Summer Internship

On

“A Statistical Study of Road Accidents in Various States in India”

Submitted to

Amity University Uttar Pradesh
In partial fulfillment of requirements for the award of the Degree of
(M.Sc. Data Science)

By
Shivam Gupta
Enrolment No: A044161824015

Under the Supervision of:
Dr. Vaidehi Singh
Department of Statistics

Amity Institute of Applied Sciences
Amity University Noida, Uttar Pradesh
Batch 2024-2026

DECLARATION

I, **Shivam Gupta**, student of **M.Sc. Data Science** hereby declare that the Summer Internship titled "**A Statistical Study of Road Accidents in Various States of India**" Which is submitted by me to Department of **Statistics**, Amity Institute of Applied Sciences, Amity University, Uttar Pradesh, Noida, in partial fulfillment of requirement for the award of the degree of **M.Sc. Data Science** has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition.

Noida

Date:

Shivam Gupta

CERTIFICATE

On the basis of declaration submitted by **Shivam Gupta**, student of **M.Sc. Data Science**, I hereby certify that the Summer Internship titled "**A Statistical Study of Road Accidents in Various States of India**" which is submitted to **Department of Statistics**, Amity Institute of Applied Sciences, Amity University, Uttar Pradesh, Noida, in partial fulfillment of requirement for the award of the degree of **M.Sc. Data Science is** a faithful record of work carried out by him/them under my guidance and supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Noida

Date:

Dr. Vaidehi Singh
Department of Statistics
Amity Institute of Applied Sciences
Amity University, Uttar Pradesh, Noida

ACKNOWLEDGEMENT

It is highly privilege for me to express my deep sense of gratitude to those entire faculty members who helped me in the completion of the "**A Statistical Study of Road Accidents in Various States of India**" under the supervision of my guide Dr. Vaidehi Singh. My special thanks to all the other faculty members, batchmates and seniors of AIAS, Amity University Uttar Pradesh for helping me in the completion of the project work and its report submission.

Shivam Gupta

ABSTRACT

In India, road accidents provide a significant obstacle to infrastructure planning and public safety. This project, titled "**A Statistical Study of Road Accidents in Various States in India**" aims to analyze accident data across various Indian states to uncover key patterns and contributing factors. The study begins with extensive exploratory data analysis (EDA) to understand the distribution of accidents based on state, weather conditions, time, vehicle types, and other variables. Subsequently, machine learning techniques are employed to predict the severity of accidents using selected features. The findings from this project can assist traffic authorities to identify high-risk locations and executing focused safety protocols. The study offers valuable insights for enhancing road safety in India by combining statistical techniques with machine learning.

List of Figures

LIST OF TABLES

TABLE OF CONTENT

1. INTRODUCTION

Traffic crashes are a substantial public health and development challenge in India and globally. India has one of the highest rates of road fatalities globally with more than 150,000 lives being lost every year in road traffic accidents," the ministry said."The Ministry of Road Transport and Highways (MoRTH) has this program has been sanctioned. Such events have both human tragedy and social and economic impact due to loss of human life, loss of productivity, increased healthcare burden, and infrastructural damage. The causes of road accidents can be attributed to the complex interaction between infrastructure, human (related to road user factors) and vehicular factors, a deeper understanding of this interaction is necessary for effective road safety strategies.

There are half a dozen reasons why there are more road accidents in Indian states.These include drunk driving, over-speeding, distracted driving and not buckling up (seatbelt/helmet). Environmental factors – inadequate road layout and visibility was the primary contributory factor in nearly half the cases (49%). Problems with the cars themselves include bad brakes or tyres, as well as the problems dealing with traffic, unmarked roads and loose enforcement of the rules on the road. Demographic characteristics, such as age, sex, and time of the day, are also important factors affecting accident risk and severity.

Road accident is such a multifaceted phenomenon, we need a statistical analysis based on data to find visual pattern, risk assessment, action plan. Title of the project Proposed" A Statistical Study of Road Accidents in Different States in India" Objectives of the Project The project is designed to analyze the road traffic accidents using Statistical tools and Machine learning techniques. Descriptive statistics are used to investigate differences in (mean/integral/percentage/ratio and their pairwise differences) with significance between states and significant contributors, and inferential methods are applied to test associations among crash characteristics. The aim is to understand how various factors like type of vehicle, weather, type of road, time etc., contribute in accidents.

Logistic Regression and Random Forest Classifier are the machine learning models to predict the severity of an accident using the extracted features.

2. Review of Literature

Road traffic accidents is a major health and safety threat for the public. With more and more access to accident data, and statistical analysis and machine learning techniques become increasingly popular in studies regarding factors influencing accident severity and prediction modeling.

Nandhini and Ramasamy (2019) also stressed the necessity of EDA in comprehending the road accident data structure and distribution. They conducted aggregate statistics, heatmaps and distribution plot analysis to identify trends and associations between variables.

Saxena et al. (2021) handled heavily class imbalanced accident datasets. They emphasized that under-representation of some severity classes (such as Fatal accidents) leads models to make poor predictions. The author suggested that to handle those imbalanced classes, sampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique) need to be applied before feed data to machine learning models.

Khan and Hussain (2022) presented a survey summarizing the evaluation metrics employed in the prediction of accident severity. They added that although accuracy is frequent, it may be misleading in case of unbalanced datasets. Instead, they suggested using precision, recall, F1-score, and confusion matrix to further evaluate performance of the model.

3. DATA PREPARATION

3.1 Dataset Description:

The road accident dataset comprises detailed records of road accidents reported across various Indian states. The dataset includes over 3,000 records (exact count to be filled after loading) and consists of approximately 20 attributes, each representing key characteristics of individual accident incidents.

The dataset captures a combination of geographical, environmental, temporal, vehicular, and demographic features that influence accident occurrence and severity.

Road and environmental features can include things like:

- Weather
- Road Type (such as an urban road or a national highway)
- Observation
- Condition of the Lighting

Vehicular and involvement-related fields may include:

- Vehicle Type(e.g., Two-Wheeler, Truck, Bus, Car)
- Number of Vehicles Involved
- Number of Persons Involved or Injured
- Use of Safety Gears(Helmet/Seatbelt)

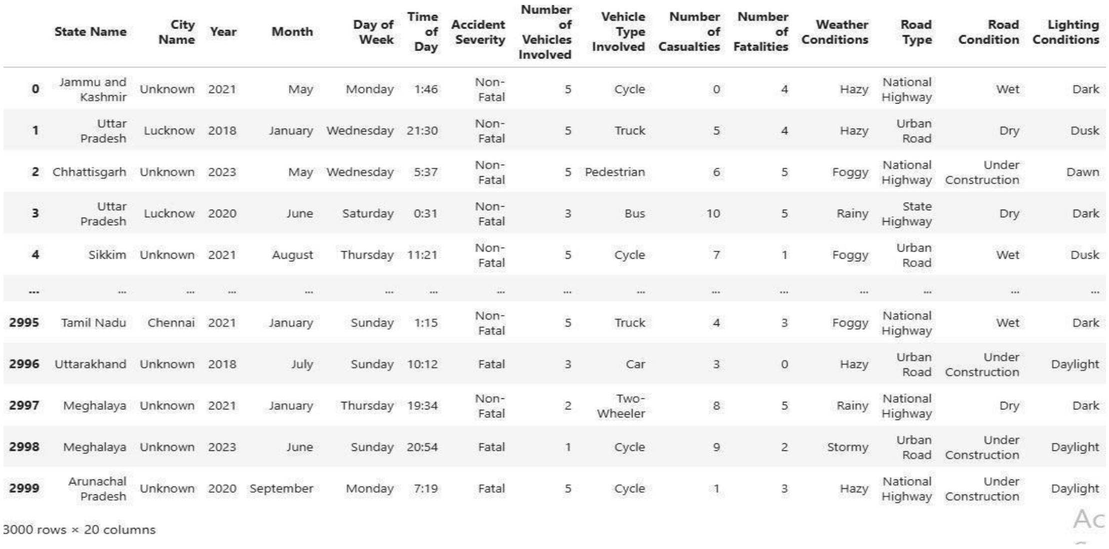
Demographic and behavioral characteristics, if present, consist of:

- Driver's age, gender, alcohol involvement, and current driver's license.

Accident Severity is the target variable and it can be multiclass (Minor, Serious, and Fatal) or binary (Fatal vs. Non-Fatal).

3.2 Collection of Data

This data was collected from Kaggle, which is an online community stage for information researchers and machine learning devotees to discover and distribute datasets.



	State Name	City Name	Year	Month	Day of Week	Time of Day	Accident Severity	Number of Vehicles Involved	Vehicle Type Involved	Number of Casualties	Number of Fatalities	Weather Conditions	Road Type	Road Condition	Lighting Conditions
0	Jammu and Kashmir	Unknown	2021	May	Monday	1:46	Non-Fatal	5	Cycle	0	4	Hazy	National Highway	Wet	Dark
1	Uttar Pradesh	Lucknow	2018	January	Wednesday	21:30	Non-Fatal	5	Truck	5	4	Hazy	Urban Road	Dry	Dusk
2	Chhattisgarh	Unknown	2023	May	Wednesday	5:37	Non-Fatal	5	Pedestrian	6	5	Foggy	National Highway	Under Construction	Dawn
3	Uttar Pradesh	Lucknow	2020	June	Saturday	0:31	Non-Fatal	3	Bus	10	5	Rainy	State Highway	Dry	Dark
4	Sikkim	Unknown	2021	August	Thursday	11:21	Non-Fatal	5	Cycle	7	1	Foggy	Urban Road	Wet	Dusk
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2995	Tamil Nadu	Chennai	2021	January	Sunday	1:15	Non-Fatal	5	Truck	4	3	Foggy	National Highway	Wet	Dark
2996	Uttarakhand	Unknown	2018	July	Sunday	10:12	Fatal	3	Car	3	0	Hazy	Urban Road	Under Construction	Daylight
2997	Meghalaya	Unknown	2021	January	Thursday	19:34	Non-Fatal	2	Two-Wheeler	8	5	Rainy	National Highway	Dry	Dark
2998	Meghalaya	Unknown	2023	June	Sunday	20:54	Fatal	1	Cycle	9	2	Stormy	Urban Road	Under Construction	Daylight
2999	Arunachal Pradesh	Unknown	2020	September	Monday	7:19	Fatal	5	Cycle	1	3	Hazy	National Highway	Under Construction	Daylight

3000 rows × 20 columns

Table 1-The original dataset

4. EDA (EXPLORATORY DATA ANALYSIS)

The handle of classifying the information utilizing statistical and visual methods in arrange to highlight noteworthy highlights for extra examination is known as exploratory data analysis, or EDA. This incorporates examining the dataset from different viewpoints and giving a portrayal and rundown of its substance without accepting anything.

