National College of Ireland

Predictive Analysis of Road Accidents in India: A Machine Learning Approach

MSc Research Project
Data Analytics

Harini Manjunatha
Student ID: x22169288

School of Computing
National College of Ireland

Supervisor: Shubham Subhnil

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Harini Manjunatha
Student ID:	x22169288
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Shubham Subhnil
Submission Due Date:	14/12/2023
Project Title:	Predictive Analysis of Road Accidents in India: A Machine Learning Approach
Word Count:	6501
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Harini Manjunatha
Date:	31st January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predictive Analysis of Road Accidents in India: A Machine Learning Approach

Harini Manjunatha
x22169288

Abstract

The road traffic accident in India poses serious threat as it impose huge socio-economic costs on a society. The increased numbers in fatalities and injuries due to road accident has forced government to look for solutions to reduce the accident rate. Predicting road accidents is crucial because it helps avoid fatalities, reduce injuries, and allocate resources effectively. Recognising locations and times when accidents are likely to occur allows authority to respond quickly, allowing for traffic control measures and the development of safer infrastructure designs. This project focuses on the development and application of machine learning algorithms for road accident prediction. Multiple machine learning(ML) models, including Random Forest, linear regression and Decision Trees were used and thoroughly analysed to assess how accurately they predicted the possibility of an accident. The Random Forest Regressor model demonstrated superior performance comparitively by predicting accidents on road based on historical patterns. The analysis of the models predictions showed that accidents peaked frequently in different parts of India between the time period of 15:00 to 21:00. The time period with the fewest accidents occurred during the night, specifically between 3 AM and 6 AM. Madhya Pradesh experiences the highest accident rate among all states in India, while Lakshadweep exhibits the lowest accident rate among them. The model's findings can significantly improve traveller safety and aid authorities in developing plans to reduce and eliminate fatal accidents on Indian roads.

Keywords: Road accident, Machine learning, Fatalities, Time interval.

1 Introduction

1.1 Background

The growth in transportation industry has resulted in rise in the volume of vehicles on road. The growing count of vehicles on road each day corresponds to a rise in the frequency of accidents. (Chirag and Supreetha; 2022). Road accidents has now pose a severe risk to the safety of public, since they cause both material and human loss (Vipin Na; 2021). According to the World Health Organization, traffic accidents rank ninth on the list of factors that lead to fatalities in human beings (Chitradevi and Rajan; 2022). India owns just 1 percent of all vehicles worldwide, however it is responsible for 6 percent of all traffic accidents. (Chitradevi and Rajan; 2022). Indian roads present a difficult transportation environment due to its large road network, varied traffic patterns, unpredictable weather, and complex geography. The number of traffic accidents in the country

has significantly increased, resulting in fatalities, serious injuries, and monetary losses. The increase in the death rate due to accident as made government pressing concern to provide relavant solution (Nawaf and Fred; 2023). This growing trend highlights the urgency of implementing effective strategies to reduce accidents and enhance road safety.

The traditional approaches to accident analysis usually involve time-consuming manual data collecting and investigation (S. Basnyat; 2006) (Jonas Lundberg; 2009). These procedures takes time and has limits in providing quick insights on preventing accidents. Machine learning has become a powerful tool in variety of industries including transportation, by making it possible to extract patterns and insights from numerous datasets (Massimo Bertolini; 2021). Employing machine learning techniques in analyzing road traffic accidents allows us to analyze the pattern in road accidents and provide relevant solutions to prevent accidents (Augustine and Shukla; 2022).

Predicting road accidents is crucial because it helps avoid fatalities, reduce injuries, and allocate resources effectively . By identifying the accident-prone areas and times, concerned authorities can react immediately, better control the traffic, and make plans for safer infrastructure (Vipin Na; 2021). The approach of accident prediction lowers the financial burden, aids in economic planning, and promotes the advancement of technology for improved road safety. Accurate forecasting can help with policy development and the search for ways to stop repeated accidents, which will ultimately lead to a safer transportation system (Parathasarathy et al.; 2019). By identifying high-risk periods, resources can be strategically allocated to increase law enforcement presence, enhance traffic management, and improve infrastructure. It also allows for the optimization of emergency response systems, ensuring that they are adequately prepared to handle any surge in incidents during peak accident-prone times. Understanding accident risks at particular time periods also offers useful insights for educational and public awareness programmes, educating people with knowledge to make safer choices when driving (Vipin Na; 2021)..

1.2 Research Question and Objective

- "How effective are machine learning algorithms in predicting the frequency of traffic accidents in India, and how might these predictive outcomes contribute to controlling and reducing the accident rate?"

Numerous studies have explored what causes accidents on roads (Nedjmedine and Tahar; 2022) (Laura Eboli; 2020) (Kraonual; 2020). It is important to know before hand how many accidents might happen at certain times and places to help us plan and take action to make things safer. For example, In areas where there are many accidents in the late hours of the night, we can put up more lights or use special devices to alert drivers who might be falling asleep. If an area experiences a higher number of accidents during the day, we can install traffic signals or implement other safety measures. By predicting how many accidents might happen soon lets us get ready to help quickly if something goes wrong. This way, we can try to prevent accidents by making roads safer and being ready to help if accidents do happen.

Several studies have considered severity analysis of an accident (Jaber; 2022) (Zhang; 2022) (Wang; 2022) In this project , we focus on predicting the frequency of accident that might happen in the coming year at specific time in each state of India. This work aid government and the concerned authorities to take necessary action to reduce the accident

rate. The dataset is taken from the Open government data website[1] (G and R; 2023) (Viswanath Dhanya K; 2021) (Chirag and Supreetha; 2022). The dataset is compiled by gathering the information collected spanning the years 2001 to 2021. The dataset undergoes various stages of data processing, including pre-processing, data cleansing, and preliminary visualisation. The Machine learning techniques such as Random Forest, Decision Tree, and Linear regressor is used. The modelling performance will be evaluated using evaluation matrices, such as the R-squared value, Mean absolute error, and mean squared error.

1.3 Document Structure

The research document is structured into seven sections, each focusing on different aspects of the investigation. The introduction part of the research is covered in the first section 1. The second section 2 is broken down into three subsections which summarises earlier studies and identifies the unique features of this project. The third section 3 provides the detailed methodology of the project, and the fourth section 4 outlines the design requirements which includes information about the project's critical performance metrics as well as the procedures and methods that were used. Section 5 5 shows how the technical solution was applied in the study while, section 6 6 explores in-depth case studies and explains how the evaluation procedure helped to achieve the research objectives, The study report is finally concluded in Section 7 7, which presents relevant findings and discusses possible directions for further research.

2 Related Work

Numerous research projects from various disciplines have carried out investigations to better understand the elements causing an increasing number of traffic accidents. This growth in research reflects the urgent need to offer practical approaches to lower accident rates, protect lives, and enhance the efficiency of the road transport system.

