       284
Getting off the vehicle improperly        197
Driving at high speed                     174
Overturning                               149
Turnover                                   78
Overspeed                                  61
Overloading                                59
Drunk driving                              27
Unknown                                    25
Improper parking                           25
Name: Cause_of_accident, dtype: int64
```

Table 4: Vehicles age for road traffic accident

```
rta_data['Service_year_of_vehicle'].value_counts()

Unknown        2883
2-5yrs         1792
Above 10yr     1324
5-10yrs        1280
1-2yr           827
Below 1yr       282
Name: Service_year_of_vehicle, dtype: int64
```

As we observe, 4 columns have more than 20% missing values. We can safely remove these columns, as these columns will not add any value to our analysis because of the high missing value rate.

Table 5: attributes of road traffic accident data

	count	unique	top	freq
Time	12316	1074	15:30:00	120
Day_of_week	12316	7	Friday	2041
Age_band_of_driver	12316	5	18-30	4271
Sex_of_driver	12316	3	Male	11437
Educational_level	11575	7	Junior high school	7619
Vehicle_driver_relation	11737	4	Employee	9627
Driving_experience	11487	7	5-10yr	3363
Type_of_vehicle	11366	17	Automobile	3205
Owner_of_vehicle	11834	4	Owner	10459
Area_accident_occured	12077	14	Other	3819
Lanes_or_Medians	11931	7	Two-way (divided with broken lines road marking)	4411
Road_allignment	12174	9	Tangent road with flat terrain	10459
Types_of_Junction	11429	8	Y Shape	4543
Road_surface_type	12144	5	Asphalt roads	11296
Road_surface_conditions	12316	4	Dry	9340
Light_conditions	12316	4	Daylight	8798
Weather_conditions	12316	9	Normal	10063
Type_of_collision	12161	10	Vehicle with vehicle collision	8774
Vehicle_movement	12008	13	Going straight	8158
Casualty_class	12316	4	Driver or rider	4944
Sex_of_casualty	12316	3	Male	5253
Age_band_of_casualty	12316	6	na	4443
Casualty_severity	12316	4	3	7076
Pedestrian_movement	12316	9	Not a Pedestrian	11390
Cause_of_accident	12316	20	No distancing	2263
Accident_severity	12316	3	Slight Injury	10415

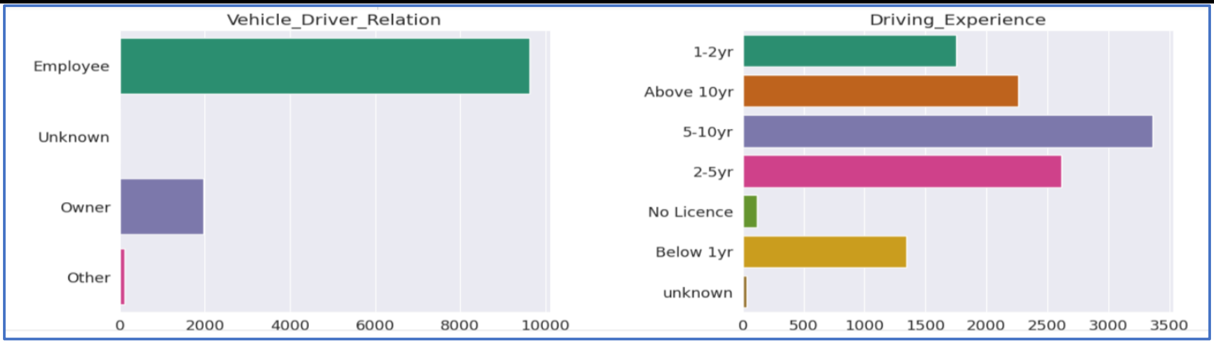

Figure 5: Vechicle driver relation and driver experience

