International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

ROAD ACCIDENT ANALYSIS USING MACHINE LEARNING

Dr. Vilas Joshi[1]

Shivansh Gautam[2], Prashant Singh[3], Sonal Raj[4], Prince Singh[5]

Computer Engineering Department, ISBM College of Engineering Nande, Pune-412115,
India Savitribai Phule Pune University

Abstract: **Today, one of the top concerns for governments is road safety. Although there are various safety precautions in place to prevent auto accidents, they cannot be completely avoided. To lessen the harm caused by traffic accidents, the primary goal now is to determine what causes them. In this study, we use machine learning techniques to identify the causes of traffic accidents. By creating precise prediction models that can automatically separate distinct unintentional instances, patterns involved in diverse situations can be identified. The development of safety measures and the application of these classification approaches will help avoid accidents. Although there are numerous inventories in the automotive sector to create and construct safety features for cars, road accidents are inevitable. Both urban and rural regions see a high rate of accidents. By creating precise prediction models that can automatically separate distinct unintentional instances, patterns involved in diverse situations can be identified. These clusters will help create safety precautions and prevent mishaps. We think we can use some ML techniques to reduce accidents as much as possible while using limited resources.**

I. INTRODUCTION

Accident-related fatalities and injuries are predicted to be an increasingly common problem. Since the invention of the vehicle, traffic safety has been a major problem. In India, road accidents claimed the lives of 1,53,972 individuals in total in 2021, according to the Ministry of Road Transport and Highways (MoRTH). This equals an average of 422 fatalities per day. Approximately 67 percent of all unintentional fatalities occur in people between the ages of 18 and 45, which is the age group most frequently affected by traffic accidents. Statistics have also shown that young individuals, who make up a significant portion of the workforce, have a relatively high death rate in traffic accidents. Various road safety measures are required to solve this issue.

Research interest in figuring out the important impact of the severity of accidents caused by traffic accidents has increased in recent years. Accident analysis is built on accurate and thorough accident records. The correctness of the data, record retention, and data analysis are some of the aspects that affect how well accident records are used. Numerous methods have been used to analyze this issue using this scenario.

The prime goal of this research paper is to analyze road accidents and determines the severity of an accident by applying advanced machine learning techniques. There exist so many developed methods in machine learning to examine this problem.

II. PROBLEM DEFINITION

To handle the enormous number of road accidents in India a precise analysis is required. This analysis will be done more deeply to determine the intensity of road accidents by using different machine learning techniques like supervised learning, unsupervised learning, etc.

Road accident analysis using machine learning to identify the factors that contribute to accidents and develop interventions to address these factors. This can be done by using machine learning to analyze historical accident data and identify patterns that would be difficult to see with the naked eye. Once the factors that contribute to accidents have been identified, interventions can be developed to address these factors. For example, if it is found that speeding is a major factor in accidents, interventions could be developed to educate drivers about the dangers of speeding and to enforce speed limits.

Machine learning is an effective tool that can be used to examine data on traffic accidents and spot trends that are difficult to spot with the unaided eye. Machine learning can help to make our roads safer by identifying the factors that cause accidents and creating interventions to address these factors.

III. PROPOSED SYSTEM

An ML-powered web app that predicts accident severity based on the current conditions. It is trained with 25 Thousand accident records over 2010-2021. More data means greater accuracy. The purpose of such a model is to be able to predict accidents based on the conditions that will be more prone to accidents, and therefore take preventive measures. We will even try to locate more precisely future accidents in order to provide faster care and precautionary service.

We will train our model with multiple algorithms and only choose the best one with high accuracy. This will ensure that our model is as accurate as possible and can provide the most accurate predictions.

IV. LITERATURE SURVEY

The research paper "Accident Analysis on National Highway-3 Between Indore to Dhamnod" by Kundan Meshram and H.S. Goliya[1]. A study of accident data on National Highway-3 between Indore and Dhamnod, India, found that the number of accidents had increased over the past few years. The most common causes of accidents were speeding drunk driving, and driver fatigue. The majority of accidents occurred during the night and on weekends. The study concluded by making a number of recommendations to improve safety on this highway, including increasing the number of speed bumps, installing more streetlights, and increasing the number of traffic police patrols.