2.1 Road Accident Prediction using statistical technique

Road crash prediction models are highly valuable tools for highway safety since they can be used in estimating the frequency and severity of accidents. Abdulhafedh (2017) highlights the statistics from world health organisation which emphasizes the severity of road accidents as a leading cause of global fatalities specifically among younger people. It provides insights into diverse approaches used to forecast accidents and address associated risk factors which are crucial for developing effective preventive measures. The various statistical methodologies were built. Poison regression model was identified to be more conventional than other approaches. Francesca La Torre (2019) focuses on improving road safety management. He introduces two accident prediction models based on specific Safety Performance functions aligning with the Highway Safety Manual (HSM) approach and adapting it to European motorways. The outcomes demonstrate both models' strong ability to characterise the analysis data set. The suggested models offer professionals a strong and dependable tool for predicting accidents along the network of Italian highways. These models might not encompass human behavioral factors, such as driver psychology

[1]Open Government data website: `https://community.data.gov.in`

or sudden individual decisions, which significantly impact accident occurrences but are challenging to quantify accurately. Jutaek Oh (2006) examines the factor associated with railroad crossing crashes using various statistical models. The gamma probability model's application to underdispersion and the knowledge gained about automobile crashes related to railway crossings are the paper's two distinctive contributions. These statistical methods heavily rely on historical data, potentially leading to biased predictions due to changing road conditions, traffic patterns, and infrastructure developments that might not be reflected in the dataset.

2.2 Geospatial Analysis and Accident Prediction

Shabani (2014) introduces geospatial approach with fuzzy classification and regression tree (FCART) model to predict and understand motor vehicle crashes. Comparative analysis against other models like CART [2] and SVM [3] reveals the superiority of the bagged-FCART model in crash severity prediction accuracy. geographic factors like curves, nearby facilities, and land use significantly influence crash severity, advocating for geographically targeted safety measures on such roadways. This comprehensive approach highlights the importance of both behavioral and geographical factors in devising effective safety strategies. limitations in data availability or accuracy of geographic information affects the model's ability to generalize beyond the studied areas or influence prediction accuracy. Shad and Moghimi (2013) illustrates the application of Geographic Information System (GIS) to predict accident probabilities at intersections focusing on ensuring safety within the transportation system in Mashhad. After gathering statistical data from traffic observations and urban transportation, the data preparation procedures are carried out in compliance with standards and requirements by applying mathematical and statistical operators including density estimation and interpolation techniques. The model validity level is established by comparing the generated findings with the frequency of accidents noted in the control points. We propose that integrating GIS (Geographic Information System) with machine learning holds the potential to enhance the predictive accuracy of the proposed system in road accident prediction. This integration presents an opportunity for future exploration and improvement in predicting road accidents.

2.3 Road Accident prediction using Machine Learning

(Bi and Yu; 2020) presented a novel heuristic prediction technique that guarantees the accuracy of road traffic accident prediction, prevents local optimisation, and enhances the generalisation capacity of small sample data prediction. The result of the frame work could be effectively applied to the direction of information control in the transportation field. Zhankaziev et al. (2022) proposed the model that predicted the occurrence of accidents on public roads which happened from collisions of two or more vehicles in real time. A machine learning model using random forest and decision tree regressor was developed by G and R (2023) to predict collisions based on collision records that have taken place in the different states of India.The hit and run, head-on collision, hit pedestrian, fog, cloudy, rainy, single lane, two-lane, four-lane, school, pedestrian crossing, market, and other parameters considered for analyzing and visualization of accidents in different states of India.

[2]CART: `https://towardsdatascience.com/cart-classification-and-regression-trees-for-clean-but-powe`
[3]SVM: `https://scikit-learn.org/stable/modules/svm.html`

Paul et al. (2020) created a multiclass model that integrated accident prediction with the severity of the relevant accidents to provide a more effective model for preventing traffic accidents. Banerjee et al. (2022) offered a comparison analysis of several of these models in an attempt to evaluate and determine a useful method for predicting the probability of traffic accidents. Since drivers are the ones in charge when driving, the study attempts to give them a traffic accident risk prediction by analysing the variables they would be aware of in advance, such as vehicle type, age, sex, time of day, and weather. Viswanath Dhanya K (2021) examined the connections between road conditions, the likelihood of accidents, and the influence of environmental variables. The Apriori algorithm and Support Vector Machines were used by the author to create an accident prediction model through the use of data mining techniques. This analysis makes use of Bangalore road accident datasets from 2014 to 2017 that are accessible online.

Augustine and Shukla (2022) suggests a technique for accident prediction that can be used to assess possible safety risks and determine whether an accident will happen or not. To find out which machine learning algorithm can assist anticipate accidents more precisely, a comparison study of several models was carried out. At eighty-seven percent, the Random Forest algorithm produced the best accuracy.t Mallahi et al. (2022) presented severity prediction model for traffic accidents, which is a huge step in road accident management in the road. This problem provides important information for emergency logistical transportation in many cities. Nedjmedine and Tahar (2022) utilised machine learning principles to estimate the event's severity and examine variables such as the annual and state-specific accident rates Road accidents by state, day and hour, and the proportion of accidents in rural versus urban areas Who was involved in the accidents, when was it most unsafe to drive? Chirag and Supreetha (2022) employed machine learning algorithms with clustering and regression techniques to forecast and examine the accident rate for all states and UT in 2022. estimating the accident rate for 2022 by applying the Linear Regression technique. utilising a variety of outside variables, including weather, alcohol, location, intersections, and vehicle defects, as the primary determinants for classification and prediction.

2.4 Takeaway from literature review

Early crash analysis models were generally based on simple multiple linear regression methods assuming normally distributed errors. statistical models struggle to accommodate complex and nonlinear relationships between numerous variables affecting accidents, leading to oversimplified representations and less accurate predictions Abdulhafedh (2017). The inability to account for unforeseen or rare events, especially outliers or extreme weather conditions, hampers the models' predictive accuracy Francesca La Torre (2019) Jutaek Oh (2006). Several studies have focused on predicting how often accidents might happen and how severe they could be Mallahi et al. (2022) Banerjee et al. (2022) . They've looked at various factors like driver details, vehicle information, weather conditions, location, road type, and the type of accident Nedjmedine and Tahar (2022). there is a noticeable gap in the specific context of predicting accident counts at distinct time intervals in each state of the country. The current literature lacks a focused examination of how temporal variations within a day impact the likelihood of accidents. Anticipating the potential count of accidents at a specific time and location for the upcoming year is of paramount importance for proactive risk management. This foresight empowers authorities to implement targeted preventive measures, enhancing overall road safety. It

allows for the strategic allocation of resources, ensuring that emergency services are well-prepared during periods identified as high-risk. Moreover, such advanced knowledge aids in crafting and adjusting policies, traffic regulations, and public awareness campaigns to address the specific challenges associated with those anticipated time frames. This proactive approach not only mitigates the severity of accidents but also contributes to the development of efficient urban planning, infrastructure maintenance, and insurance strategies. By staying ahead of potential risks through predictive analysis, communities can foster a safer and more resilient environment for all road users. The project aims in identifying the time interval and the location at which there is highest probable frequency of road accidents. and develop a model which can give the better result for this prediction.

3 Methodology

This section provides an detailed explanation of the methodology employed in the study. Figure 1 shows the flowchart representation of the methodology.