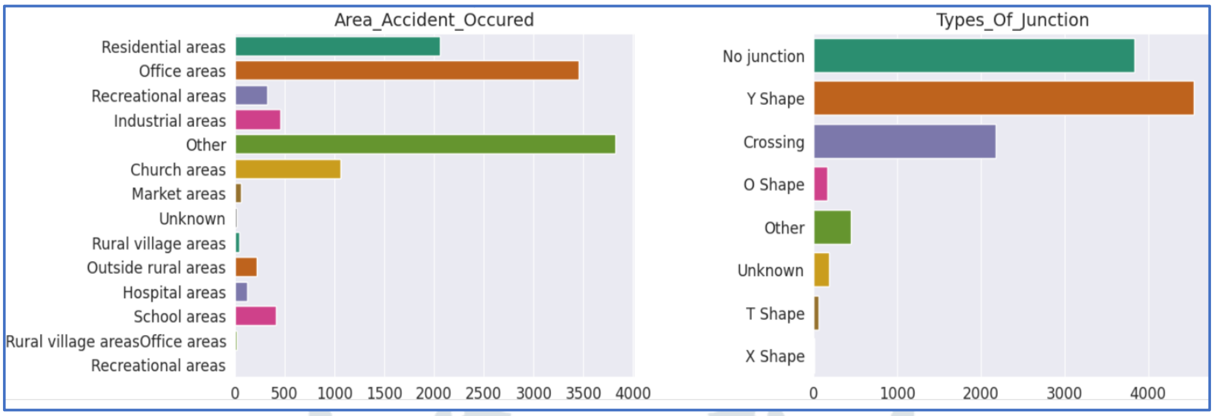

Figure 6: Area of accident and types of junctions

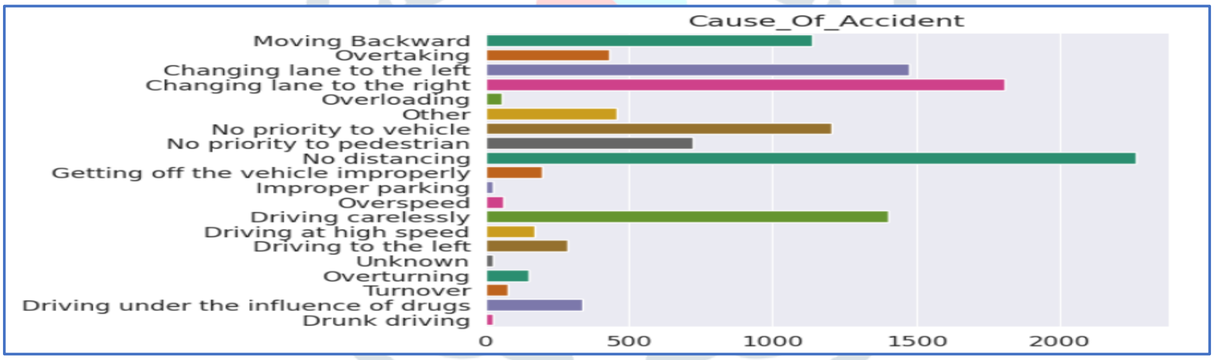

Figure 7: Visualization of data based on different road traffic accidents

3.3.1 Observations of Road Traffic Accidents

1. **Most of the accidents**

- Occurred on Friday
- Occurred at 8AM and 5PM (office & school hours)
- Occurred at two-way lines
- Sunday has a smaller number of accidents
- Severity of accident is slight injury

2. **Causality**

- Avg. Causality number is 1
- The severity range of causality is 3
- Age Range is 18-30
- Male causality is more compared to female causality
- Major causality is the driver himself
- Fatality occurred on Saturdays and Sundays.

3. **Drivers**

- Most of the drivers are male between the 18-30 age group and with 5-10 years of driving experience.
- Majority of the drivers who got into accidents are employees.
- The educational level of the driver is jr. high school.

4. **Most of the accidents occurred in personally owned passenger vehicle**

5. **Accident Area**

- Majority of accidents occurred in office areas rather than residential areas.