Make Rundowns: Summarize the information to get a common thought of its substance, such as normal values, common values, or esteem dispersions. Calculating quantiles and checking for skewness can give experiences into the data's distribution.

Visualize the Information: Utilize intuitively charts and charts to spot patterns, designs, or irregularities. Bar plots, diffuse plots, and other visualizations help get it variables connections. NumPy, Matplotlib, Seaborn, and pandas are the Python libraries which we used in EDA .

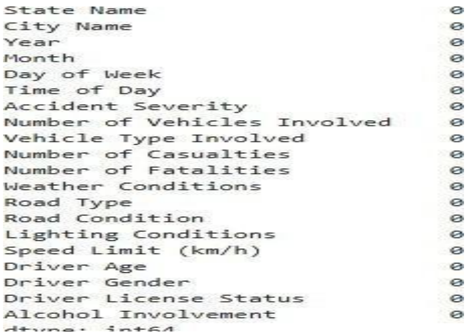
4.1 Preprocessing Data

Data preprocessing is the crucial stage for getting the dataset ready for modelling.It guarantee that the data was clean, consistent, and prepared for machine learning algorithms, the following procedures were followed.

4.2 Dealing with Missing Values

Python code was used to check the dataset for missing or null values.

df.isnull().sum()



```
State Name                    0
City Name                     0
Year                          0
Month                         0
Day of Week                   0
Time of Day                   0
Accident Severity             0
Number of Vehicles Involved   0
Vehicle Type Involved         0
Number of Casualties          0
Number of Fatalities          0
Weather Conditions            0
Road Type                     0
Road Condition                0
Lighting Conditions           0
Speed Limit (km/h)            0
Driver Age                    0
Driver Gender                 0
Driver License Status         0
Alcohol Involvement           0
dtype: int64
```

Table2 -Finding missing values

There are no missing values in any of the dataset's columns.

4.3 Elimination of Outliers

The following numerical characteristics that are likely to influence accident severity and risk were subjected to outlier analysis:

- Age of the Driver
- The quantity of casualties
- The quantity of fatalities
- Limit of Speed (km/h)

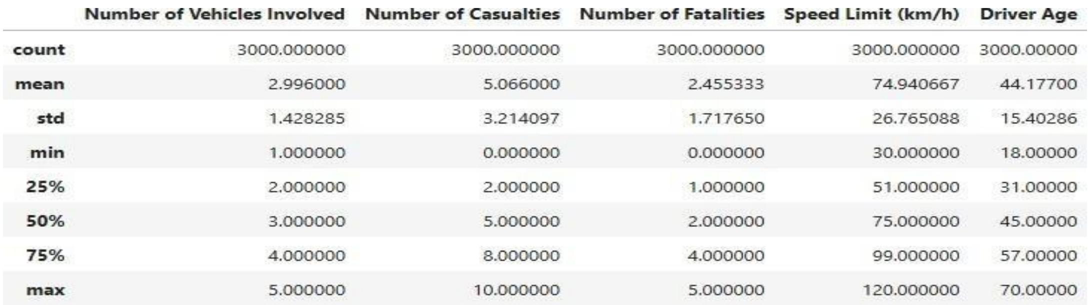
These factors were chosen because they have a direct effect on the results of accidents and are frequently employed in traffic safety analytics.

Outlier detection results:

- Shape of the original dataset: (3000, 20)
- Following IQR outlier removal: (3000, 20)

No records are removed as none of the values in the selected columns exceeded the IQR thresholds for identify outliers.

4.4 Descriptive Statistics



	Number of Vehicles Involved	Number of Casualties	Number of Fatalities	Speed Limit (km/h)	Driver Age
count	3000.000000	3000.000000	3000.000000	3000.000000	3000.00000
mean	2.996000	5.066000	2.455333	74.940667	44.17700
std	1.428285	3.214097	1.717650	26.765088	15.40286
min	1.000000	0.000000	0.000000	30.000000	18.00000
25%	2.000000	2.000000	1.000000	51.000000	31.00000
50%	3.000000	5.000000	2.000000	75.000000	45.00000
75%	4.000000	8.000000	4.000000	99.000000	57.00000
max	5.000000	10.000000	5.000000	120.000000	70.00000

Table 3 – Descriptive statistics

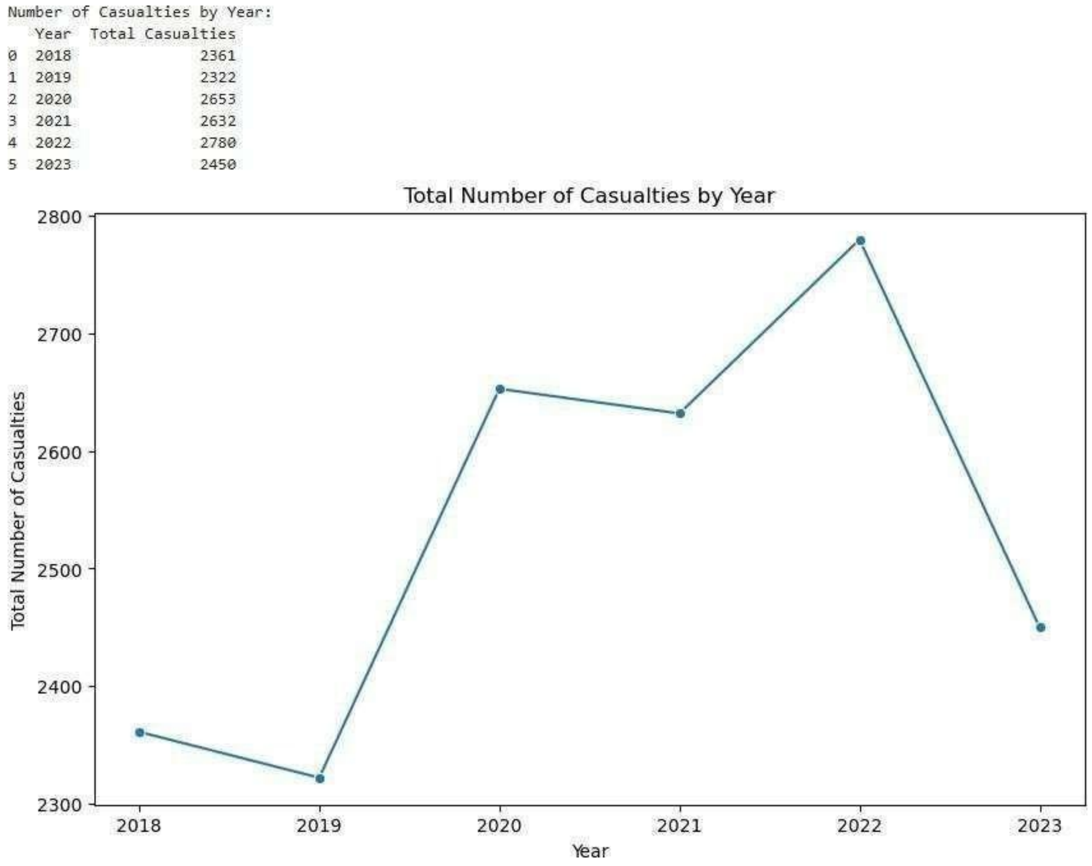
- **Number of Vehicles Involved:** Three or so vehicles are typically involved in an accident. The majority of collisions involve two to four vehicles, and no more than five.
- **Number of Casualties:** In an accident is approximately five. Although some accidents resulted in up to ten casualties.
- **Number of Fatalities**: The average fatalities is 2.46, with a maximum of five. Half of the accidents had two or fewer fatalities, as indicated by the median of 2.
- **Speed Limit (km/h)**: At accident scenes, the speed limits vary widely, ranging from 30 to 120 km/h.
- **Driver Age**: Average age of driver is 44. The dataset's drivers range in age from 18 to 70.

4.5 UNIVARIATE ANAYLSIS

Plotting Bar Charts for Categorical Columns to visualize these categorical columns, we used plots such as bar plot created using the seaborn library.

4.5.1 Road Accident Trend (2018-2023)

The line graph shows how many people died in traffic accidents in India between 2018 and 2023.The number of accidents in India from 2018 to 2023 is shown in the below graph.



```
Number of Casualties by Year:
   Year  Total Casualties
0  2018              2361
1  2019              2322
2  2020              2653
3  2021              2632
4  2022              2780
5  2023              2450
```

Fig 1 –Trend of Road Accident

Interpretation:

- **2018–2019:** Casualties slightly decreased from 2361 to 2322, suggesting marginal improvement in road safety or reporting.
- **2019–2020:** Given the COVID-19 lockdowns, it may come as a surprise that the number of casualties increased significantly to 2653; this increase may be the result of serious accidents involving heavy or necessary vehicles that were still in use during lockdowns.
- **2020–2021:** A **minor drop** to **2632** casualties, showing a temporary decline.
- **2021–2022**: With 2780 casualties, 2022 had the largest jump, maybe because of
  - o Post-pandemic traffic surges
  - o Increased reckless driving or poor infrastructure maintenance
  - o Insufficient road safety reforms
- **2022–2023:** A **significant drop** to **2450 casualties**, indicating recent **positive changes** such as:
  - o Implementation of better traffic laws
  - o Improved emergency response systems
  - o Awareness campaigns or infrastructure improvements.