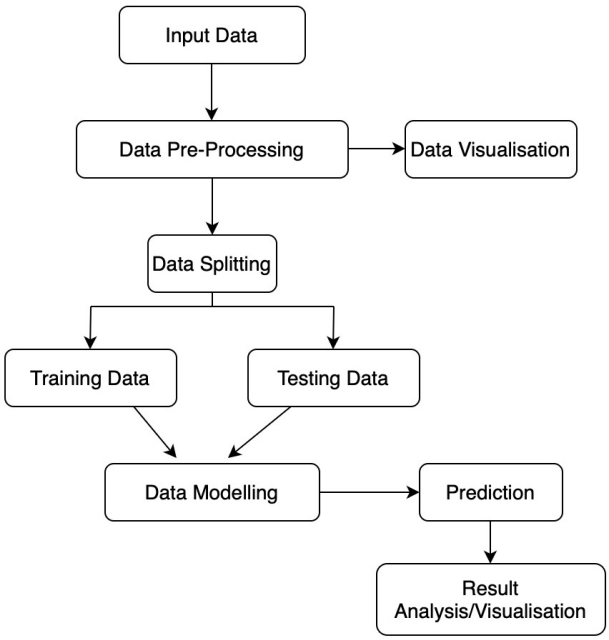


Figure 1: Flowchart Representation of Methodology

3.1 Dataset

The dataset used in this project is taken from the Open Government Data (OGD) Platform, India [4] (Chitradevi and Rajan; 2022) (G and R; 2023) (Viswanath Dhanya K; 2021) (Chirag and Supreetha; 2022). The dataset captures the frequency of traffic accidents that occur within particular time intervals of the day in numerous Indian states over a period of 2001 to 2021. The time intervals are categorized in segments of three

[4]Open Government data website: `https://community.data.gov.in`

hours each 6-9 AM, 9-12 PM, 12-3 PM, 3-6 PM, 6-9 PM, 9-12 AM (midnight), 12-3 AM, and 3-6 AM. There is noticeably absence of data for the year 2015 in the dataset. To address this absence, the missing values are replaced with mean of the observed data. This dataset can be used to provide a thorough analysis of the time distribution of road accidents in different parts of India. Figure 2 is the dataset that we have considered for our study. The dataset consist of 708 rows and 11 columns of count of accidents in each state of india from the year 2001 to 2021 for each time period.

	State	Year	6-9	9-12	12-15	15-18	18-21	21-24	0-3	3-6
0	ANDHRA PRADESH	2001	2239.0	3265.0	3198.0	3729.0	3604.0	3792.0	4098.0	3263.0
1	ANDHRA PRADESH	2002	2931.0	3857.0	3671.0	4255.0	4153.0	4778.0	4844.0	4088.0
2	ANDHRA PRADESH	2003	3158.0	4865.0	3749.0	4319.0	4266.0	4853.0	5218.0	4109.0
3	ANDHRA PRADESH	2004	3191.0	4770.0	4598.0	5030.0	4033.0	4971.0	6031.0	4454.0
4	ANDHRA PRADESH	2005	3826.0	6011.0	5002.0	4137.0	4261.0	4524.0	5096.0	4432.0
...	...	...	...	...	...	...	...	...	...	...
703	Delhi	2021	573.0	578.0	467.0	499.0	733.0	717.0	574.0	473.0
704	jammu & kashmir	2021	512.0	972.0	1159.0	1259.0	1041.0	302.0	96.0	105.0
705	Ladakh	2021	34.0	44.0	50.0	46.0	28.0	17.0	17.0	0.0
706	Lakshadweep	2021	0.0	2.0	0.0	1.0	0.0	1.0	0.0	0.0
707	Puducherry	2021	135.0	139.0	146.0	151.0	192.0	131.0	49.0	68.0

Figure 2: Dataset

3.2 Data Preprocessing

In the preprocessing phase, an initial check was conducted to identify if there are any NA/null values within the dataset (Viswanath Dhanya K; 2021). This dataset was found to be having no null/NA values. However, it contained instances where certain states lacked accident count data for all 20 years, which is shown in Figure 3. Consequently, states with incomplete data across the 20-year period were eliminated. After removing this unwanted data, the dataset's count was reduced to 680 instances.

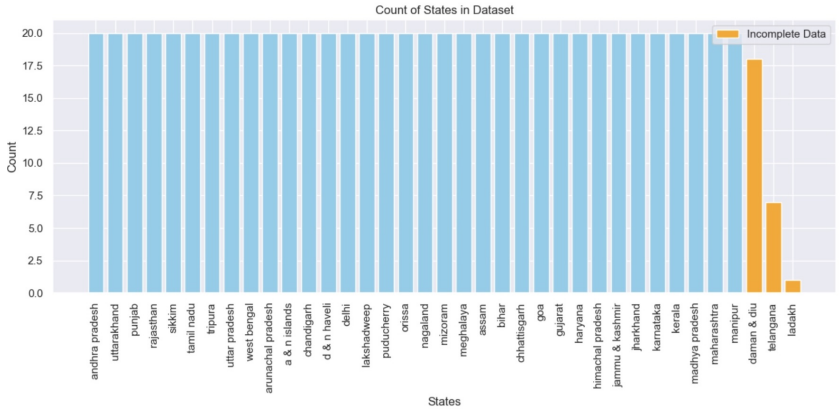
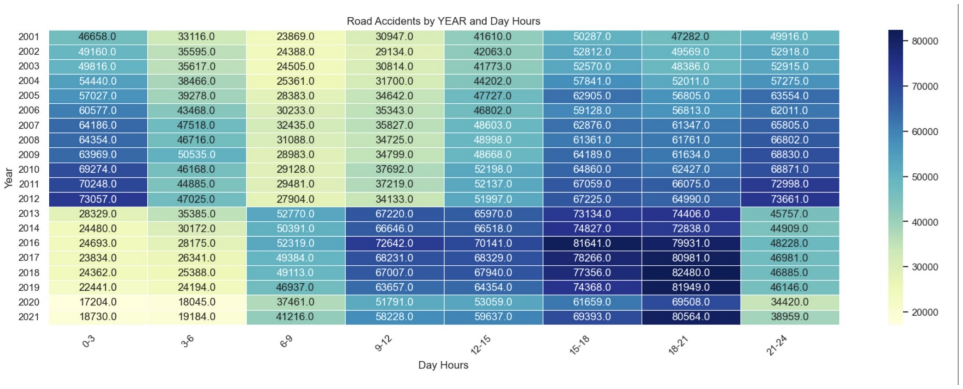


Figure 3: Representation of State-wise Data Distribution

3.3 Data Visualisation

Data visualisation refers to graphical representation of data to communicate information effectively and efficiently. It enhances understanding, facilitates communication, and helps users gain actionable insights from data analysis.