- Majority of accidents occurred in normal daylight and Y junction.

6. Type of Collision

- Majority of accidents occurred in vehicle-vehicle collision.
- The number of vehicles involved is 2 in the majority of accidents.
- The major cause of accidents is not keeping sufficient distance between vehicles and lane changing.

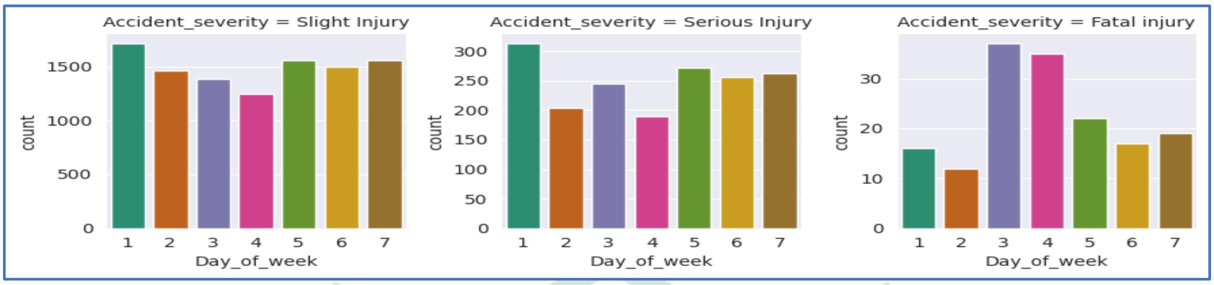

Figure 8: Accident severities basedon day

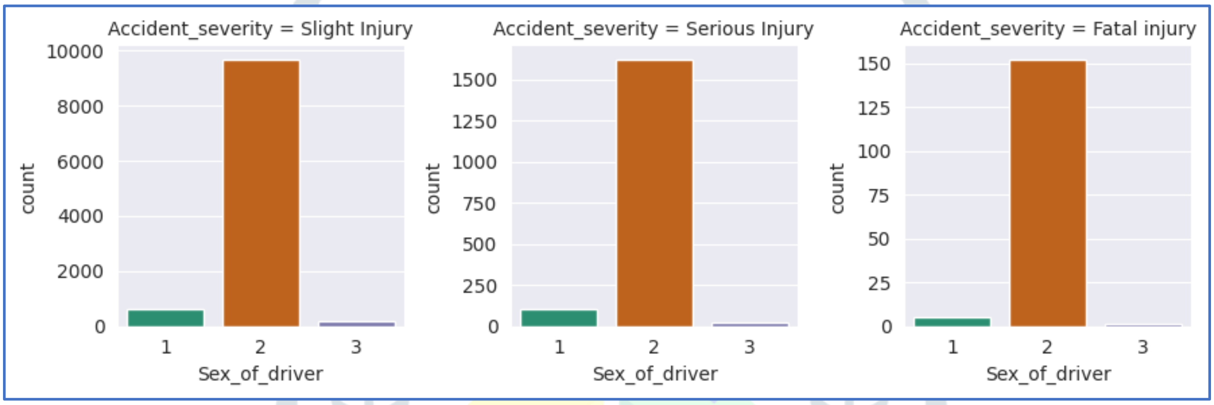

Figure 9: Accident severities based on sex type

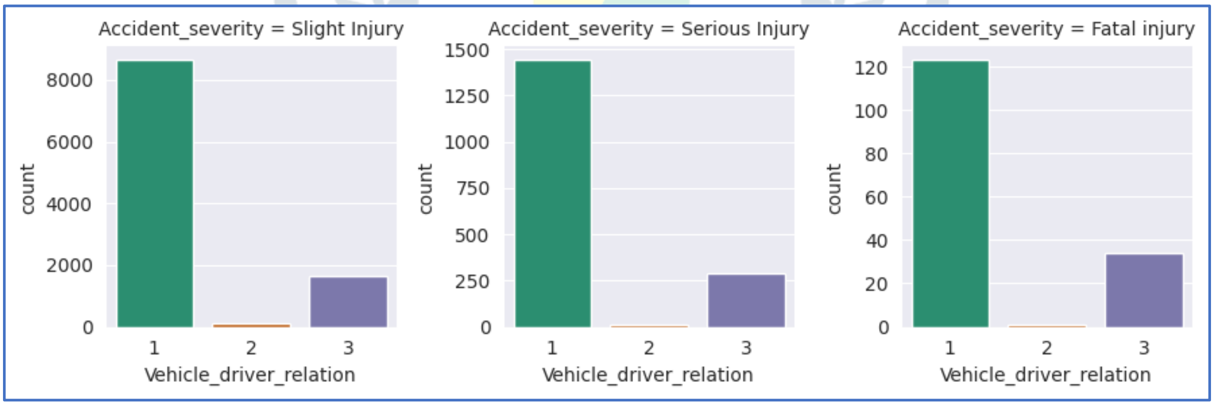

Figure 10: Accident severities based driver relation with vechicle

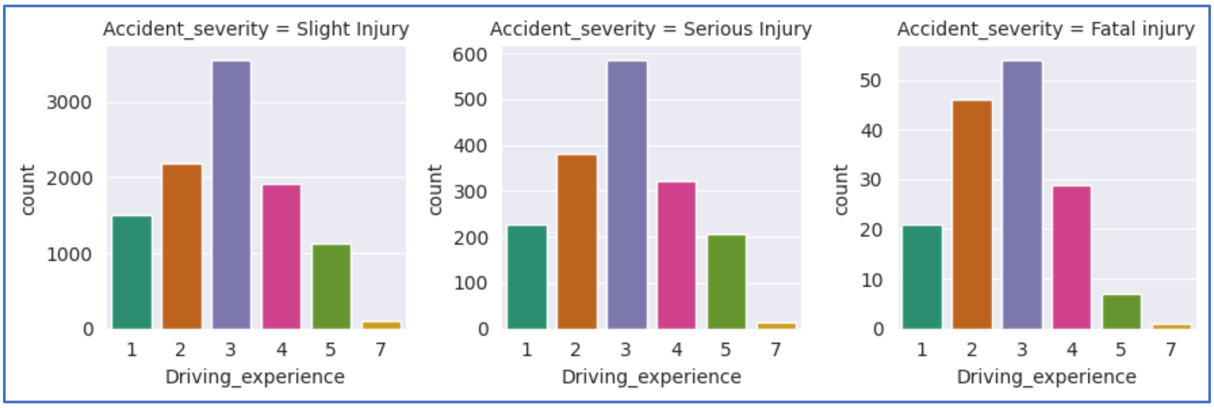

Figure 11: Accident severities based driver experience

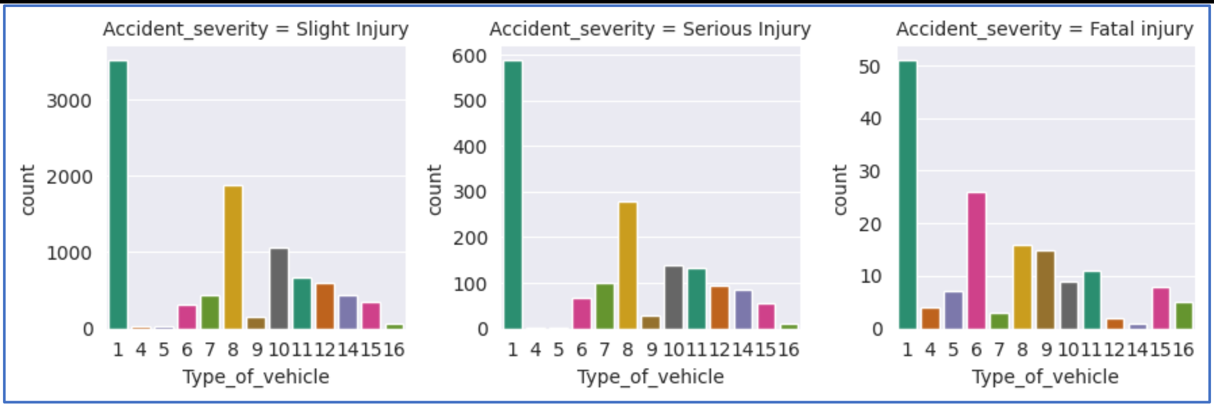

Figure 12: Accident severities based on type of vehicle

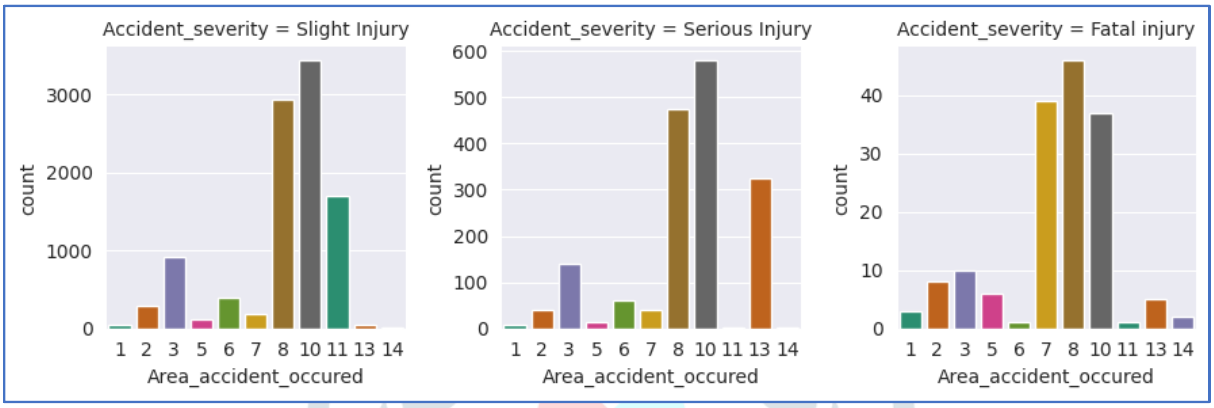

Figure 13: Accident severities based on area of accident

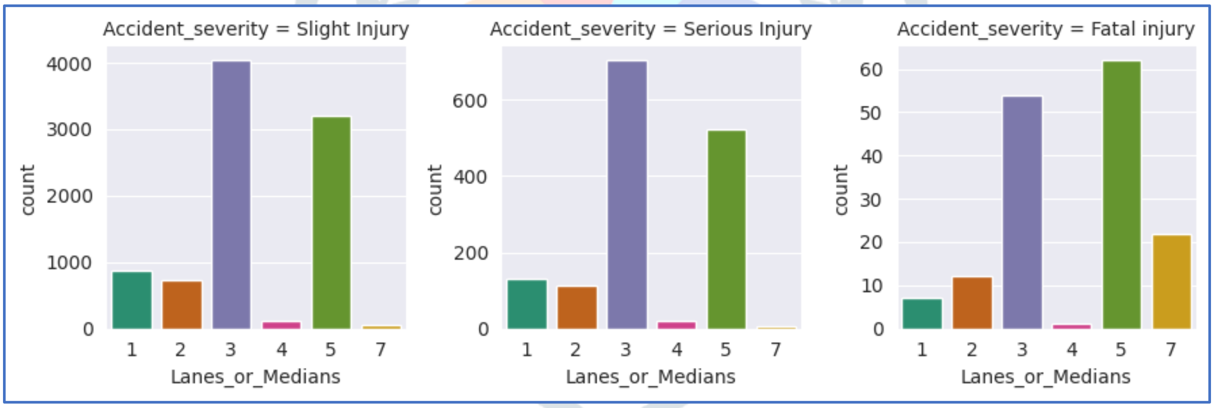

Figure 14: Accident severities based on type of lanes

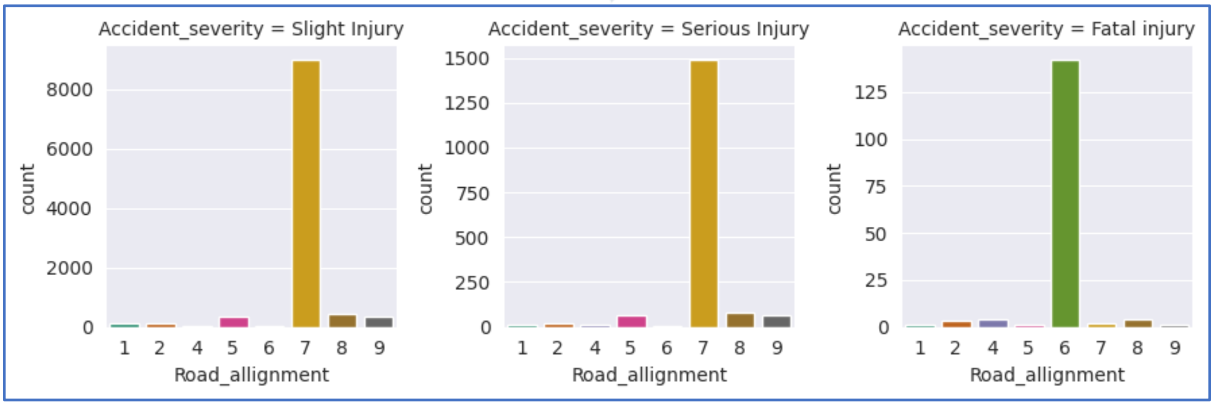

Figure 15: Accident severities based on road alignment