4.5.2 Accidents by Driver Gender

The bar plot above displays the breakdown of traffic accidents by the gender of the driver involved.

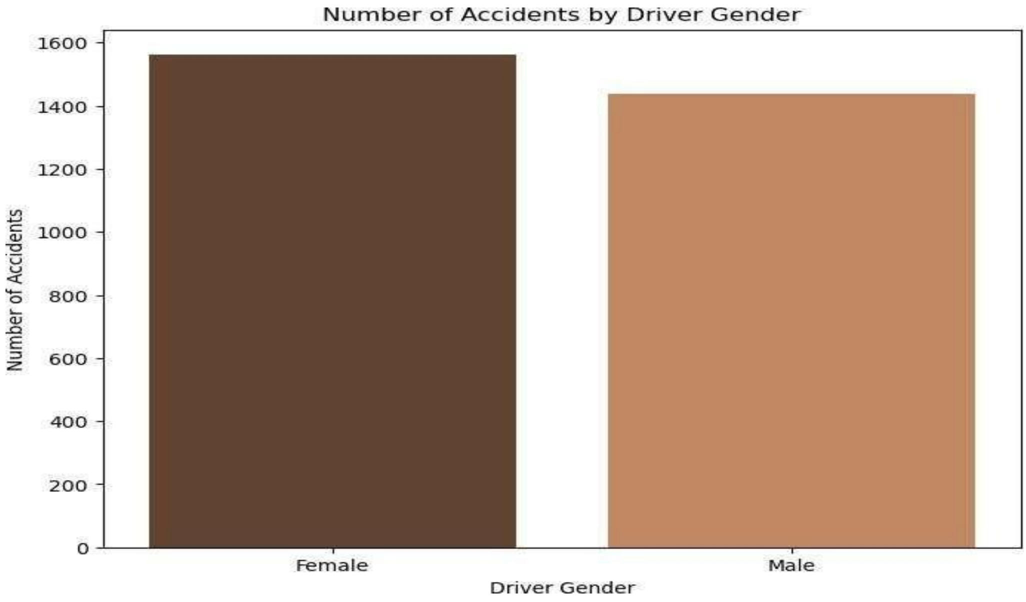


Fig2 – Accidents by Driver Gender

Interpretation:

- Female Drivers: Involved in approximately 1,570 accidents, female drivers were the most common gender in this dataset.

- Male Drivers: There were roughly 1,440 accidents involving male drivers, slightly fewer than those involving female drivers.

4.5.3 Vehicle Types Involved in Accidents

The bar chart above illustrates the different types of vehicles that have been involved in traffic accidents. It also represents the number of times each category has been involved in such incidents.

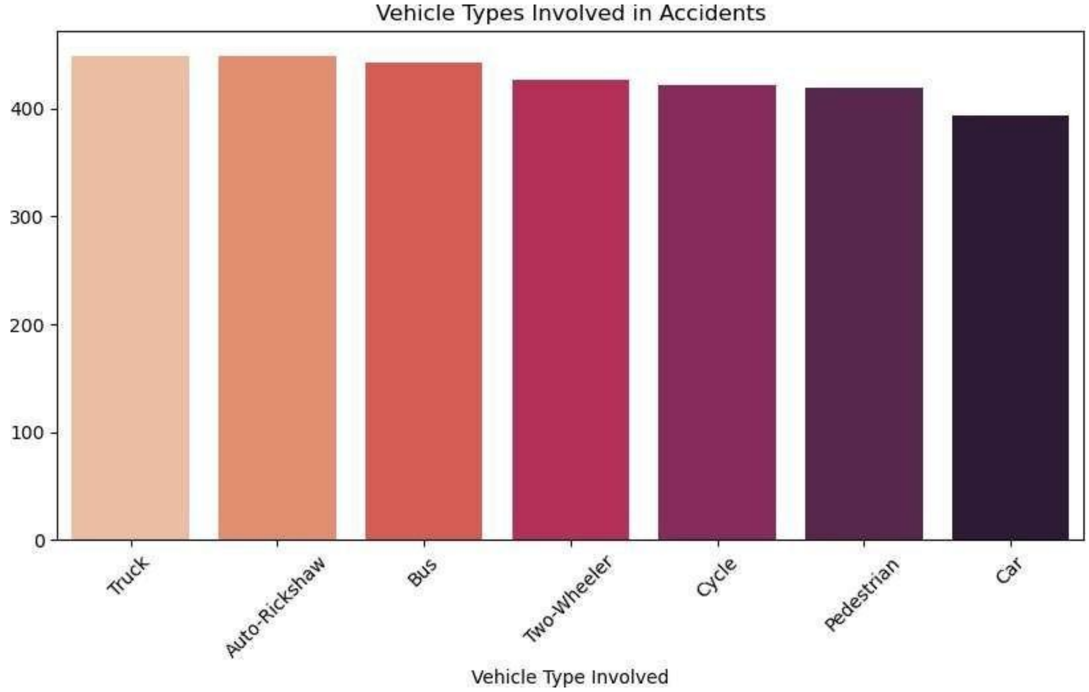


Fig3 - Vehicle Types Involved in Accidents

Interpretation:

- The frequent vehicle types involved in accidents are Trucks and Auto-Rickshaws.
- In addition, Buses have the accident count that is almost equal to their accidents, which means that accidents with heavy and public transport vehicles make up a large part of the total number of accidents.
- The figure of accidents which involve Two-Wheelers, Cycles, and even Pedestrians is very high, hence, it can be inferred that the vulnerable road users are still at a great risk.

Conclusion:

- The chart represents road accidents among road users who cover a wide spectrum of the composition of the road users ranging from heavy vehicles to non-motorized and pedestrian participants.
- The frequency of accidents involving commercial and public transport vehicles is high. This may indicate that there are some problems with vehicle maintenance, driver training, or traffic congestion in urban areas.

4.5.4 States/UT with Most Accidents

The bar graph shows the top 10 Indian states and union territories with the highest number of traffic accidents in the dataset.

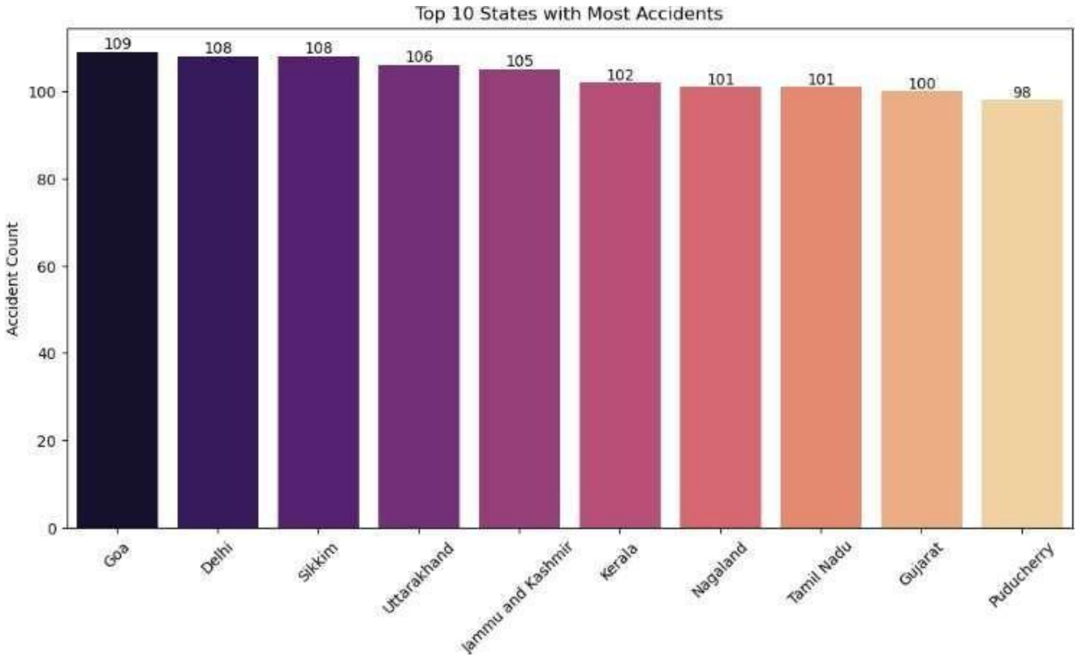


Fig4 - States/UT with Most Accidents

Interpretation:

- At 109, Goa has the most accidents, followed by Delhi and Sikkim, both of which have 108.
- Other areas with high accident rates—between 101and 106 cases—include Uttarakhand, Jammu and Kashmir, Kerala, and Nagaland.
- Indicating that both metropolitan areas and smaller states/UTs are greatly impacted, Tamil Nadu, Gujarat, and Puducherry are also included in the top 10.

Conclusion:

- According to these findings, accidents are common in smaller states and hilly areas like Sikkim and Uttarakhand, rather than just in densely populated or highly urbanized areas. Possible contributing factors involve tourist traffic, difficult surfaces, traffic jams in cities, and the quality of the infrastructure.
- Factors involve tourist traffic, difficult surfaces ,traffic jams in cities and the quality of infrastructure.

4.5.5 Accidents by Road Condition

Below horizontal bar chart shows the distribution of traffic accidents according to different road conditions.