Year	0-3	3-6	6-9	9-12	12-15	15-18	18-21	21-24
2001	46658.0	33116.0	23869.0	30947.0	41610.0	50287.0	47282.0	49916.0
2002	49160.0	35595.0	24388.0	29134.0	42063.0	52812.0	49569.0	52918.0
2003	49816.0	35617.0	24505.0	30814.0	41773.0	52570.0	48386.0	52915.0
2004	54440.0	38466.0	25361.0	31700.0	44202.0	57841.0	52011.0	57275.0
2005	57027.0	39278.0	28383.0	34642.0	47727.0	62905.0	56805.0	63554.0
2006	60577.0	43468.0	30233.0	35343.0	46802.0	59128.0	56813.0	62011.0
2007	64186.0	47518.0	32435.0	35827.0	48603.0	62876.0	61347.0	65805.0
2008	64354.0	46716.0	31088.0	34725.0	48998.0	61361.0	61761.0	66802.0
2009	63969.0	50535.0	28983.0	34799.0	48668.0	64189.0	61634.0	68830.0
2010	69274.0	46168.0	29128.0	37692.0	52198.0	64860.0	62427.0	68871.0
2011	70248.0	44885.0	29481.0	37219.0	52137.0	67059.0	66075.0	72998.0
2012	73057.0	47025.0	27904.0	34133.0	51997.0	67225.0	64990.0	73661.0
2013	28329.0	35385.0	52770.0	67220.0	65970.0	73134.0	74406.0	45757.0
2014	24480.0	30172.0	50391.0	66646.0	66518.0	74827.0	72838.0	44909.0
2016	24693.0	28175.0	52319.0	72642.0	70141.0	81641.0	79931.0	48228.0
2017	23834.0	26341.0	49384.0	68231.0	68329.0	78266.0	80981.0	46981.0
2018	24362.0	25388.0	49113.0	67007.0	67940.0	77356.0	82480.0	46885.0
2019	22441.0	24194.0	46937.0	63657.0	64354.0	74368.0	81949.0	46146.0
2020	17204.0	18045.0	37461.0	51791.0	53059.0	61659.0	69508.0	34420.0
2021	18730.0	19184.0	41216.0	58228.0	59637.0	69393.0	80564.0	38959.0

Figure 4: Correlation plot of dataset

Figure 4 represents the correlation plot of the dataset. This visualization offers insights into the patterns and variations within the annual accident counts. The plot highlights a decline in midnight accidents from 2012, which previously exhibited the highest occurrence in the previous decade.It also indicates that the accident rate reaches its peak during the 15-21 interval over the span of two decades, while remaining low during the 3-9 interval.

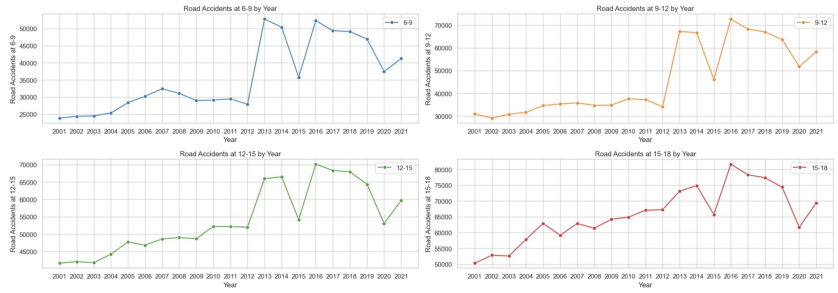


Figure 5: Plot of variation in accident count over 2 decades at different time interval

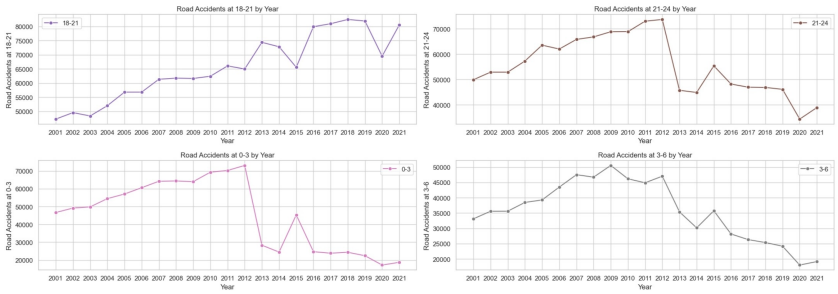


Figure 6: Plot of variation in accident count over 2 decades at different time interval

Figure 5 and Figure 6 is the line graph which represents the variation in the count of accident each year for each time interval. This visualization aids in comprehending the yearly variations in accident occurrences within specific time frames.

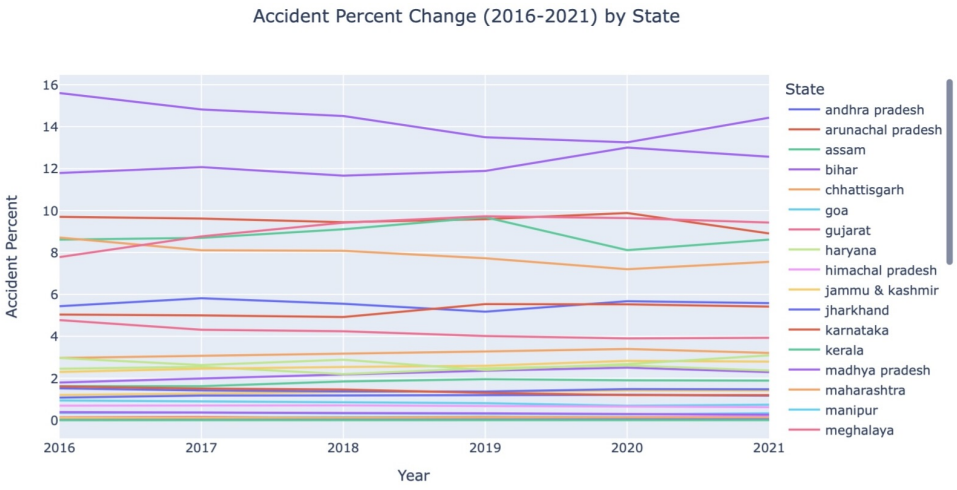


Figure 7: Accident percent change (2016-2021) by state

Figure 7 represents the variation in accident count in each state and union territories spanning year 2016 t 2021. This plot helps us in understanding the variation in the frequency of accident in each state for last 5 years. This helps us to take necessary precautions for reducing accidents as well as fatalities.

3.4 Splitting the data

Two sets of the dataset are used to train the model. The model's performance is tested using data from 2021, while training data is derived from 2001 to 2020. The best-performing model will forecast the data for 2022 based on its performance.

3.5 Model Training

This phase is called as model selection phase. Here we decide which model/algorithm best suits the given problem. In this project, we have used multi-target variable which is also called as multi-output or multi-dimensional target variable prediction. It is used in the scenarios where the goal is to predict multiple target variables simultaneously rather than a single outcome where each target variable represents a distinct aspect or dimension of the prediction problem (Pourroostaei Ardakani and Cheshmehzangi; 2023). This multi-target technique allows for a more granular and detailed prediction of accident counts across different time periods. It can be particularly beneficial in cases where decision-making or resource allocation depend on an awareness of changes or trends in incidents over specified intervals . In this project, the target variable represents accident counts at each time interval in steps of 3. There are multiple algorithms to consider for this analysis. Based on prior research findings, the chosen algorithms for this study are Random Forest G and R (2023) Augustine and Shukla (2022), Linear Regression Chirag and Supreetha (2022) and Decision Tree G and R (2023) . Different approaches to hyperparameter tuning is applied to enhance the performance of these algorithms.