The paper "Road Accident Analysis using Machine Learning" by Jayesh Patil, Mandar Prabhu, Dhaval Walavalkar, and Vivian Brian Lobo. This paper explores the application of machine learning (ML) to road accident analysis. The authors used a dataset of over 100,000 road accidents in India to train a variety of ML algorithms, including decision trees, support vector machines, and random forests. The authors found that ML was able to accurately predict the severity of accidents, as well as the factors that contributed to accidents. The authors concluded that ML has the potential to be a valuable tool for road safety and that it can be used to identify accident hotspots, predict the severity of accidents, identify the factors that contribute to accidents, and develop interventions to prevent accidents.

The research paper "Road Accident Analysis" by Dr. Anitha Patila, Prithvish Kumbleb, Naresh Kc, and Sriharid, investigates the causes of road accidents in India. The most frequent causes of accidents are identified by the paper using a dataset of traffic accidents from 2010 to 2014. The study's findings indicate that driver errors, such as speeding, drunk driving, and distracted driving, are the most frequent causes of accidents. weather, the condition of the road, traffic congestion, defects in the vehicle, such as bad tires, brakes, or steering, and environmental elements like rain, fog, and darkness. The implications of these findings for India's road safety are covered in the paper's conclusion. The study's authors urge the government to take action against the most frequent causes of collisions, including educating motorists about the risks of speeding, drunk driving, and distracted driving, enforcing speed limits and other traffic laws, enhancing road conditions, making cars safer, and improving weather forecasting and procedures for road closures. raising public awareness of issues related to road safety. The study's authors are of the opinion that the government can lessen the frequency and severity of road accidents in India by taking these actions.

V. ASSUMPTIONS AND DEPENDENCIES

Here are some of the assumptions and dependencies that we have to consider while creating an analysis model using machine learning:

Data quality: The quality of the data used to train the model is critical. The data should be accurate, complete, and representative of the real world.

Model complexity: The complexity of the model should be appropriate for the data. A complex model may be more accurate, but it may also be more difficult to train and interpret.

Training data: The training data must be accurate and realistic. A wide range of potential accident-causing variables, such as driver behaviour, road conditions, and weather, should be included in the data.

Evaluation data: The evaluation data should be separate from the training data. This data is used to evaluate the accuracy of the model.

Deployment: The model should be deployed in a way that is accessible to users. The model should be easy to use and interpret.

Bias: The model should be free of bias. The model should not be biased against any particular group of people.

Interpretability: The model should be interpretable. The model should be able to explain its predictions.

By considering theseassumptions and dependencies, we can create a road accident analysis model that is accurate, reliable, and fair.

VI. DATA FLOW

A data-flow diagram(DFD) is a visual representation of how data moves through a process or system (usually an information system). Additionally, the DFD provides details about the inputs and outputs of each entity as well as the process itself for Road Accident Analysis using Machine Learning. A data-flow diagram lacks loops, decision rules, and control flows. Using a flowchart, specific operations based on the data can be depicted. Dataflow diagrams can be displayed using a variety of notations. A process must contain at least one of the endpoints (source and/or destination) for each data flow. Another data-flow diagram that divides a process into sub-processes can be used to represent a process in more detail. The structured analysis modeling tools include the dataflow diagram.

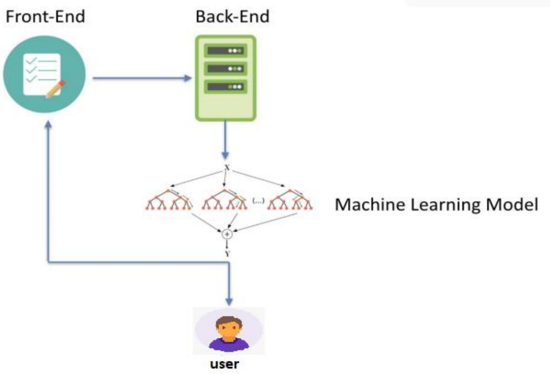


Figure 1: System model

Data is collected from a variety of sources, such as police reports, government web pages, and researchers. The data is cleaned, pre-processed, and engineered to create features that are used to train a machine-learning model. The model is evaluated and deployed so that users can use it to predict the probability of an accident. The model is monitored over time to ensure that it is still accurate.