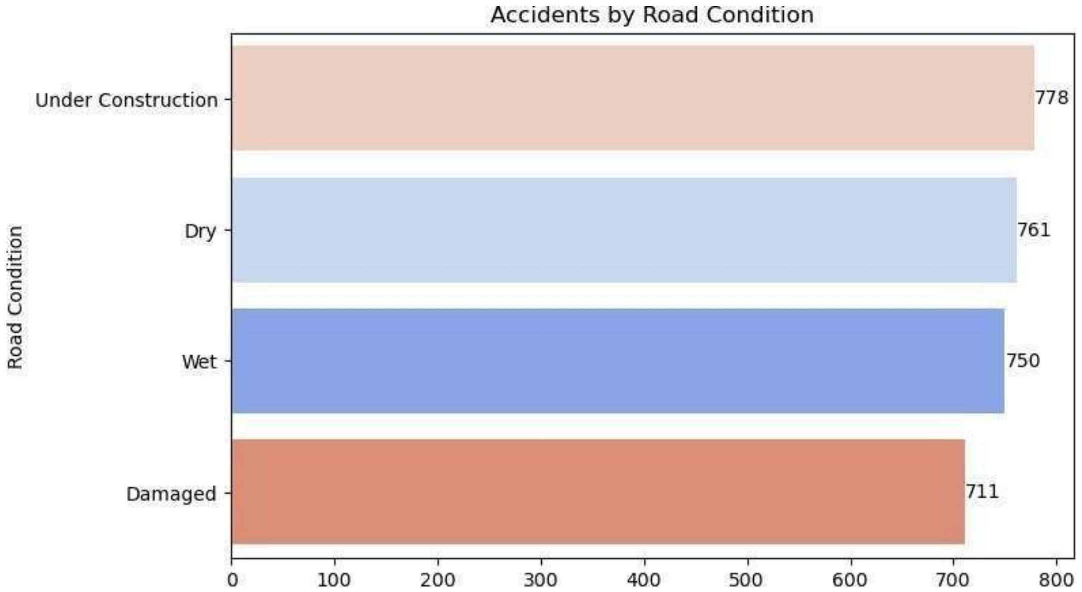


Fig5 - Accidents by Road Condition

Interpretation:

- Under Construction roads account for the highest number of accidents (778), indicating a strong link between ongoing roadwork and increased accident risk. This may be due to:
  - Lack of proper signage or lighting
  - Poor traffic management in construction zones
  - Obstructions and irregular surfaces
- Unexpectedly, dry roads rank second (761), most likely due to their prevalence rather than their dangers. Under ideal conditions, drivers might also become overconfident, which could lead to reckless driving or speeding.
- Because of the 750 accidents that happen on wet roads,slippery roads shows a significant impact on road safety.
- 711 accidents are caused by damaged roads, underscoring the risks posed by potholes, cracks, and uneven surfaces.

Conclusion:

- Accident rates are directly impacted by road conditions.
- Reducing accident rates requires driver awareness, clear construction zone markings, and infrastructure maintenance.
- Priorities for authorities should be:
  - Quick fixes for damaged roads
  - Improved safety procedures on building sites

4.6 BIVARIATE ANALYSIS

Visualizing Relationships Between Categorical and Target Variables To explore the relationship between each categorical feature and the target variable, we used count plots (grouped bar charts) from the seaborn library. These plots help in identifying how different categories of a feature relate to the outcome variable.

4.6.1 Accident Severity by Driver Gender

The distribution of accidents by driver gender (male and female) under the two severity classes (fatal and non-fatal) is displayed in the bar chart. The number of accidents in each severity level can be directly compared thanks to side-by-side bars for each gender group.

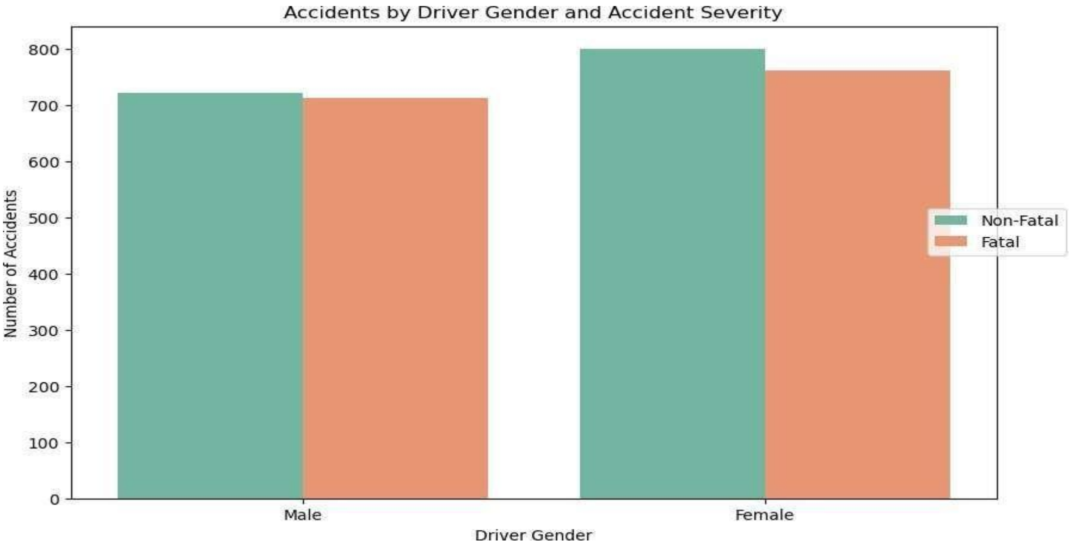


Fig6 - Accident Severity by Driver Gender

Interpretation:

- Most incidents involving male drivers do not end in fatalities, as evidenced by the fact that male drivers are marginally more likely to be involved in non-fatal accidents than fatal ones.
- Similarly, female drivers are more likely to be involved in non-fatal collisions than fatal ones. This dataset, however, indicates that female drivers have a slightly higher chance than male drivers of being involved in both fatal and non-fatal collisions.
- The female group exhibits somewhat higher participation in both accident types, although overall there isn't much of a gender difference.

Conclusion:

- Male and female drivers appear to have similar accident severity profiles, according to the chart, with female drivers having somewhat more accidents overall.

- Numerous factors, including exposure to driving, traffic density in regions with a higher proportion of female drivers, or possible sampling biases in the dataset, could have an impact on these findings.
- Policy Implication: Data-driven, gender-aware initiatives, such as awareness campaigns, training courses, or insurance risk assessments customized for various driver demographics, may be informed by these gender-based trends.

4.6.2 Fatality Distribution Across Vehicle Types in Road Accidents

In accordance with the type of automobile involved, the bar chart compared the frequency of traffic accidents, which have been classified as either fatal or non-fatal. Two distinct severity levels across vehicle types can be visually compared given that each vehicle category is represented through two bars.

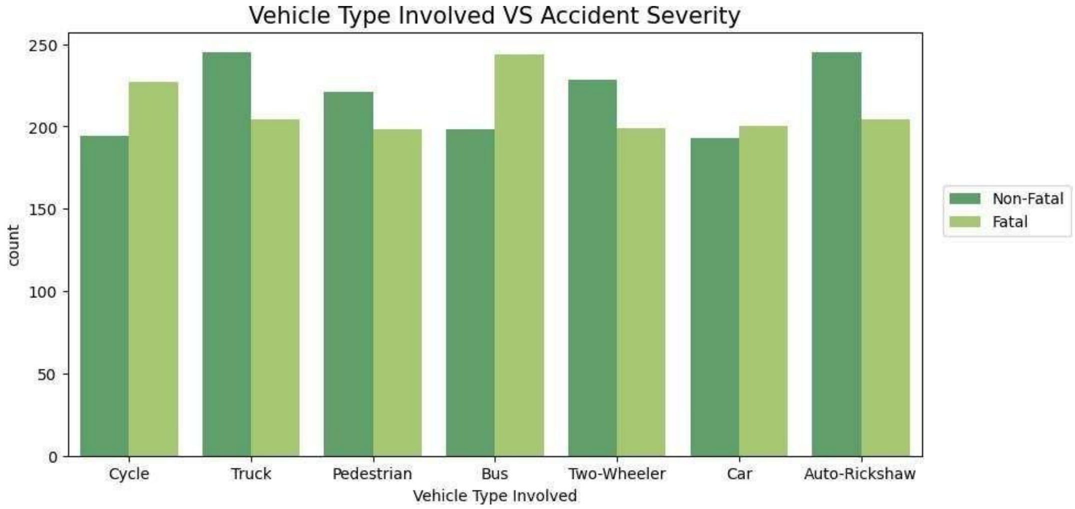


Fig7 - Fatality Distribution Across Vehicle Types in Road Accidents

Interpretation:

- The vehicles with highest accident rates—both fatal and non-fatal—are trucks, two-wheelers, and rickshaws and. Notably, a little higher number of non-fatal cases are reported by trucks and rickshaws.
- Buses have a disproportionately high number of fatal accidents compared to non-fatal ones, suggesting that accidents involving buses usually carry more serious consequences, perhaps as a result of their size, speed, and or urban working patterns.
- Although there is a noticeable fatality rate to earn cyclists and pedestrians, indicating their vulnerability, cars, bicycles, and pedestrians show an additional balanced breakdown of fatal and non-fatal accidents.

Conclusions:

- The vehicle type has an important impact on the severity of an accident, as the graph illustrates. Larger or more visible modes of transport carry a greater chance of fatal accidents.
- The details provided are essential because:
  - o Planning for urban transportation (including bus lanes and pedestrian zones),
  - o Campaigns for vehicle-specific safety .
  - o Policy changes aimed at reducing the number of fatalities among vulnerable road user groups

5. DETERMINING THE RESPONSE VARIABLE

Accident Severity is the response variable and depending on variables like the quantity of accidents, fatalities, and property damage, it can be classified into levels like Non-Fatal or Fatal.Driver age, driver gender, vehicle type, road conditions, weather, accident time, day of the week, accident location, speed limit, etc. are some of the predictor variables.