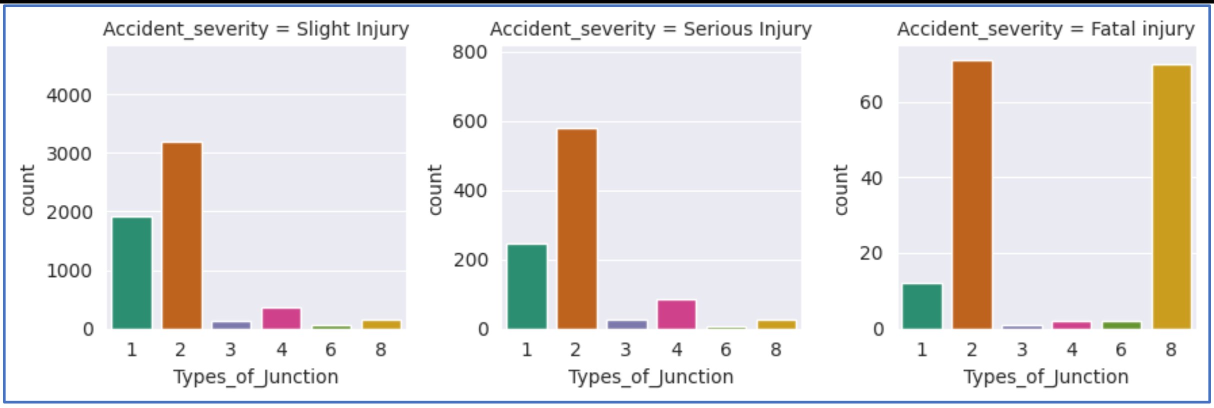

Figure 16: Accident severities based on type of junction

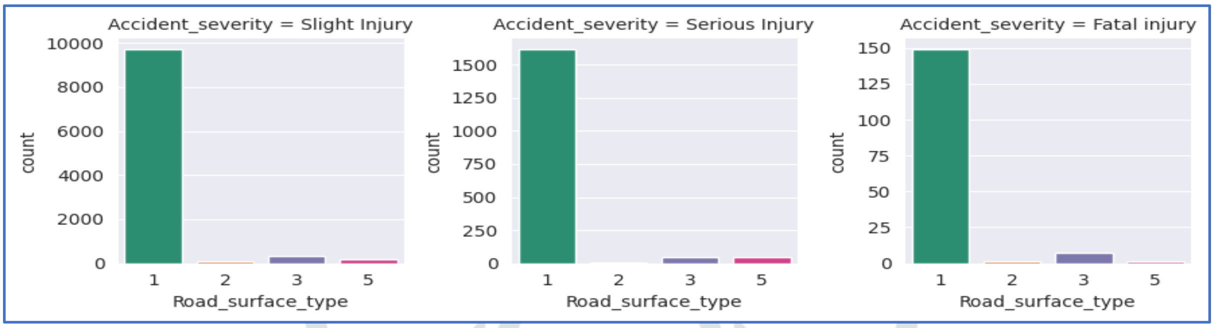

Figure 17: Accident severities based on road surface type

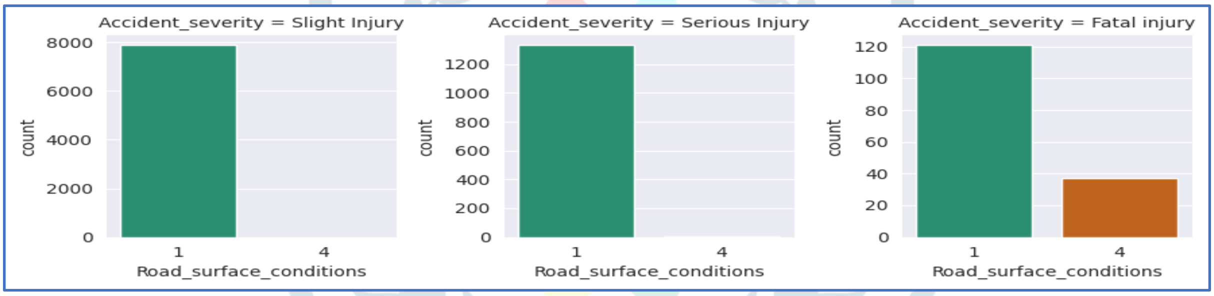

Figure 18: Accident severities based on weather conditions

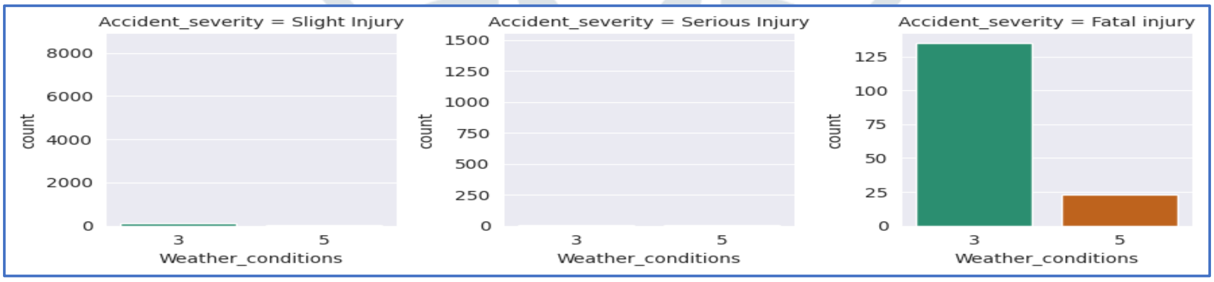

Figure 19: Accident severities based on weather conditions

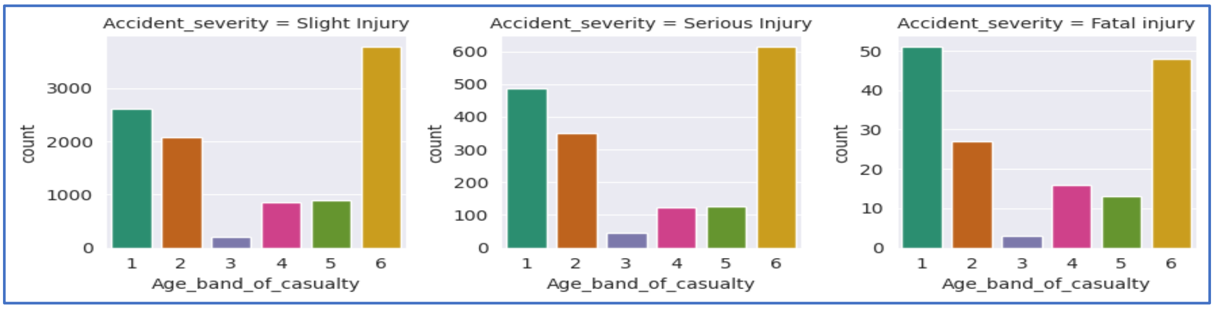

Figure 20: Accident severities based on ageof vehicle

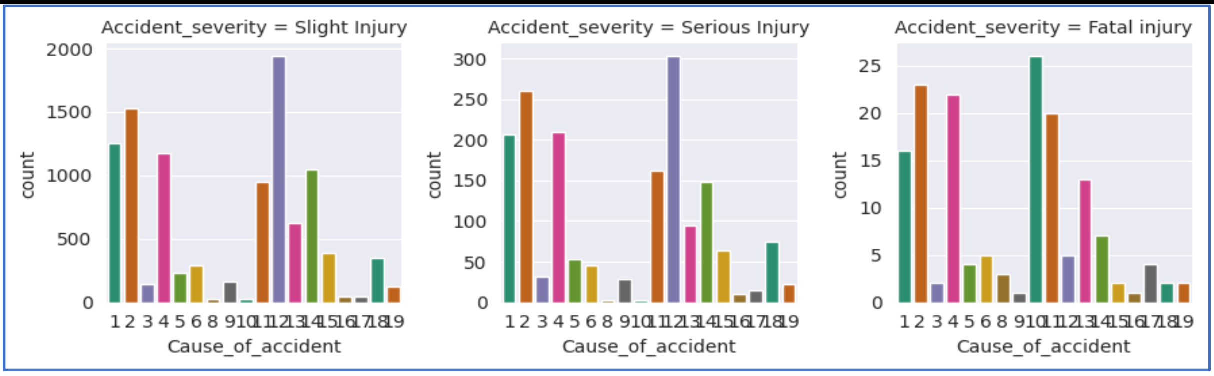

Figure 21: Accident severities based on cause of accident

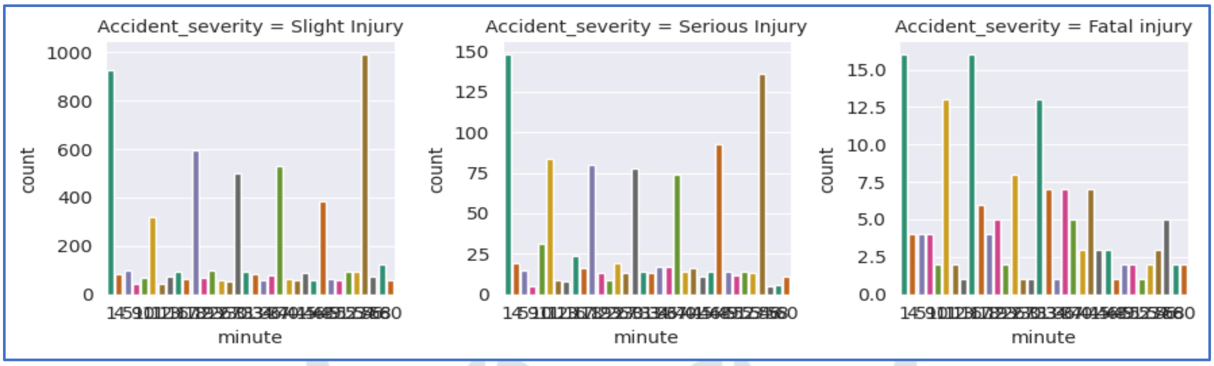

Figure 22: Accident severities based on time in minute

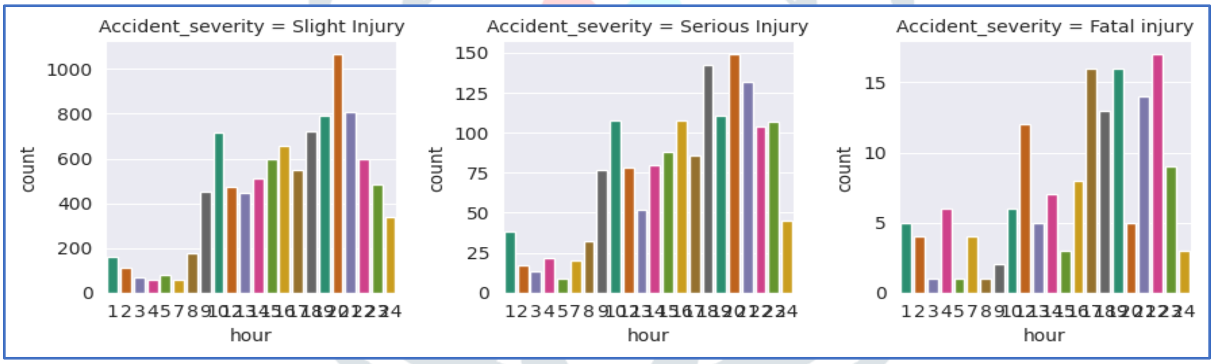

Figure 23: Accident severities based on time in an hour

3.3.2 Correlation

A heatmap (aka heat map) depicts values for a main variable of interest across two axis variables as a grid of colored squares. The axis variables are divided into ranges like a bar chart or histogram, and each cell's color indicates the value of the main variable in the corresponding cell range [14].