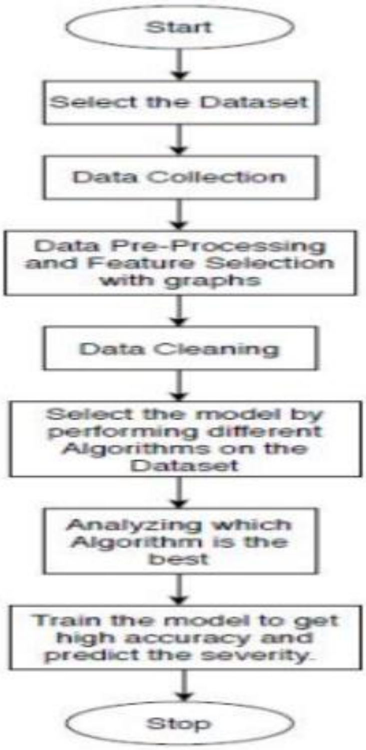

Start
Select the Dataset
Data Collection
Data Pre-Processing and Feature Selection with graphs
Data Cleaning
Select the model by performing different Algorithms on the Dataset
Analyzing which Algorithm is the best
Train the model to get high accuracy and predict the severity
Stop

Figure 2: Data Flow Diagram

VI. PROPOSED MODEL

PLANNING AND ANALYSIS:

We have read various road accident reports and documents and found that speed, gender, weather condition, and road condition are some of the basic factors that are involved in road accidents. These factors will help us to build an accurate model that can help to predict road accidents and prevent them from happening. Speed, gender, weather condition, and road condition are some of the basic factors that are involved in road accidents. By taking these factors into account, we can build an accurate model that can help to predict road accidents and prevent them from happening.

DATA GATHERING:

We will use a variety of sources to gather data for machine learning, including Kaggle. Kaggle is a great resource for finding high-quality data sets, and it makes it easy to import data into our model. We have found a data set on Kaggle that contains 25 thousand accident records from 2010 to 2021. This data set includes a variety of factors that led to the accident, such as the speed of the vehicle, the weather condition, and the road condition. This data set is high quality and well-curated, and we believe that it will be very helpful in building an accurate model that can help to predict road accidents and prevent them from happening.

DATA PRE-PROCESSING:

A key stage of the data mining process is data preparation. It is the process of preparing raw data for analysis by cleaning, formatting, and other changes. The reliability, uniformity, and completeness of the data can be improved via data preparation. This may result in simpler and more accurate data analysis outcomes. The data set we have gathered lacks proper formatting and has missing values. This indicates that before using the data for analysis, preparation is required. Among the things we must accomplish are the following:

Data cleaning involves removing errors and differences from the data.

Putting the data into a format that is compatible with our data mining tools is known as formatting the data.

Data transformation: This entails putting the data in a format that is better suited for analysis.

We will have a tidy, formatted, and transformed data set that is suitable for analysis once we have completed the data pre-processing. As a result, our results from data analysis will be simpler and more accurate.

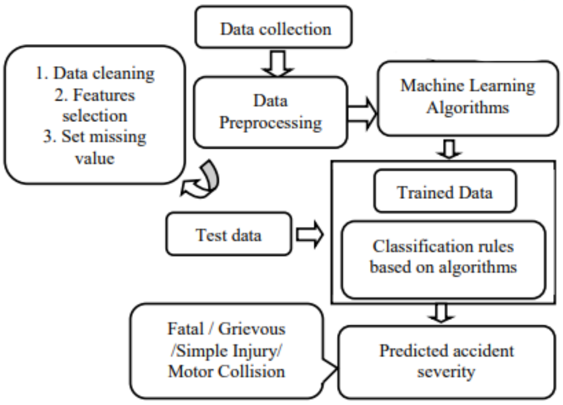

Data collection
1. Data cleaning
2. Features selection
3. Set missing value
Data Preprocessing
Machine Learning Algorithms
Trained Data
Test data
Classification rules based on algorithms
Fatal / Grievous /Simple Injury/ Motor Collision
Predicted accident severity

Figure 3: Diagram Of Model Pipeline Process

DATA VISUALIZATION:

The visual representation of data is known as data visualization. Data patterns and trends that would be challenging to spot without them are now visible to us. Finding hidden patterns in data sets with the aid of data visualization can aid in the selection of features. For example, let's say we have a data set of accidents. We can use data visualization to see how the different features of the accidents are related to each other. For example, we can see if there is a relationship between the age of the driver and the severity of the accident. We can also see if there is a relationship between the weather conditions and the severity of the accident. Once we have found hidden insights in the data set, we can use this information to select features that are important for predicting accident severity. For example, if we find that the age of the driver is a significant predictor of accident severity, we can include this feature in our model.