3.6 Model Evaluation

Assessing the performance of the trained model is the crucial step in model building. It is important to examine trained model using test / unseen data. This is usually performed using set of evaluation matrices. The evaluation matrices vary based on the type of problem that we are solving such as regression and classification. The section where we evaluate the model's performance on fresh data is essential as it allows us to figure out how well the model can handle recent data. AIt also allows us to assess the relative efficacy of many models and select the best fit for the given situation. Some of the common metrics used for regression problems include Mean Absolute Error (MAE), R-Squared value, and Root Mean Squared Error (RMSE) G and R (2023). The summary of all outputs will be compared to identify the best performing model.

4 Design Specification

This section outlines the fundamental components that directs the implementation of the project. It includes detailed explanations of the methods and algorithms used and outlines on expected performance metrics used during analysis of the models.

4.1 Modelling Technique

The machine learning algorithm serves as a method through which AI systems execute operations, typically predicting output values based on given input data (Banerjee et al.; 2022). Classifying the data using Regression or classification methods are the two main key function of machine learning. In this project, we perform regression analysis using multiple machine learning models to determine the best performing algorithm.

1. Decision Tree Regressor (DT)

   Decision Trees is supervised learning techniques that can be applied to both regression and classification problems (G and R; 2023). It follows a tree structure, where the nodes within the structure represent features present in a dataset. Decision rules are represented by branches of the tree, while each leaf node signifies an outcome or result (Paul et al.; 2020).Decision tree regression analyzes the attributes of an entity and constructs a model within a tree-like format to predict future data, offering valuable continuous output. This output/result is not categorical, it does not rely only on a discrete, predefined set of numbers or values (G and R; 2023). It constructs a tree-like structure by recursively partitioning the data into smaller subsets based on the most significant predictor variables.

2. Random Forest Regressor (RF)

   Random forest is an collective approach that can perform both classification and regression tasks using multiple decision trees. The core concept of the algorithm is to train numerous decision trees and then utilise the average of each individual tree's results for regression tasks hence it reduces the impact of noise findings (Paul et al.; 2020). The algorithms advantages lie in its ability to resist overfitting, manage nonlinear patterns effectively, and offer insights into feature importance. This technique is based on using many decision trees in concert rather than depending on individual trees to determine the final outcome Augustine and Shukla (2022).

3. Linear Regressor (LR)

   Linear regressor is the Supervised machine learning algorithm. A linear relationship between independent and dependent variables is often represented by a linear regression model.It assumes that the relationship between variables is approximately linear, and the model aims to fit a line that best represents this relationship. The main advantage of linear regression is that it works very effectively with data that demonstrates a linear relationship. Whereas establishing the assumption of linearity between dependent and independent variables can be difficult (Chirag and Supreetha; 2022).

4.2 Evaluation Technique

1. Mean Absolute Error (MAE) [5]: It is a statistic for calculating the average absolute discrepancies between a dataset's actual values and its anticipated values. The average of the absolute discrepancies between the values that were seen and those that were predicted is used to calculate it. It is robust to outliers since it gives a clear picture of the average magnitude of mistakes without taking into account their direction.

2. R-Squared (R^2) value: It is a statistical measure that shows how much of the variance in the dependent or target variable in a regression model can be accounted for by the independent variables. On a scale of 0 to 1, 1 denotes flawless prediction of the target variable by the model. R-squared values that are closer to 1 indicate that the model explains more of the variance in the data, whereas values that are closer to 0 show that the model explains very less of the variance.

3. Root Mean Squared Error (RMSE): The average magnitude of the residuals, or errors between anticipated and observed values, is measured by the RMSE. The square root of the average of the squared discrepancies between the expected and actual numbers is used to calculate it. When assessing a prediction model's accuracy, root mean square error (RMSE) is frequently employed to penalise greater errors more severely than smaller errors.

5 Implementation

In this paper, we have built a predictive models which is designed to estimate the occurrence rate of road accidents in different states of India at specific time interval. The multiple procedures have been carried out to develop a effective machine learning model.

5.1 Tools Used

The project's development and implementation was been performed with the help of variety of tools. Python was used as a primary programming language due to its adaptability and extensive library options for modelling and visualisations, which makes programming simple to analyse the results. The main development environment was

[5]MAE: `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html`

Jupyter Notebooks, which allowed for effective project management and interactive coding. Essential libraries such as Pandas[6] and NumPy [7] enabled seamless data manipulation and numerical computations, while Scikit-learn [8] made it possible for the creation and evaluation of machine learning models. Visualization needs were met by Matplotlib [9] and Seaborn [10], aiding in data representation.

5.2 Data Preperation

The data is taken from the Open Government Data (OGD) Platform, India which is accessible to public (G and R; 2023). It contains data of road accident counts in different states of India at each interval of time from the year 2001 to 2021. Information spanning each year between 2001 and 2021 was downloaded and combined into a dataset using Excel. The dataset contains 10 columns which includes year, state ans time intervals 6-9 AM, 9-12PM, 12-3 PM, 3-6 PM, 6-9 PM, 9-12 AM, 12-3 AM, and 3-6 AM. It contains data for 37 different states and union territories of India. In total there are 708 rows and 10 columns in the dataset. This dataset is be analysed to provide a thorough analysis of the time distribution of road accidents in different parts of India.

5.3 Data Cleaning and splitting

The Initial step in data cleaning is checking for null values and NA's and clearning them by using dropna() function. In the next step, the plot in the Figure 3 depicts the distribution of data available for each state in India. This visualization aids in assessing the completeness of data across all mentioned states for the years spanning from 2001 to 2021. The plot reveals that three states lack data for the entire 20-year duration. Consequently, these three states are excluded from further analysis due to incomplete data coverage. There is noticeably absence of data for the year 2015 in the dataset. To address this absence, the missing values are replaced with mean of the observed data.

The data is split into training and testing set. the data from the year 2001 to 2020 is taken as training set. whereas data of 2021 is considered as a testing set.

5.4 Hyper parameter Tuning

Finding the best hyperparameter settings for a machine-learning model is known as hyperparameter tuning. Hyperparameters govern the behaviour and performance of the model.

• Random forest Regressor:

Hyperparameter tuning is done with the help of Grid Search Cross-Validation [11], which searches the specified parameter grid to find the optimal combination. Parameters like n-estimators, max-depth,min-samples-split, and min-samples-leaf are varied to identify the best-performing configuration. n-estimators and max-depth usually control the model's complexity and its ability to learn intricate patterns. min-samples-split and

[6]Pandas: `https://pandas.pydata.org`

[7]Numpy: `https://numpy.org`

[8]Scikit-learn: `https://scikit-learn.org/stable/`

[9]Matplotlib: `https://www.simplilearn.com/tutorials/python-tutorial/matplotlib`

[10]Seaborn: `https://seaborn.pydata.org`

[11]gridSearchCV: `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html`

min-samples-leaf regulate the granularity and avoids overfitting by controlling the model tree's expansion.