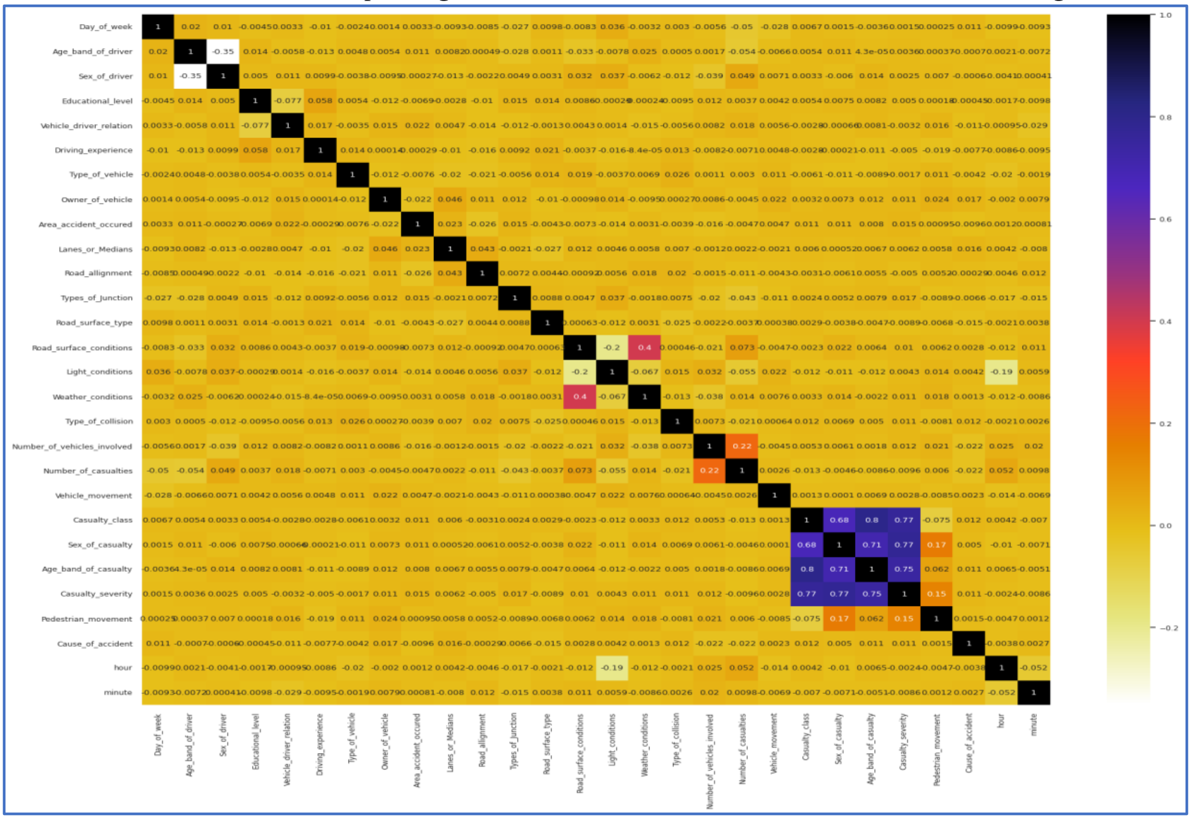

Figure 24: heatmap for data visualization

IV. COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS

4.1 Gradient Boosting Classifier

Table 6: Classification report of gradient boosting

The classification report:

	precision	recall	f1-score	support
1	0.16	0.15	0.16	52
2	0.27	0.20	0.23	552
3	0.85	0.89	0.87	3091
accuracy			0.78	3695
macro avg	0.43	0.42	0.42	3695
weighted avg	0.76	0.78	0.77	3695

4.2 Random Forest Classifier

Table 7: Classification report of random forest

The classification report:

	precision	recall	f1-score	support
1	0.44	0.08	0.13	52
2	0.30	0.21	0.24	552
3	0.86	0.91	0.88	3091
accuracy			0.80	3695
macro avg	0.53	0.40	0.42	3695
weighted avg	0.77	0.80	0.78	3695

4.3 Decision Tree Classifier

Table 8: Classification report of decision tree

The classification report:

	precision	recall	f1-score	support
1	0.22	0.38	0.28	52
2	0.23	0.36	0.28	552
3	0.87	0.78	0.82	3091
accuracy			0.71	3695
macro avg	0.44	0.50	0.46	3695
weighted avg	0.76	0.71	0.73	3695

4.4 Logistic Regression

Table 9: Classification report of logistic regression