Finding hidden insights in data sets can be done with the help of the potent tool known as data visualization. We can choose features that are crucial for estimating accident severity by using data visualization. This can aid in the development of a more precise model that can aid in averting of accidents.

Our finding from the data visualization:

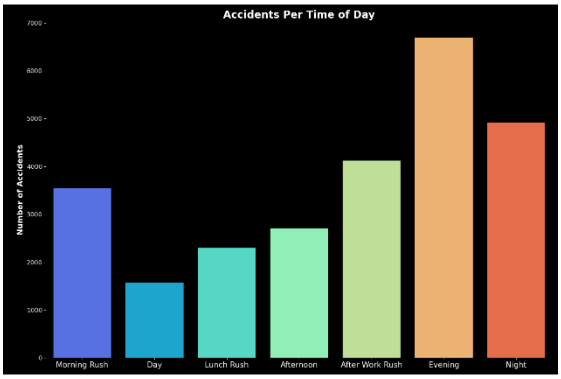


FIGURE 4: Accidents At Russ hours of the day

The histogram represents the number of accidents that occurred each month.

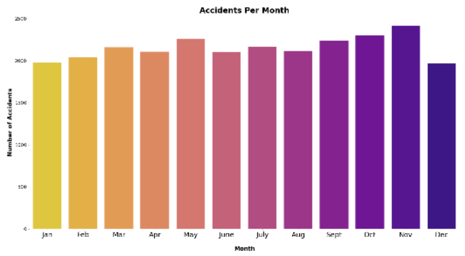


Figure: Accident per month

FEATURE SELECTION:

We used the chi-square test and visualization of various features to see which of the features are important for model building. The chi-square test is a statistical test that is used to determine the independence of two categorical variables. We used the chi-square test to see if there was a significant relationship between each feature and the accident severity. We also used visualization techniques, such as bar charts and heat maps, to see how the different features were related to each other.

Based on the results of the chi-square test and visualization techniques, we selected the following features to build our model:

- Age band of driver
- Vehicle type
- Age of vehicle
- Weather conditions
- Day of week
- Road surface conditions
- Light conditions
- Sex of driver
- Season
- Speed limit
- Accident seriousness

We believe that these features are the most important for predicting accident severity. We will use these features to build a machine learning model that can help to predict accidents and prevent them from happening.

MODEL SELECTION:

We have selected various models for model selection that can handle large amounts of data and are able to map complex relationships between the features. The models that we have selected are:

- Random forest classifier
- Support vector machine (SVM)
- XGBoost
- Gradient boosting classifier
- K-nearest neighbors (KNN)
- Decision tree classifier

We believe that these models are the most suitable for our data set and will be able to provide us with accurate predictions. We will train each model on our data set and then select the model with the highest accuracy. We will then tune the hyperparameters of the selected model to further improve its accuracy.

MODEL TRAINING:

We will first perform label encoding on the data to convert the categorical features into numerical features. We will then deal with the unbalanced categorical data of prediction. We will split the data set into train and test sets. We will then train the model on the train set and evaluate the model on the test set. We will select the model with the highest accuracy. We will then tune the hyperparameters of the selected model to further improve its accuracy. Finally, we will save the pickle file of the model, which we will use in the front end to predict the seriousness of an accident.

MODEL EVALUATION:

Model evaluation is the process of assessing the quality of a machine-learning model. It involves using the evaluation data to measure the accuracy of the model. Many different methods can be used to evaluate a machine learning model, some of the most common methods include accuracy, precision, recall, F1 score, and area under the curve (AUC). The best method for evaluating a machine learning model will depend on the specific application.

The table below shows the different models and their accuracy before hyperparameter tuning:

Algorithm	Accuracy
Random Forest Classifier	78.40%
Gradient Boosting Classifier	80.70%
SVC (Support Vector Machine)	78.40%
K-Neighbors Classifier	75.09%
Logistic Regression	79.69%
Decision Tree Classifier	79.69%
XGBoost	80.90%

Table 1: Accuracy of Different Models

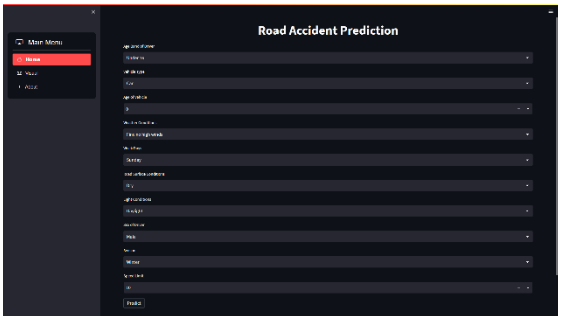
We found that the random forest classifier had the highest accuracy, precision, recall, and F1 scores. This means that the model was able to correctly predict the outcome of the test cases most often, and it was also able to identify positive cases accurately. We believe that the random forest classifier is the best model for our application because it has the highest accuracy and F1 scores. This means that the model is likely to make accurate predictions on new data.