Best Hyperparameters for Road accident data: max depth: None, min samples leaf: 1, min samples split: 2, n estimators: 200

• Decision Tree Regressor:

Node Splitting: At each step, it selects the feature that best separates the data based on certain criteria Leaf Nodes: Here the process works until a stopping criterion is met, resulting in leaf nodes containing predicted values. Hyperparameter tuning is carried out with Grid Search Cross-Validation, which searches the specified parameter grid to find the optimal combination. Hyperparameters: The hyperparameters being tuned include max-depth: Controls the maximum depth of the decision tree. min-samples-split: Defines the minimum number of samples required to split an internal node. min-samples-leaf: Specifies the minimum number of samples required to be at a leaf node.

Best Hyperparameters for Decision Tree: max depth: None, min samples leaf: 1, min samples split: 10

6 Evaluation

The crucial phase in machine learning pipeline is evaluation process in order to guage the effectiveness of the models performance. The results of the regression models are evaluated using evaluation matrix such as MAE, R-squared and RMSE. The experiments are conducted using different algorithms such as Linear regression, decision tree regressor and random forest regressor. The results of these models are compared to identify the best suited model to predict the accident rate.

6.1 Experiment using Decision tree

The generated output in Figure 8 represents the assessment metrics of the Decision Tree model applied to predict accident rates across various time intervals. Each row in the presented table corresponds to a distinct time period, spanning from 6 AM to 6 AM the following day, exhibiting how effectively the model forecasts accident frequencies during these intervals. The result obtained from this model appears relatively average based on the evaluation metrics displayed in the table. The Mean Absolute Error (MAE) values range from approximately 177 to 260, indicating the average magnitude of errors in predicting accident counts within specific time intervals. Generally, lower MAE values signify better accuracy. R-squared values ranges around 0.87 to 0.98. Higher R-squared values closer to 1 imply a better fit of the model to the actual data. Additionally, Root Mean Squared Error (RMSE) values range from approximately 298 to 522, reflecting the square root of the average squared differences between predicted and actual values. Lower RMSE values indicate better accuracy in prediction. Overall, the model exhibits average predictive capability, as seen through the evaluation metrics across various time intervals.

6.2 Experiment using Linear regression

The output provided in Figure 9 presents the performance evaluation metrics for a Linear Regression model applied to predict accident frequencies within various time intervals. The evaluation metrics include Mean Absolute Error (MAE), R-squared, and Root Mean Squared Error (RMSE) for each segmented time period (e.g., 6-9 AM, 9-12 PM, etc.).

	Time	MAE	R-squared	RMSE
0	6-9	186.928	0.960102	314.626
1	9-12	173.684	0.982027	298.445
2	12-15	177.397	0.983781	294.656
3	15-18	222.207	0.978393	396.986
4	18-21	249.932	0.979635	474.813
5	21-24	271.152	0.878487	522.059
6	0-3	263.045	0.524352	505.74
7	3-6	269.022	0.609901	469.322

Figure 8: Evaluation matrix of Decision Tree

The Mean Absolute Error (MAE) values range from approximately 577 to 2513. The R-squared values are consistently low across all time intervals close to zero, signifying that the model explains very little variance in the data. The Root Mean Squared Error (RMSE) values are notably high, ranging from around 732 to 3281. Overall, these metrics suggest that the Linear Regression model shows diminished prediction results in predicting accident frequencies across different time intervals.

	Time	MAE	R-squared	RMSE
0	6-9	1317.96	0.0094158	1567.71
1	9-12	1825.83	0.00737236	2217.95
2	12-15	1883.81	0.0148463	2296.45
3	15-18	2186.17	0.0166207	2678.18
4	18-21	2513.35	0.0271615	3281.7
5	21-24	1244.75	0.0138433	1487.24
6	0-3	577.307	0.00183072	732.632
7	3-6	635.301	-0.0409616	766.656

Figure 9: Evaluation matrix of Linear regression

6.3 Experiment using Random forest regressor

The output presented in Figure 10 the evaluation metrics for a Random Forest model utilized to predict accident frequencies within different time intervals. The 'MAE' (Mean Absolute Error) values for different time periods range from approximately 177 to 304, signifying the average magnitude of errors between predicted and actual accident counts. The 'R-squared' values range between 0.92 and 0.99, indicating the goodness of fit of the

model to the actual data. the Root Mean Squared Error (RMSE) values are observed to be relatively low, ranging from about 276 to 581. These metrics collectively suggest that the Random Forest model outperforms the linear regression model previously employed, showcasing better predictive capabilities in understanding the relationship between time intervals and traffic accidents, potentially due to its ability to capture non-linear patterns and interactions among various features within the dataset.

	Time	MAE	R-squared	RMSE
0	6-9	177.799	0.969439	275.36
1	9-12	181.637	0.977278	335.567
2	12-15	206.618	0.975737	360.394
3	15-18	266.766	0.970287	465.538
4	18-21	304.562	0.969541	580.682
5	21-24	277.118	0.871549	536.755
6	0-3	273.255	0.528	503.796
7	3-6	263.259	0.668395	432.707

Figure 10: Random Forest Regressor

6.4 Experiment using random forest regressor - After outlier removal

The Random Forest model's evaluation across different time intervals illustrates its varied performance in predicting accident rates. For time slots like 6-9, 9-12, 12-15, and 15-18, the model showcases relatively low Mean Absolute Error (MAE) values, indicating an average error range of around 56 to 148 in predicting accident counts. The R-squared values demonstrate high performance across most time intervals, with values ranging between 0.962 and 0.994, signifying an excellent fit of the model to the data. the Root Mean Squared Error (RMSE) values have decreased, ranging from about 118 to 266, further indicating the improvement in prediction accuracy. This suggests that the Random Forest model, after the outlier removal, maintains its robust predictive capability and offers enhanced accuracy in understanding the relationship between time intervals and traffic accidents within the dataset.

Figure 12 represents the normal distribution graph also known as bell curve of a best performed model. This graphical representation is important in understanding, analyzing, and making predictions about data in diverse fields. A bell curve showing up in our results is the best case outcome for our model. This bell-shaped curve indicates that our data distribution resembles a normal distribution, in which most data points are located around the mean and less are distributed towards both ends. This occurrence is positive since it shows that the observed data and the model's predictions are in balance, indicating accuracy and consistency. In statistics, the bell curve pattern is crucial because it shows that the data follows a predictable pattern, which is necessary for drawing valid conclusions and forecasts about the performance of the model.