The classification report:

	precision	recall	f1-score	support
1	0.04	0.50	0.07	52
2	0.17	0.30	0.22	552
3	0.86	0.56	0.68	3091
accuracy			0.52	3695
macro avg	0.36	0.45	0.32	3695
weighted avg	0.74	0.52	0.60	3695

4.5 Support Vector Machine

Table 10: Classification report of support vector machine

The classification report:

	precision	recall	f1-score	support
1	0.03	0.31	0.06	52
2	0.17	0.29	0.21	552
3	0.85	0.62	0.72	3091
accuracy			0.57	3695
macro avg	0.35	0.40	0.33	3695
weighted avg	0.74	0.57	0.63	3695

4.6 Extra Trees Classifier

Table 11: Classification report of extra trees

The classification report:

	precision	recall	f1-score	support
1	0.67	0.04	0.07	52
2	0.29	0.13	0.18	552
3	0.85	0.95	0.89	3091
accuracy			0.81	3695
macro avg	0.60	0.37	0.38	3695
weighted avg	0.76	0.81	0.78	3695

Table 12: Checking the accuracy score of different models

	Model	Acc_Score
5	ExtraTreesClassifier	0.8103
1	Random Forest Classifier	0.7973
0	Gradient Boosting Classifier	0.7792
2	Logistic Regression	0.7069
4	SVC	0.5654
3	Decision Tree Classifier	0.5210

4.5 Ensemble learning

4.5.1 Ensemble model (Extra Trees + Random Forest)