MODEL DEPLOYMENT:

The Streamlit library was used to deploy the model. Making web applications for machine learning models is simple with the help of the free and open-source Python library known as Streamlit. With three categories—fatal, serious, and not serious—we used the pickle model of the classifier to predict the accident's seriousness. By entering accident details and receiving a prediction of the severity, users can easily interact with the model on the webpage.

Users can predict the severity of an accident, view the data, and learn more about the project using our Streamlit web application, which is a multi-page application.

By entering the accident's characteristics on the main page of the application, users can forecast the accident's severity. To determine how serious an accident will be, the application uses a machine learning pickle model. An existing dataset of accidents is used to train the model.



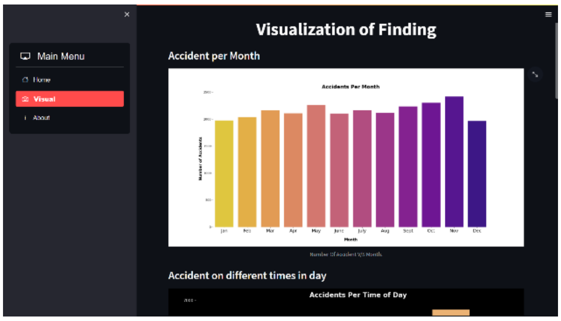
Road Accident Prediction

Age band of driver: Under 18
Vehicle type: Car
Age of vehicle: 0
Weather Conditions: Fine no high winds
Week Day: Sunday
road surface conditions: Dry
Light conditions: Daylight
Sex of Driver: Male
Season: Winter
Speed limit: 10

Predict

Figure: A Screenshot Of The Home Web-Page

The visual page of the application allows users to visualize the data. The data can be visualized in a variety of ways, including charts, graphs, and maps. The visual page allows users to explore the data and identify trends.



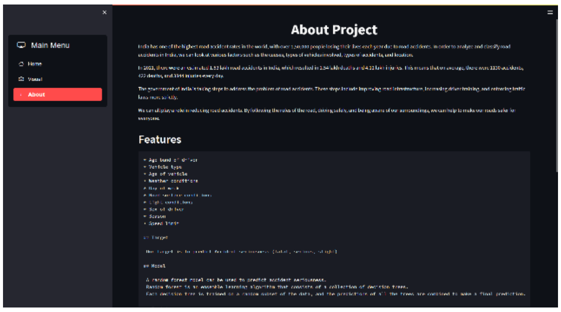
Visualization of Finding

Accident per Month

Accident on different times in day

Figure: A Screenshot Of The Visual Web-Page

The third page of the application provides information about the project. The page includes information about the goals of the project, the methods used, and the results of the project.



About Project

India has one of the highest road accident rates in the world, with over 1,50,000 people losing their lives each year due to road accidents. In order to analyze and classify road accidents in India, we can look at various factors such as the causes, types of vehicles involved, types of accidents, and location.

In 2021, there were an estimated 1.53 lakh road accidents in India, which resulted in 1.54 lakh deaths and 4.33 lakh injuries. This means that on average, there were 1130 accidents, 422 deaths, and 3344 injuries every day.

The government of India is taking steps to address the problem of road accidents. These steps include improving road infrastructure, increasing driver training, and enforcing traffic laws more strictly.

We can all play a role in reducing road accidents. By following the rules of the road, driving safely, and being aware of our surroundings, we can help to make our roads safer for everyone.

Features

```
* Age band of driver
* Vehicle type
* Age of vehicle
* Weather conditions
* Day of week
* Road surface conditions
* Light conditions
* Sex of driver
* Season
* Speed limit

## Target

 Our target is to predict Accident seriousness (fatal, serious, slight)

## Model

 A random forest model can be used to predict accident seriousness.
 Random forest is an ensemble learning algorithm that consists of a collection of decision trees.
 Each decision tree is trained on a random subset of the data, and the predictions of all the trees are combined to make a final prediction.
```

Figure: A Screenshot Of The About Web-Page

The application is created to be informative and user-friendly. Anyone looking to forecast the severity of accidents, visualize data, or learn more about the project will find the application to be a useful tool.