	Time	MAE	R-squared	RMSE
0	6-9	68.1767	0.989902	126.066
1	9-12	96.6517	0.98507	207.736
2	12-15	82.6955	0.992874	141.925
3	15-18	96.7515	0.993629	154.137
4	18-21	147.665	0.985408	266.167
5	21-24	72.4964	0.988301	125.945
6	0-3	55.9112	0.968126	117.951
7	3-6	64.641	0.96164	131.255

Figure 11: RF after outliers removal

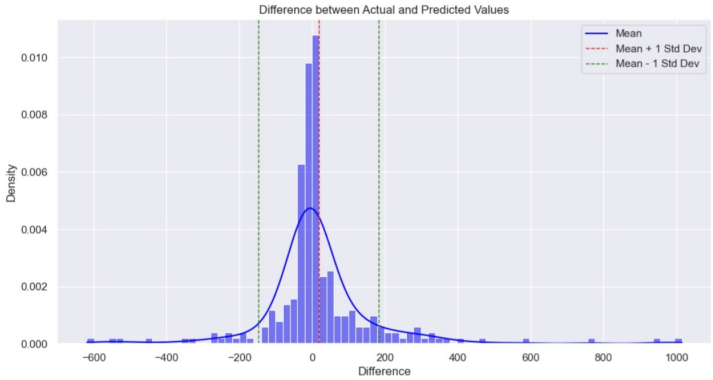
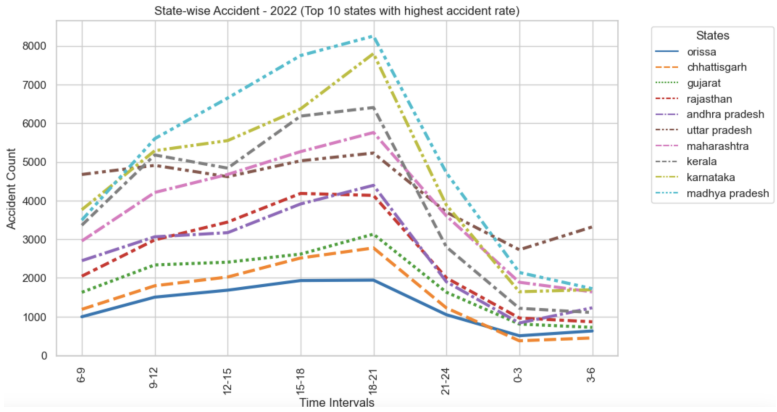


Figure 12: Normal Distribution

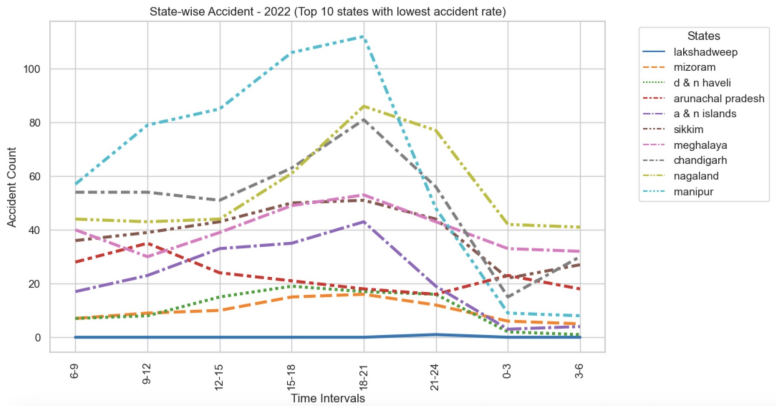
6.5 Discussion

In this project, we employed three distinct machine learning algorithms to forecast the accident rate in India. The Decision Tree Regressor model was constructed. The model showed moderate predictive capabilities. However, the outcomes were unsatisfactory. The application of the Linear Regressor model resulted in notably poor performance, rendering it unsuitable for our predictive needs. The Random Forest algorithm, used as the third model, demonstrated superior performance compared to the other models, exhibiting an R-squared value consistently around 0.9, indicating strong predictive accuracy. In an attempt to refine predictions, we explored the effect of outlier removal from the dataset on the Random Forest model's performance. Small perturbations in removing the outliers contributed to improvement in the model's performance.

In contrast to prior studies focused on predicting road accidents (Li et al.; 2020), (Lv et al.; 2015), this project has demonstrated notable advancements and valuable contributions. It has outperformed earlier efforts in offering an improved and comprehensive analysis of accident occurrences by utilising dataset that span two decades and include a variety of time intervals. This approach to learning traffic accidents over a long period of time across Indian states certainly contributed an important difference to the field



(a) Top 10 States with Highest Accident rate in 2022



(b) Top 10 States with Lowest Accident Rate in 2022

Figure 13: Predicted output analysis

by outperforming earlier studies and establishing a standard by offering better value in terms of data reliability.

The predicted accident frequency for the year 2022 is visualised in the Figure 13 and Figure 14. In Figure 13, the plot illustrates the accident count across various time intervals in each state and union territory of India.Figure 13a represents the states with highest accident rate in 2022 whereas Figure 13b represents the states with lowest accident rate in 2022. It clearly displays that Madhya Pradesh has the highest contribution to accidents in 2022, while Lakshadweep has notably fewer accidents.

Figure 14 presents a correlation plot specifically for the accident count in 2022. This plot showcases that accidents reached their peak during the time interval of 15-21 and were at their lowest between 0-3 hours.

7 Conclusion and Future Work

The research aims to develop an effective machine learning algorithm to predict the frequency of traffic accidents in India, and analyse whether these predictive outcomes can contribute to controlling and reducing the accident rate. India accounts to 6 percent of global road accidents while owning only 1 percent of global vehicle population. This growing trend of death rate due to accident in India highlights the urgency of implementing effective strategies to reduce accidents and enhance road safety. The data for this research is taken from the open government website spanning from the year 2001 to

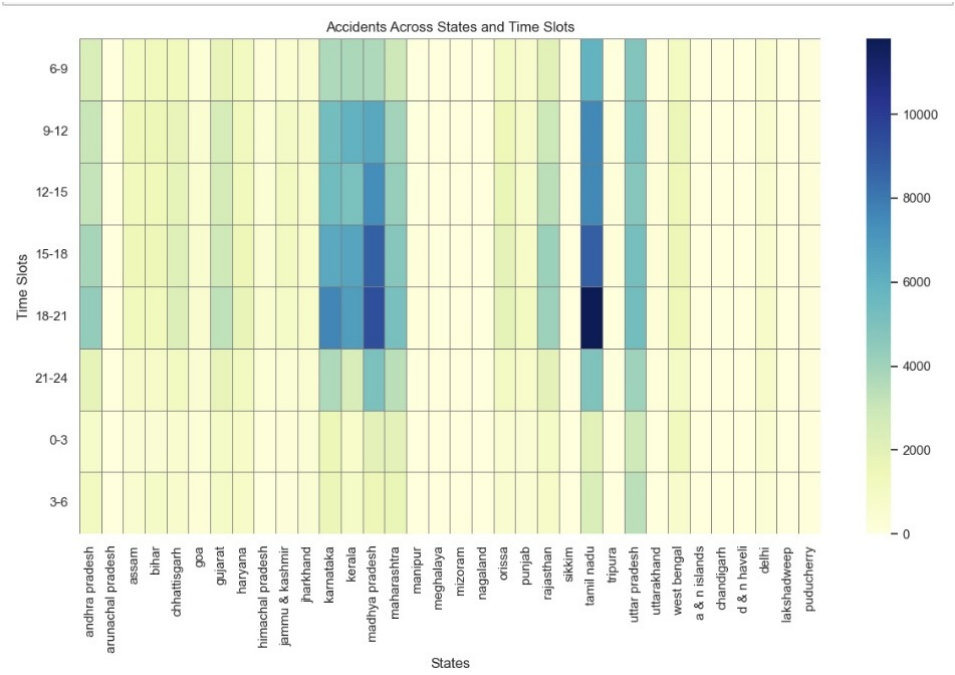


Figure 14: Correlation plot for 2022

2021. Python programming language is used for analysis and model development. The dataset is visualised to understand the trends and patterns in the count of road accident over 2 decades. Several machine learning models such as random forest, decision tree and linear regressor were built to identify the best performing model using evaluation matrix. Random forest is found to give the best result compared to other algorithms. The prediction of the model for the year 2022 was analysed. the analysis showed that accidents peaked frequently in different parts of India between the time period of 15:00 to 21:00. The time period with the fewest accidents occurred during the night, specifically between 3 AM and 6 AM. The model's insights have the potential to significantly improve traveller safety and aid authorities in developing plans to reduce and eliminate fatal accidents on Indian roads. The project holds several promising avenues for future exploration. The government can effectively use these optimised models to lower traffic accidents and execute traffic safety regulations.