```
from sklearn.ensemble import VotingClassifier

extree = ExtraTreesClassifier()
rfc = RandomForestClassifier(random_state = 0)

ensemble_model = VotingClassifier(estimators=[('extra_tree', extree), ('random_forest', rfc)], voting='hard')

ensemble_model.fit(X_train, y_train)

# Make predictions on the testing data
predictions = ensemble_model.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, predictions)
print(f"Accuracy: {accuracy}")

Accuracy: 0.7878213802435724
```

4.5.2 Ensemble model (Gradient Boost + Logistic Regression)

```
# Define and train the Gradient Boosting model
gb_model = GradientBoostingClassifier(n_estimators=100, max_depth=3, random_state=0)
gb_model.fit(X_train, y_train)

# Define and train the Logistic Regression model
lr_model = LogisticRegression(C=1.0, penalty='l2', random_state=0)
lr_model.fit(X_train, y_train)

# Make predictions using both models
gb_predictions = gb_model.predict(X_test)
lr_predictions = lr_model.predict(X_test)

# Combine predictions using a simple averaging approach
ensemble_predictions = (gb_predictions + lr_predictions) / 2

# Round the predictions to the nearest integer (assuming classes are integers)
ensemble_predictions = ensemble_predictions.round().astype(int)

# Evaluate the performance of the ensemble model
ensemble_accuracy = accuracy_score(y_test, ensemble_predictions)
print(f"Ensemble Model Accuracy: {ensemble_accuracy}")

Ensemble Model Accuracy: 0.5190798376184033
```

Table 13: comparative study of ensemble models vs individual models

S. No.	Name of the Model	Accuracy in %
1	Extra trees	81
2	Random forest Tree	79.7
3	Gradient Boosting	77.9
4	Logistic regression	70.6
5	Support vector Machine	56.5
6	Decision Trre	52.1
7	Extra Trees+ Random Forest	78.7
8	Gradient Boosting +Logistic Regression	51.9

In our research study, compare all models and ensemble models with the road traffic accident dataset. We find the accuracy of all models. We observe support vector machines and decision trees predict a lower accuracy rate compared with other models. Ensemble models also do not give much accuracy compared to individual models. Finally, extra trees predict the highest accuracy rate.

V. CONCLUSION

Traffic is a major reason for road accidents. Due to road accidents occurred injuries and lives loss both. So safe driving and observe the road traffic to find information regarding road accidents. If you understand this situation, study road accidents and it helped us develop novel strategies to avoid road accidents. So many factors like road conditions, and traffic accidents impact accidents. To overcome this problem, make an accident prediction model. In our research, we use machine learning and ensemble learning. From our research study, compare all models and ensemble models with the road traffic accident dataset. From our research study, compare all models and ensemble models with the road traffic accident dataset. We find the accuracy of all models. We observe support vector machines and decision trees predict a lower accuracy rate compared with other models. Ensemble models also do not give much accuracy compared to individual models. Finally, extra trees predict the highest accuracy rate.

REFERENCES

[1] Sachin K and D Toshniwal," A data mining framework to analyze road accident data", Journal of Big Data, 2005.

[2] Savolainen P, Mannering F, and Quddus," The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives", Accid Anal Prev., 43:1666–76,2011.

[3] Depaire B, Wets G, and Vanhoof K," Traffic accident segmentation using latent class clustering, accident analysis, and prevention", vol. 40, Elsevier, 2008.

[4] Karlaftis M and Tarko A," Heterogeneity considerations in accident modeling, Accid Anal Prev, 30(4):425–33,1998.

[5] Ma J and Kockelman K," Crash frequency and severity modeling using clustered data from Washington state", IEEE Intelligent Transportation Systems Conference, Toronto, Canadá,2008.

[6] Jones B and Janssen L," Analysis of the frequency and duration of freeway accidents in Seattle, accident analysis and prevention", Elsevier, vol. 231991.

[7] Miaou SP and Lum H," Modelling vehicle accidents and highway geometric design relationships, accident analysis and prevention", Elsevier, vol. 25,1993.

[8] Subhani Shaik, "DM Algorithms Based Clustering for Road Accident Data Analysis," International Journal of Computer Sciences and Engineering, Vol.-6, Issue-9, Sept. 2018.

[9] Dr. Sunil Bhutada and Subhani Shaik, "IPL Match Prediction using Machine Learning", IJAST, Vol.29, Issue 5, April-2020.

[10] Mr. Sujan Reddy, Ms. Renu Sri and Subhani Shaik," Sentimental Analysis using Logistic Regression", International Journal of Engineering Research and Applications (IJERA), Vol.11, Series-2, July-2021.

[11] Ms. Mamatha, Srinivasa Datta and Subhani Shaik," Fake Profile Identification using Machine Learning Algorithms", International Journal of Engineering Research and Applications (IJERA), Vol.11, Series-2, July-2021.

[12] Subhani Shaik and Dr. Uppu Ravibabu, "Detection and Classification of Power Quality Disturbances Using Curvelet Transform and Support Vector Machines", in the 5th IEEE International Conference on Information Communication and Embedded System (ICICES-2016) at S.A Engineering College, Chennai, India on 25th -26th, February 2016.

[13] J. Lavanya, M. Ramesh, J. Sravan Kumar, G. Rajaramesh and Subhani Shaik," Hate Speech Detection Using Decision Tree Algorithm", Journal of Advances in Mathematics and Computer Science, Volume 38, Issue 8, Page 66-75, June-2023.

[14] Neeraja, Anupam, Sriram, Subhani Shaik and V. Kakulapati," Fraud Detection of AD Clicks Using Machine Learning Techniques", Journal of Scientific Research and Reports, Volume 29, Issue 7, Page 84-89, June-2023.

[15] P. Pranathi, V. Revathi, P. Varshitha, Subhani Shaik and Sunil Bhutada," Logistic Regression Based Cyber Harassment Identification", Journal of Advances in Mathematics and Computer Science, Volume 38, Issue 8, Page 76-85, June-2023.