MAINTENANCE:

We will maintain the model by regularly adding new data to the training set and retraining the model. This will help to ensure that the model remains accurate as the underlying data changes. We will also monitor the performance of the model and make adjustments as needed.

VIII. RISK ANALYSIS

Risk analysis is a process of identifying, assessing, and controlling risks. It is an important part of any project, but it is especially important for projects that involve safety. In this project, we conducted a risk analysis to identify and assess the risks associated with using our model to predict the severity of accidents. We identified a number of risks, including:

- The model could make a wrong prediction, which could lead to an accident.
- The model could be biased, which could lead to unfair or inaccurate predictions.
- The model could be hacked, which could allow someone to manipulate the predictions.

We took a number of steps to mitigate these risks, including:

- We trained the model on a large and diverse dataset.
- We used a variety of techniques to prevent bias in the model.
- We implemented security measures to protect the model from hacking.

We believe that the risks associated with using our model are manageable. However, we are committed to continuously monitoring and improving the model to ensure that it is as accurate and safe as possible.

Finding safety measures to guard against accidents taking human lives is the aim of this project. Human casualties may result if our model's prediction is incorrect. As a result, we are committed to making sure that the model is as precise and secure as possible and are taking this project very seriously.

We believe that our model has the potential to save lives by helping to prevent accidents. We are committed to using the model responsibly and to ensuring that it is used in a way that protects human life.

IX. SOFTWARE AND HARDWARE REQUIREMENTS

Software Requirements:

- Jupyter notebook.
- VS code.
- Colab.
- Streamlit.
- Any operating system (Windows/Linux).
- Other required Python libraries (NumPy, Pandas, Matplotlib, etc.)

Hardware Requirements:

- RAM: 4 GB (minimum requirement).
- Hard Disk: 100 GB working space (minimum requirement).
- Processor: Intel Core i5 or higher, or AMD Ryzen 5 or higher
- CPU with at least 4 cores (8 cores or more is recommended)

X. OTHER SPECIFICATION

A. ADVANTAGES

1. Identify the causes of accidents: Road accident analysis can help to identify the factors that contribute to accidents, such as driver error, road conditions, and weather. This information can be used to develop safety measures that can reduce the number of accidents and injuries.
2. Develop safety measures. Road accident analysis can be used to develop safety measures that can reduce the number of accidents and injuries. For example, if an analysis finds that a particular type of intersection is dangerous, safety measures such as traffic lights or speed bumps can be installed to make the intersection safer.
3. Governments can use road accident analysis to improve road design and construction. This can be done by identifying and addressing dangerous road features, such as poorly-marked intersections or blind spots. By making roads safer, governments can help to reduce the number of accidents and injuries.
4. Road accident analysis can be used to train automatic vehicles to be more cautious and avoid mistakes. This can help to reduce the number of accidents involving automatic vehicles and make roads safer for everyone.

B. LIMITATIONS

1. Data on road accidents is often incomplete or inaccurate. This can make it difficult to identify the causes of accidents and develop effective safety measures.
2. Road accidents are often complex events with multiple contributing factors. This can make it difficult to identify all of the factors that contributed to an accident and to develop effective safety measures.
3. Road accident analysis projects often rely on data that is collected after an accident has occurred. This can make it difficult to identify the causes of accidents and to develop effective safety measures in a timely manner. Collecting real-time data about road accidents can help to address this limitation and improve road safety.
4. Risk analysis models are only as good as the data they are trained on. If a model is not accurate, it can lead to inaccurate predictions of risk, which can put people in danger. It is important to make sure that risk analysis models are as accurate as possible before using them to make decisions about risk.

C. CHALLENGES

1. Data collection. One of the biggest challenges is data collection. We need to collect a large and diverse dataset of accidents. This dataset should include information about the location, time, weather conditions, and other factors that may have contributed to the accident.
2. Data cleaning. The data that we collect will be messy and incomplete. We need to clean the data and remove any errors or inconsistencies.
3. Feature engineering. We need to engineer features from the raw data. This will allow us to train the model on features that are relevant to the prediction of accident severity.
4. Model selection. There are many different machine learning models that we can use to predict accident severity. We need to select the model that is most appropriate for our data.
5. Model training. We need to train the model on a large dataset of accidents. This process can take a long time and is required high-performance hardware, especially for complex models.