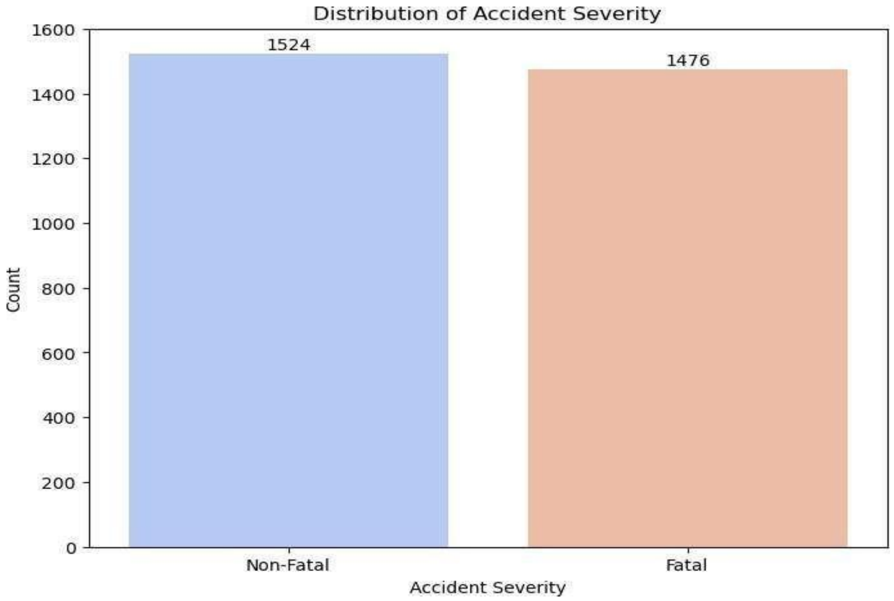


Fig8 – Distribution of Accident Severity

6. STUDYING FACTORS OR CORRELATION

To understand the relationships between different numerical variables in the dataset and the target variable (Accident Severity), a correlation matrix was generated. The matrix displays how strongly each variable is linearly related to others, with values ranging from -1 to +1 . It was observed that:

- Most variables showed very weak or no correlation with Accident Severity.

- Number of Fatalities and Number of Vehicles Involved showed a very slight positive correlation (≈ 0.04), suggesting that higher fatalities may occur when more vehicles are involved.
- Speed Limit (km/h) and Driver Age had almost no significant correlation with any other numeric variables.
- Since the correlation between the number of vehicles involved and the number of fatalities was moderate, it was the strongest of any pair of variables (correlation = 0.04), showing that accidents that involve multiple vehicles usually end in greater deaths.

In the end, the correlation matrix indicates that unlike a single numerical variable having a dominating effect, Accident Severity may be affected by a combination of factors. This indicates that so as to better understand and predict the severity of traffic accidents, advanced analytical techniques like logistic regression are needed.

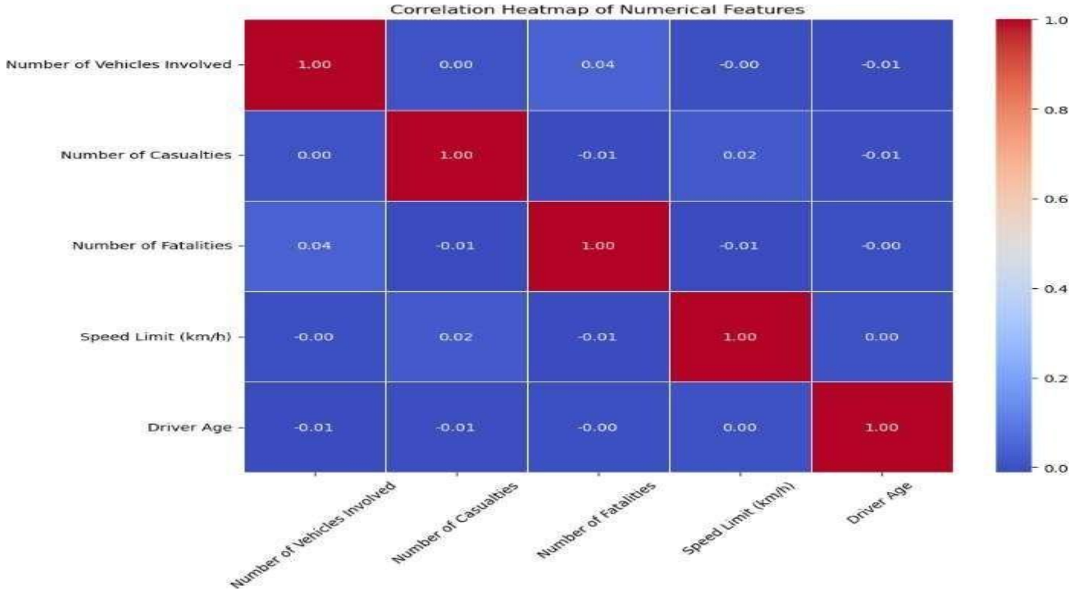


Fig9 – Correlation matrix

7. IDENTIFYING IMPORTANT FEATURES

The top eight features determined by a Random Forest model to have the greatest impact on accident severity classification are shown in the bar chart above. Each feature's role in making accurate predictions is calculated by the importance score.

- **Speed Limit (km/h)** : The most important factor looked out to be the speed limit (km/h). The results of accidents are seriously affected by speed limits, with higher speeds typically raising the risk of fatal accidents.
- **Driver Age** : Since age is linked with experience or reflexes, driver age is essential. Teenagers or old drivers can have a higher probability of getting in an accident.
- **Number of Casualties**: The amount of injured is correlated to the extent of the accident, as more injured often implies a more severe occurrence.

- **Year** is a temporal indicator and may capture changes in road safety laws, vehicle technology, or driving behavior over time.
- **Number of Fatalities** is inherently linked to accident severity and was rightly identified as an important predictor.
- **Number of Vehicles** involved raises the risk that multi-vehicle collisions are more complicated and possibly more serious than single-vehicle ones.
- **Driver Gender** was found to be moderately significant, suggesting that there might be differences in behavior/ exposure to driving between gender.
- **Alcohol Involvement** While not nearly as significant as the other factors, alcohol involvement (yes) is just as important in that it can triple the likelihood of a fatal or serious outcome in an altercation.

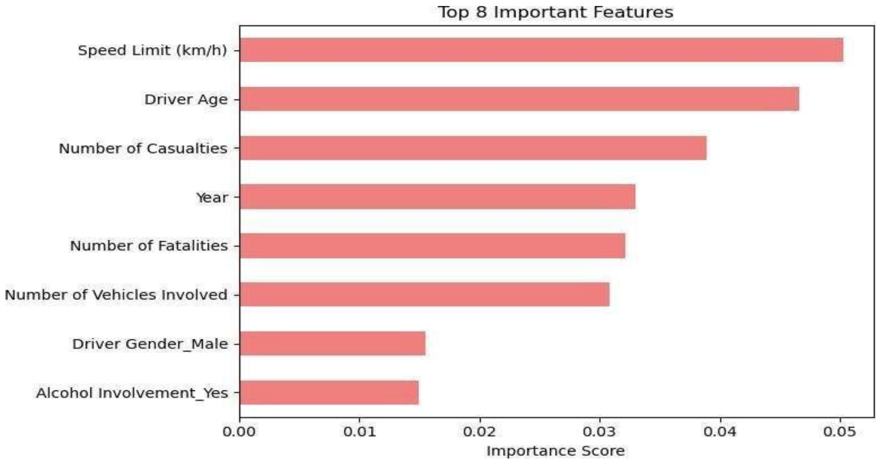


Fig10 – Important Features

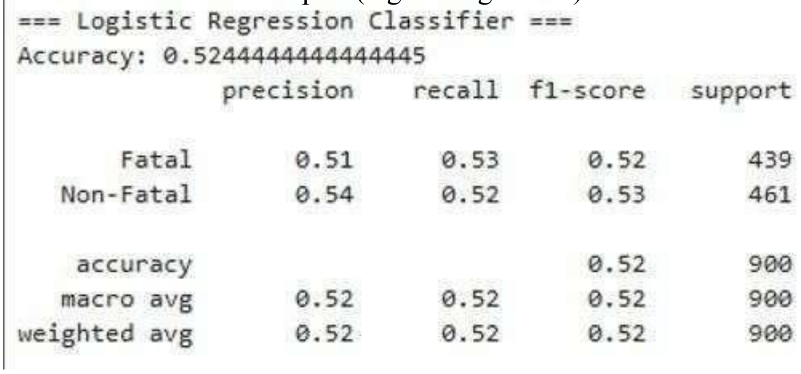
8. FITTING STATISTICAL MODELS FOR THE ANALYSIS

8.1 Logistic Regression

- Using the Logistic Regression Classifier as the classification model, the following results were obtained while predicting the severity of traffic accidents categorized into two classes: fatal and non-fatal.
- Slightly over half of the estimates were precise according to the model's overall accuracy of 52.44%. Based on the information given and features, this indicates the model's ability to distinguish between fatal and non-fatal accidents is only moderately effective.

Classification report was as follows

Table 4- Classification Report (logistic regression)



```
=== Logistic Regression Classifier ===
Accuracy: 0.5244444444444445
              precision    recall  f1-score   support

       Fatal       0.51      0.53      0.52       439
   Non-Fatal       0.54      0.52      0.53       461

    accuracy                           0.52       900
   macro avg       0.52      0.52      0.52       900
weighted avg       0.52      0.52      0.52       900
```

Precision, F1-score and recall for the Fatal class were 0.51,0.52 and 0.53 and similarly,Non-Fatal class got an F1-score, precision and recall of 0.53,0.54 and 0.52. These results show that the model fails to show strong predictive power for either class, instead it performs fairly evenly across both. The macro and weighted averages of accuracy, recall, and F1-score are all 0.52; this shows that the classifier treats both classes similarly. To increase predictive accuracy, these metrics' moderate values, however, indicate to the necessary of better feature selection, model tuning, or the use of improved algorithms.