7.1 Acknowledgement

I want to express my sincere gratitude to Professor Shubham Shubham for all of his help and advice with my research. His profound insights significantly influenced the course of this work. I express my gratitude to the National College of Ireland for providing me with the opportunity to carry out this study.I express my heartfelt thanks to my family, and my friends for their unwavering support and encouragement during my pursuit of a master's degree. Their constant backing has been helping me finish my research project.

References

Abdulhafedh, A. (2017). Road crash prediction models: Different statistical modeling approaches, *Journal of Transportation Technologies* **07**(2): 31–32.

Augustine, T. and Shukla, S. (2022). Road accident prediction using machine learning approaches, pp. 808–811.

Banerjee, K., Bali, V., Sharma, A., Aggarwal, D., Yadav, A., Shukla, A. and Srivastav, P. (2022). Traffic accident risk prediction using machine learning, pp. 76–82.

Bi, W. and Yu, F. (2020). Research on the prediction framework of road traffic accidents based on idwpso, pp. 106–111.

Chirag, P. and Supreetha, M. (2022). Road accident prediction and classification using machine learning, pp. 1–8.

Chitradevi, C. K. D. and Rajan, A. (2022). Predictive analytics of road accidents using machine learning, *2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* pp. 1782–1786.

Francesca La Torre, Monica Meocci, L. D. V. B. N. T. A. P. (2019). Development of an accident prediction model for italian freeways, *Accident Analysis Prevention* **124**: 1–11.

G, M. and R, R. H. (2023). Prediction of road accidents in the different states of india using machine learning algorithms, pp. 1–6.

Jaber, A. (2022). Severity of pedestrian crashes in developing countries: Analysis and comparisons using decision tree techniques, *Southeast Asian Journal of Tropical Medicine and Public Health* **11**.

Jonas Lundberg, Carl Rollenhagen, E. H. (2009). What-you-look-for-is-what-you-find ‚Äì the consequences of underlying accident models in eight accident investigation manuals, *Safety Science* **47**(10): 1297–1311.

Jutaek Oh, Simon P. Washington, D. N. (2006). Accident prediction model for railway-highway interfaces, *Accident Analysis Prevention* **38**(2): 346–356.

Kraonual, Sunee Lim, A. (2020). Factors associated with hospital mortality due to road traffic accidents among pedestrians in southern thailand, *Southeast Asian Journal of Tropical Medicine and Public Health* **51**: 763–770.

Laura Eboli, Carmen Forciniti, G. M. (2020). Factors influencing accident severity: an analysis by road accident type, *Transportation Research Procedia* **47**: 449–456. 22nd EURO Working Group on Transportation Meeting, EWGT 2019, 18th ‚Äì 20th September 2019, Barcelona, Spain.

Li, W., Zhao, X. and Liu, S. (2020). Traffic accident prediction based on multivariable grey model, *Information* **11**(4).
URL: *https://www.mdpi.com/2078-2489/11/4/184*

Lv, Y., Duan, Y., Kang, W., Li, Z. and Wang, F.-Y. (2015). Traffic flow prediction with big data: A deep learning approach, *IEEE Transactions on Intelligent Transportation Systems* **16**(2): 865–873.

Mallahi, I. E., Dlia, A., Riffi, J., Mahraz, M. A. and Tairi, H. (2022). Prediction of traffic accidents using random forest model, pp. 1–7.

Massimo Bertolini, Davide Mezzogori, M. N. F. Z. (2021). Machine learning for industrial applications: A comprehensive literature review, *Expert Systems with Applications* **175**: 114820.

Nawaf, A. and Fred, M. (2023). An analysis of day and night bicyclist injury severities in vehicle/bicycle crashes: A comparison of unconstrained and partially constrained temporal modeling approaches, **40**(100301).

Nedjmedine, O. and Tahar, M. (2022). Analysis of road accident factors using decision tree algorithm: a case of study algeria, pp. 1–6.

Parathasarathy, Soumya, Das, J. and Saravanakumar, Merjora, A. (2019). Using hybrid data mining algorithm for analysing road accidents data set, pp. 7–13.

Paul, J., Jahan, Z., Lateef, K. F., Islam, M. R. and Bakchy, S. C. (2020). Prediction of road accident and severity of bangladesh applying machine learning techniques, pp. 1–6.

Pourroostaei Ardakani, S. and Cheshmehzangi, A. (2023). Data-driven multi-target prediction analysis for driving pattern recognition: A machine learning approach to enhance prediction accuracy, pp. 63–80.
URL: *https://doi.org/10.1007/978-981-99-6620-2*$_4$

S. Basnyat, N. Chozos, P. P. (2006). Multidisciplinary perspective on accident investigation, *Reliability Engineering System Safety* **91**(12): 1502–1520. Complexity in Design and Engineering.

Shabani, M. E. F. H. S. (2014). Prediction of crash severity on two-lane, two-way roads based on fuzzy classification and regression tree using geospatial analysis, *Journal of Computing in Civil Engineering* **29**.

Shad, R., M. A. and Moghimi, R. (2013). Extraction of accidents prediction maps modeling hot spots in geospatial information system,, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* .

Vipin Na, R. T. (2021). Road traffic accident mortality analysis based on time of occurrence: Evidence from kerala, india, *Clinical Epidemiology and Global Health 2021* **11**.

Viswanath Dhanya K, Preethi R, N. R. B. (2021). A road accident prediction model using data mining techniques, pp. 1618–1623.

Wang, Xuelian Zhang, Z. L. H. X. J. D. M. M. Q. (2022). Research on the prediction of traffic accident severity based on bp neural network, *Advances in Transdisciplinary Engineering* **30**: 1117 – 1126.

Zhang, Daowen Wang, C. J. J. L. H. (2022). Analysis of the severity of vehicle to vehicle accidents considering the interaction of factors, *Journal of Automotive Safety and Energy* **13**: 643 – 650.

Zhankaziev, S. V., Zamytskih, A., Vorobyev, A. I., Gavrilyuk, M. V. and Pletnev, M. G. (2022). Predicting traffic accidents using the conflict coefficient, pp. 1–6.