XI. FUTURE PROSPECTS

The future prospects of this project are very promising. By aiding in accident prevention, the model has the potential to save lives. It may be applied in several ways, such as:

- By governments to identify areas where road safety improvements are needed and to develop policies and programs to reduce accidents.
- By insurance companies to assess the risk of accidents and to set premiums accordingly.
- By car manufacturers to design safer cars and to develop driver assistance systems.
- By drivers to learn how to avoid accidents and to improve their driving skills.
- By self-driving car companies train self-driving cars to avoid accidents.

governments could use the model to identify areas where accidents are more likely to occur. This information could then be used to improve road safety by installing speed bumps, traffic lights, or other safety measures. Insurance companies could use the model to assess the risk of accidents for different drivers. This information could then be used to set premiums accordingly. Car manufacturers could use the model to design safer cars and develop driver assistance systems. Drivers could use the model to learn how to avoid accidents and improve their driving skills.

Self-driving car companies can use the model to improve the safety of their vehicles. This is important because self-driving cars are the future of the automobile industry, and they need to be as safe as possible. The model can be used to train self-driving cars to identify potential hazards and to take evasive action to avoid accidents. The model can be used to collect real-time data from self-driving cars and analyze it to identify potential hazards. This information can then be used to take evasive action to avoid accidents. This would make self-driving cars safer for everyone on the road.

XII. CONCLUSION

In this project, we built a machine-learning model to predict the severity of road accidents. We collected a large and diverse dataset of accidents and used a variety of techniques to train the model. The model was able to predict the severity of accidents with a high degree of accuracy.

We think that by helping in accident prevention, our model has the potential to save lives. The organization may use the model to pinpoint locations where accidents are more likely to happen and then create prevention plans for those locations. The model can also be used to train drivers on different accident-prone conditions as precautionary measures to help them stay safe and prevent them from being involved in an accident.

We think that by helping in accident prevention, our model has the potential to save lives. The organization may use the model to pinpoint locations where accidents are more likely to happen and then create prevention plans for those locations. We are excited to see how this approach is applied to save lives since we think it has the power to truly change the world.

XIII. ACKNOWLEDGMENT

We would like to take this opportunity to thank all the people who were part of this seminar in numerous ways, people who gave unending support right from the initial stage.

In particular, we wish to thank Dr. Vilas Joshi as an internal project guide who gave their co-operation timely and precious guidance without which this project would not have been a success. We thank them for reviewing the entire project with painstaking efforts and more of his, unbanning ability to spot the mistakes.

We would like to thank our H.O.D Prof. B. B. Gite for his continuous encouragement, support and guidance at each and every stage of the project.

And last but not least we would like to thank all my friends who were associated with me and helped us in preparing our project. The project named “Road Accident Analysis” would not have been possible without the extensive support of people who were directly or indirectly involved in its successful execution.

XIV. REFERENCES

- K Meshram, and S.H Goliya “Accident analysis on national highway 3 between Indore to Dhamnod”. International Journal of Application or Innovation in Engineering & Management (IJAIEM) ISSN 2319 – 4847.
- Road Accident Analysis using Machine Learning" by Jayesh Patil, Mandar Prabhu, Dhaval Walavalkar, and Vivian Brian Lobo. IEEE Pune Section International Conference 978-1-7281-9600-8/20/$31.00 ©2020 IEEE.
- Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning .by- Md. Farhan Labib, Ahmed Sady Rifat, Md. Mosabbir Hossain, Amit Kumar Das, and Faria Nawrine. Published by International Conference on Smart Computing & Communications (ICSCC). 978-1-7281-1557-3/19/$31.00 ©2019 IEEE.
- Road Accident Analysis by Dr. Anitha Patila, Prithvish Kumbleb, Naresh Kc, and Srihari M. Dadkhah. Published by Turkish Journal of Computer and Mathematics Education. Vol.12 No.10(2021), 392-396.
- Road accident in India 2020 published by the Ministry of Road Transport and Highways (MoRTH).
- Road accident in India 2021 published by the Ministry of Road Transport and Highways (MoRTH).
- Streamlit book library to create interactive books and presentations. By Sebastian Flores Benner