Confusion Matrix Analysis:

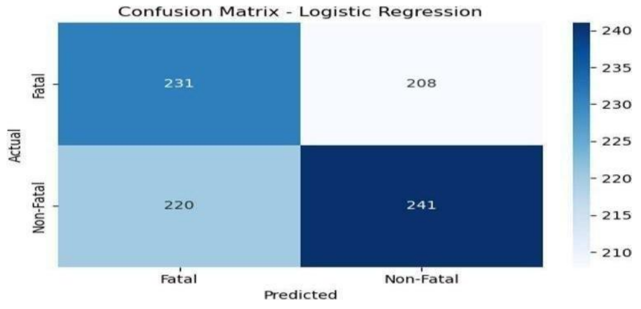


Fig12 – Confusion Matrix (logistic Regression)

The classifier's ability to identify between the two classes—fatal and non-fatal is further showed by the Logistic Regression model's confusion matrix.

Actual \ Predicted	Fatal	Non-Fatal
Fatal	231	208
Non-Fatal	220	241

This matrix reveals that:

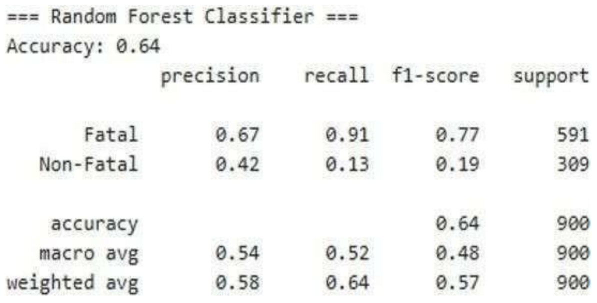
- True Positives (231): These are the instances in which the model accurately predicted a fatal accident and the actual accident was fatal.
- True Negatives (241): These are the Non-Fatal accidents correctly predicted as Non-Fatal.
- False Positives (220): These are Non-Fatal accidents that were incorrectly predicted as Fatal.
- False Negatives (208): These are Fatal accidents that the model mistakenly predicted as Non-Fatal.

This matrix reveals that:

- For both classes, the model's accuracy is identical, with the correct predictions (231 and 241) little better than the wrong ones (208 and 220).
- The nearly equal number of mistakes suggests that the model is not very much in support of either class.
- 472 of the 900 total predictions were right, and 428 were wrong, which is consistent with the previously reported 52.44% overall accuracy.

8.2 Random Forest Classifier

The Random Forest Classifier was applied to predict the severity of road accidents, categorized into Fatal and Non-Fatal. The following results were obtained:



```
=== Random Forest Classifier ===
Accuracy: 0.64
              precision    recall  f1-score   support

       Fatal       0.67      0.91      0.77       591
   Non-Fatal       0.42      0.13      0.19       309

    accuracy                           0.64       900
   macro avg       0.54      0.52      0.48       900
weighted avg       0.58      0.64      0.57       900
```

Table 5 – Classification Report for Random Forest

Approximately 64% of the predictions were accurate, according to the model's 64% overall accuracy. Overall performance is better because this accuracy is higher that that of the logistics regression model (52.44%)

Class (Accident Severity)	Precision	Recall	F1-Score	Support
Fatal	0.67	0.91	0.77	591
Non-Fatal	0.42	0.13	0.19	309
Overall / Avg	0.54	0.52	0.48	900

Key Metrics:

Precision:

- In terms of Fatal: 67% of the anticipated fatal incidents were accurately classified as such.
- Only 42 percent of the non-fatal accidents that were anticipated turned out to be non-fatal.

Recall:

- In terms of Fatal, 91% of real fatal accidents were accurately classified as such, which is very good.
- Regarding Non-Fatal: This class performed poorly, as only 13% of real non-fatal accidents were accurately classified as such.

F1-Score:

- For Fatal: The classifier works well in predicting fatal accidents, as indicated by the F1-score of 0.77.
- To non-fatal accidents, the model's low F1-score of 0.19 indicates how badly it can forecast non-fatal accidents.

Conclusion:

- According to the very low recall and F1-score for this class, the Random Forest Classifier has a very difficult time predicting non-fatal accidents, even though its strong ability to predict fatal accidents with high precision and recall (91% recall for fatal).

- The observed inequality in performance means that the model may require class balancing (e.g., over sampling for non-fatal accidents) or additional adjusting to enhance predictions for non-fatal accidents.
- While the non-fatal predictions performed poorly, the overall precision of 64% is a significant improvement over the 52.44% of the Logistic Regression model.

Confusion Matrix Analysis:

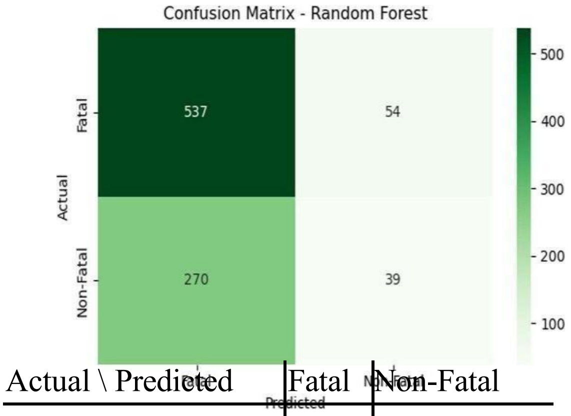


Fig14 – Confusion Matrix (Random Forest)

The confusion matrix for the Random Forest Classifier offers a clear picture of the model's prediction results for the two accident severity classes: Fatal and Non-Fatal.

Actual \ Predicted	Fatal	Non-Fatal
Fatal	537	54
Non-Fatal	270	39

Interpretation:

- True Positives (537): Fatal incidents that were accurately classified as such.
- True Negatives (39): Non-fatal mishaps were accurately classified as such.
- False Positives (270): Accidents that were not fatal but were mistakenly classified as such.
- False Negatives (54): Accidents that are fatal but are mistakenly classified as non-fatal.

Observations:

- The model has a high recall (91%) for the Fatal class and does a very good job of identifying Fatal cases, correctly predicting 537 out of 591 Fatal cases.
- The model only correctly predicts 39 out of 309 Non-Fatal cases, which results in a very low recall (13%) for the Non-Fatal class.
- A significant bias towards the Fatal class is suggested by the high number of false positives (270), which shows that many non-fatal accidents are being incorrectly classified as Fatal.
- This disparity is consistent with the report on classification, where the the F1-score for Fatal is 0.77 while for Non-Fatal it drops drastically to 0.19.

9. MODEL PERFORMANCE EVALUATION

In this analysis, the Random Forest Classifier is the model that performs the best at predicting the severity of traffic accidents, according to the performance evaluation of the two models. At 64%, it surpassed the Logistic Regression model in terms of overall accuracy, which was 52.44%. In particular, Random Forest showed a high recall of 91% for fatal cases, showing its strong ability to accurately identify serious accidents. It is especially important in practical applications where correct fatal accident identification is necessary for emergency response planning and public safety.

It's important to understand that Random Forest's recall of only 13% shows that it performed not well in predicting non-fatal cases. Despite this flaw, it has an advantage due to its high predictive ability for fatal accidents, especially in situations where identifying serious incidents is more crucial than attaining perfect balance across all classes.

The Logistic Regression model, on the other the same direction, given better performance in both classes, but it was less accurate overall and performed greater in each class. Because of its consistently moderate metrics, it is better suited for cases in which both Fatal and Non-Fatal classes are given equal weight, but it is less useful for issues requiring high precision or sensitivity to Fatal outcomes.

In summary, the random forest is the best model in this situation because random forest classifier is more accurate at identifying fatal accidents, but there some additional work which is required to make it better.

10. CONCLUSION

We used real-world data, and this statistical analysis project sought to determine the severity of traffic accidents and use predictive modeling techniques to categorize them as fatal or non-fatal. In order to forecast accident outcomes based on multiple contributing factors, the analysis first involved a thorough exploratory data analysis, after which Logistic Regression and Random Forest Classifier models were implemented.

Important insights into the dataset's distributions and patterns were revealed by preliminary statistical summaries and visualizations. Key factors affecting accident severity were identified with the use of boxplots, correlation heatmaps, and summary statistics. To get the dataset ready for modeling, feature selection and preprocessing were done with care.

With an overall accuracy of 52.44% and nearly equal precision and recall for both the Fatal and Non-Fatal classes, the Logistic Regression model showed a balanced but modest performance. While it had little predictive power, it behaved properly by treating both classes equally.

However, with a recall of 91% and an F1-score of 0.77, the Random Forest model displayed strong performance in predicting fatal accidents and achieved a higher accuracy of 64%. Its considerably poorer performance for non-fatal predictions, however, indicates a model bias to the benefit of the Fatal class.
The Random Forest Classifier is more successful for uses where finding fatal accidents is a top priority, even though both models have various advantages, according to this analysis. However, class imbalance, model parameter efficiency, and the study of other algorithms or combination methods should be the main focus of future work in order to create a more accurate and balanced prediction system.

Finally, this project shows how statistical and machine learning methods can be used to extract useful information from accident data. This information can then be used by emergency services, traffic authorities, and policymakers improve safety measures and help to reduce the severity of accidents.

11. REFERENCES

- Kaggle Dataset: Source of raw data used for analysis and modeling. https://www.kaggle.com
- Scikit-learn: Scikit-learn: Machine Learning in Python, Pedregosa et al., Journal of Machine Learning Research, 2011. https://scikit-learn.org/stable/
- Pandas: pandas: Powerful Python data analysis toolkit, The Pandas Development Team. https://pandas.pydata.org/
- Matplotlib: Matplotlib: Visualization with Python, Hunter, J. D., Computing in Science & Engineering, 2007. https://matplotlib.org/ 🎬 Seaborn: Seaborn: Statistical Data Visualization, Waskom, M., Journal of Open Source Software, 2021. https://seaborn.pydata.org/
- Ministry of Road Transport and Highways (MoRTH), Government of India — Road Accident Statistics — Reports and statistics published annually on road accidents in India were referred to for understanding real- world accident severity distributions and trends. https://morth.nic.in/road-accident-in-india
- National Crime Records Bureau (NCRB) — Accidental Deaths & Suicides in India Reports — Used to understand accident causes, driver behavior, and victim statistics in India.

Shivam Gupta

A Statistical Study of Road Accidents in Various States in India.docx

Amity University, Noida

Document Details

Submission ID

trn:oid:::16158:104682870

Submission Date

Jul 16, 2025, 3:06 PM GMT+5:30

Download Date

Jul 16, 2025, 3:08 PM GMT+5:30

File Name

A Statistical Study of Road Accidents in Various States in India.docx

File Size

699.7 KB

24 Pages

4,587 Words

25,868 Characters

0%Overall Similarity

Thecombinedtotalofallmatches,includingoverlappingsources, for each database.

Filtered from the Report

- Bibliography
- Cited Text
- Small Matches (less than 14 words)
- Submitted works

Match Groups

0 Not Cited or Quoted0%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations0%
Matches that are still very similar to source material

0 Missing Citation0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted0%
Matches with in-text citation present, but no quotation marks

Top Sources

0%	Internet sources
0%	Publications
0%	Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags forReview

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.
A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

0 Not Cited or Quoted0%
Matches with neither in-text citation nor quotation marks

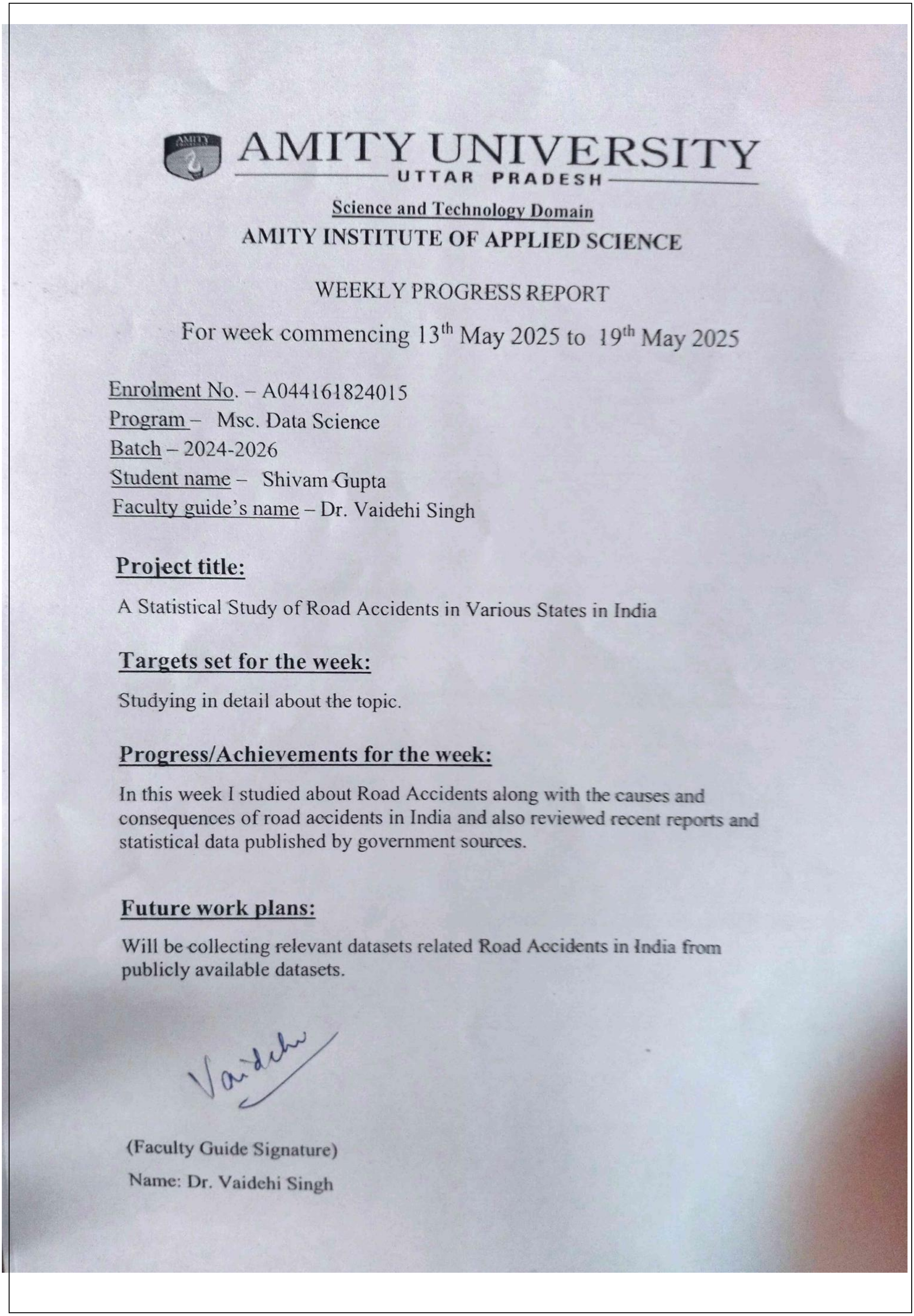
0 Missing Quotations0%
Matches that are still very similar to source material

0 Missing Citation0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted0%
Matches with in-text citation present, but no quotation marks

Top Sources

0%	Internet sources
0%	Publications
0%	Submitted works (Student Papers)



AMITY UNIVERSITY
UTTAR PRADESH

Science and Technology Domain

AMITY INSTITUTE OF APPLIED SCIENCE

WEEKLY PROGRESS REPORT

For week commencing 13th May 2025 to 19th May 2025

Enrolment No. – A044161824015
Program – Msc. Data Science
Batch – 2024-2026
Student name – Shivam Gupta
Faculty guide's name – Dr. Vaidehi Singh

Project title:

A Statistical Study of Road Accidents in Various States in India

Targets set for the week:

Studying in detail about the topic.

Progress/Achievements for the week:

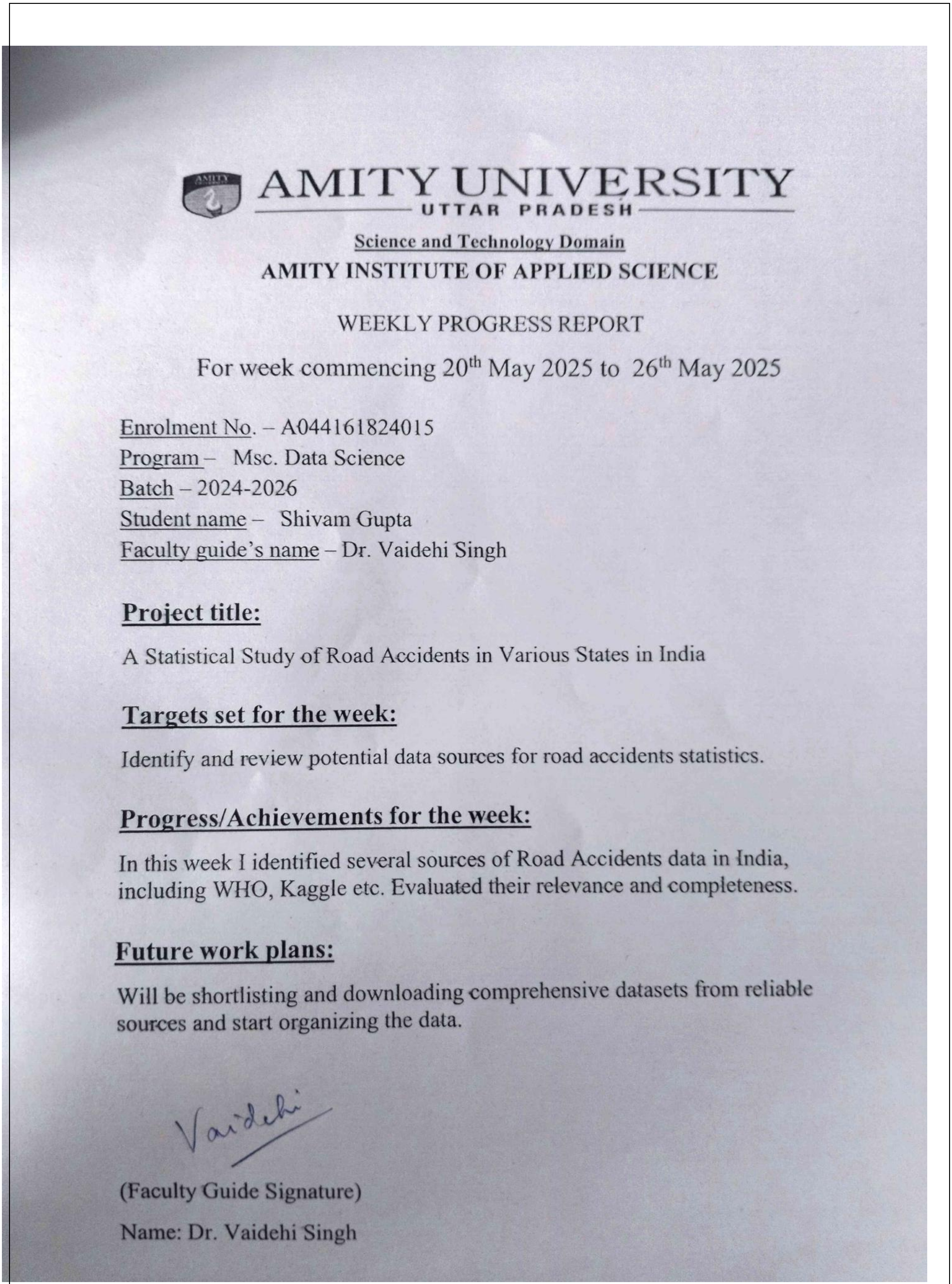
In this week I studied about Road Accidents along with the causes and consequences of road accidents in India and also reviewed recent reports and statistical data published by government sources.

Future work plans:

Will be collecting relevant datasets related Road Accidents in India from publicly available datasets.

Vaidehi

(Faculty Guide Signature)
Name: Dr. Vaidehi Singh



AMITY UNIVERSITY
UTTAR PRADESH

Science and Technology Domain

AMITY INSTITUTE OF APPLIED SCIENCE

WEEKLY PROGRESS REPORT

For week commencing 20th May 2025 to 26th May 2025

Enrolment No. – A044161824015
Program – Msc. Data Science
Batch – 2024-2026
Student name – Shivam Gupta
Faculty guide's name – Dr. Vaidehi Singh

Project title:

A Statistical Study of Road Accidents in Various States in India

Targets set for the week:

Identify and review potential data sources for road accidents statistics.

Progress/Achievements for the week:

In this week I identified several sources of Road Accidents data in India, including WHO, Kaggle etc. Evaluated their relevance and completeness.

Future work plans:

Will be shortlisting and downloading comprehensive datasets from reliable sources and start organizing the data.

Vaidehi

(Faculty Guide Signature)

Name: Dr. Vaidehi Singh

AMITY UNIVERSITY
UTTAR PRADESH

Science and Technology Domain

AMITY INSTITUTE OF APPLIED SCIENCE

WEEKLY PROGRESS REPORT – 3

For week commencing 27th May 2025 to 2nd June 2025

Enrolment No. – A044161824015
Program – Msc. Data Science
Batch – 2024-2026
Student name – Shivam Gupta
Faculty guide's name – Dr. Vaidehi Singh

Project title:

A Statistical Study of Road Accidents in Various States in India

Targets set for the week:

Download and Organise datasets along with it begin cleaning and preparing selected dataset for the analysis.

Progress/Achievements for the week:

In this week I downloaded multiple datasets related to road accidents categorized by state, year, cause, and type. Organised data in excel for preliminary examination along with it created a clean version for analysis.

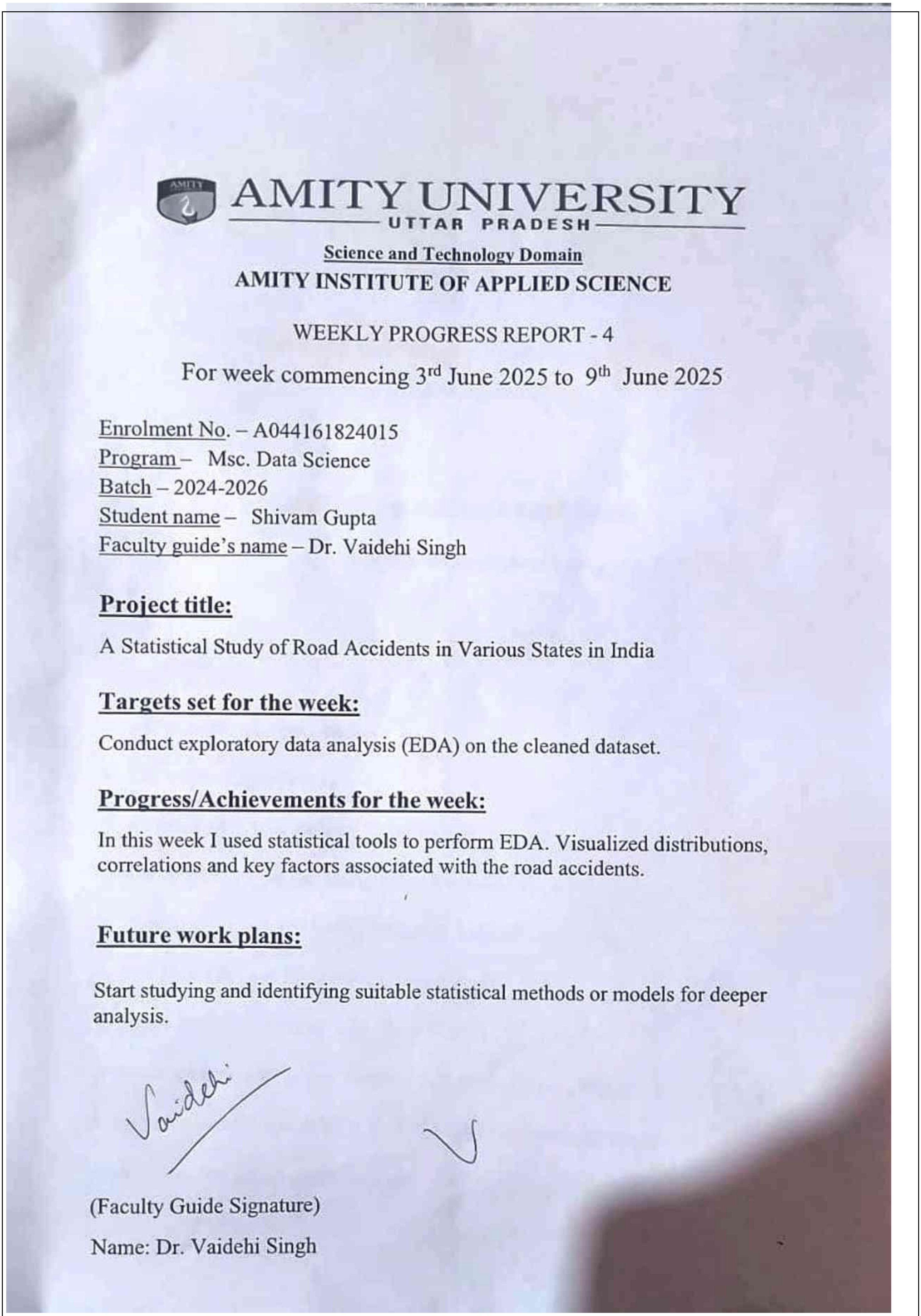
Future work plans:

Perform exploratory data analysis (EDA) to understand patterns and relationships in the data.

Vaidehi

(Faculty Guide Signature)

Name: Dr. Vaidehi Singh



AMITY UNIVERSITY
UTTAR PRADESH

Science and Technology Domain

AMITY INSTITUTE OF APPLIED SCIENCE

WEEKLY PROGRESS REPORT - 4

For week commencing 3rd June 2025 to 9th June 2025

Enrolment No. – A044161824015
Program – Msc. Data Science
Batch – 2024-2026
Student name – Shivam Gupta
Faculty guide's name – Dr. Vaidehi Singh

Project title:

A Statistical Study of Road Accidents in Various States in India

Targets set for the week:

Conduct exploratory data analysis (EDA) on the cleaned dataset.

Progress/Achievements for the week:

In this week I used statistical tools to perform EDA. Visualized distributions, correlations and key factors associated with the road accidents.

Future work plans:

Start studying and identifying suitable statistical methods or models for deeper analysis.

Vaidehi

(Faculty Guide Signature)

Name: Dr. Vaidehi Singh

AMITY UNIVERSITY
UTTAR PRADESH

Science and Technology Domain

AMITY INSTITUTE OF APPLIED SCIENCE

WEEKLY PROGRESS REPORT - 5

For week commencing 10th June 2025 to 16th June 2025

Enrolment No. – A044161824015
Program – Msc. Data Science
Batch – 2024-2026
Student name – Shivam Gupta
Faculty guide's name – Dr. Vaidehi Singh

Project title:

A Statistical Study of Road Accidents in Various States in India

Targets set for the week:

Studied about statistical methods and different models.

Progress/Achievements for the week:

In this week I studied about various statistical methods that are appropriate for the analysis that I'm conducting, along with it reviewed regression and classification models (like logistic regression) based on the nature of the data.

Future work plans:

Will be focusing more on the appropriate model selection.

Vaidehi

(Faculty Guide Signature)

Name: Dr. Vaidehi Singh

AMITY UNIVERSITY
UTTAR PRADESH

Science and Technology Domain

AMITY INSTITUTE OF APPLIED SCIENCE

WEEKLY PROGRESS REPORT - 6

For week commencing 17th June 2025 to 23rd June 2025

Enrolment No. – A044161824015
Program – Msc. Data Science
Batch – 2024-2026
Student name – Shivam Gupta
Faculty guide's name – Dr. Vaidehi Singh

Project title:

A Statistical Study of Road Accidents in Various States in India

Targets set for the week:

Choose appropriate statistical techniques for analyzing the dataset.

Progress/Achievements for the week:

In this week I selected Classification models (like KNN and Random Forest) based on the nature of the data. Started implementing these models.

Future work plans:

Evaluating the performance of different models.

for.

(Faculty Guide Signature)

Name: Dr. Vaidehi Singh

AMITY UNIVERSITY
UTTAR PRADESH

Science and Technology Domain

AMITY INSTITUTE OF APPLIED SCIENCE

WEEKLY PROGRESS REPORT - 7

For week commencing 24th June 2025 to 30th June 2025

Enrolment No. – A044161824015
Program – Msc. Data Science
Batch – 2024-2026
Student name – Shivam Gupta
Faculty guide's name – Dr. Vaidehi Singh

Project title:

A Statistical Study of Road Accidents in Various States in India

Targets set for the week:

Evaluate model performances and interpret results along with the comparison of the model performance.

Progress/Achievements for the week:

In this week I tested models and evaluated them using accuracy, precision, and recall. I also compared model performances.

Future work plans:

Draft findings and begin writing the analysis and results of the final report.

for.

(Faculty Guide Signature)

Name: Dr. Vaidehi Singh

AMITY UNIVERSITY
UTTAR PRADESH

Science and Technology Domain

AMITY INSTITUTE OF APPLIED SCIENCE

WEEKLY PROGRESS REPORT - 8

For week commencing 1st July 2025 to 7th July 2025

Enrolment No. – A044161824015
Program – Msc. Data Science
Batch – 2024-2026
Student name – Shivam Gupta
Faculty guide's name – Dr. Vaidehi Singh

Project title:

A Statistical Study of Road Accidents in Various States in India

Targets set for the week:

Finalizing the report work along with giving conclusion.

Progress/Achievements for the week:

In this week I have finalised the report along with the compilation.

Future work plans:

Submitting the final report.

for.

(Faculty Guide Signature)

Name: Dr. Vaidehi Singh