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.

CONTENT

1 INTRODUCTION 01

1.1 Introduction 02

1.2 Problem Definition 02

1.3 Proposed System 03

1.4 Objective 03

2 LITERATURE SURVEY 04

2.1 Literature 05

3 SOFTWARE REQUIREMENTS SPECIFICATION 11

3.1 Assumptions and dependencies 12

3.2 Project Scheduling 13

3.3 Technologies Used 14

3.3.1 Python 14

3.3.2 NumPy 14

3.3.3 Google collab 15

3.3.4 Scikit-learn 15

3.3.5 Streamlit 16

3.3.6 Jupyter Notebook 16

3.4 Software and Hardware Requirements 17

3.4.1 Software Requirements 17

3.4.2 Hardware Requirements 17

4 SYSTEM DESIGN 18

4.1 The software development life cycle (SDLC) 19

4.2 Data flow Diagram 20

4.3 System Architecture 22

4.4 Use Case Diagram 22

5 IMPLEMENTATION 24

5.1 Implementation 25

5.2 Phase 1 26

5.3 Phase 2 28

5.4 Phase 3 32

5.5 Phase 4 36

6 RISK ANALYSIS 39

6.1 Risk Analysis Introduction 40

6.2 Steps of Risk Analysis 40

6.3 Technical Feasibility 41

6.4 Real-World Risks 41

6.5 Conclusion of Risk Analysis 42

7 RESULT 44

8 CONCLUSION 49

8.1 Final Conclusion 50

8.2 Future Prospects 50

8.3 Advantages 51

8.4 Limitations 52

8.5 Challenges 52

8.6 Application 53

REFERENCES 54

LIST OF FIGURES

FIGURE	ILLUSTRATION	PAGE NO.
4.1	System model	20
4.2	Data Flow Diagram	21
4.3	System Architecture	22
4.4	Use Case Diagram	23
4.5	Web Application Interaction	23
5.1	Model Pipeline Process	27
5.2	Accidents at Russ hours of the day	29
5.3	Accidents per month	29
5.4	Heat Map od Selected Features	31
5.5	Random Forest Model	34
5.6	Confusion Matrix	35
5.7	A Screenshot of the Home Web-Page	37
5.8	A Screenshot of the Visual Web-Page	37
5.9	A Screenshot of the About Tab of Web-Page	38
7.1	Home page with Serious Prediction	45
7.2	Home page with Fatal Prediction	46
7.3	Home page with Not-Serious Prediction	46
7.4	Home page Navagation Bar	47
7.5	Visual page screenshot-1	47
7.6	Visual page screenshot-2	48
7.7	About Page	48

LIST OF TABLES

TABLE	ILLUSTRATION	PAGE NO.
5.1	Accuracy of Different Model	33

CHAPTER 1
INTRODUCTION

1.1 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) [1]. 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.

1.2 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.

1.3 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 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.

1.4 Objective

The objective of Project is to develop an ML-powered web app that predicts accident severity based on the current conditions. The app will be trained with 25,000 accident records over 2010-2021. 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. The app will also try to locate future accidents more precisely in order to provide faster care and precautionary service.

The app will be developed using a variety of machine learning algorithms, and the best one with high accuracy will be chosen. This will ensure that the app is as accurate as possible and can provide the most accurate predictions.

The app will be made available to the public for free, and it will be used to help make our roads safer. The app will also be used to provide valuable data to researchers and policymakers, which can be used to develop new strategies for preventing accidents.

The development of Project is a significant step towards making our roads safer. The app will provide a valuable tool for predicting accidents and taking preventive measures. The app will also provide valuable data to researchers and policymakers, which can be used to develop new strategies for preventing accidents.

CHAPTER 2
LITERATURE SURVEY

2.1 Literature

Many researchers have carried out research work in the area of road accidents. some of them have analyzed accident data in different ways. some of them identification of blackspot zone. some of them have developed accident models for forecasting future accident trends. They have also proposed strategies for road safety. in the present chapter literature review is carried out covering the different issues related to road accidents and road safety.

A review of the effect of traffic and weather characteristics on road safety. Despite the existence of generally mixed evidence on the effect of traffic parameters, a few patterns can be observed. for instance, traffic flow seems to have a non-linear relationship with accident rates, even though some studies suggest a linear relationship with accidents. regarding weather effects, the effect of precipitation is quite consistent and leads generally to increased accident frequency but it does not seem to have a consistent effect on severity. The impact of other weather parameters on safety, such as visibility, wind speed, and temperature is not found straightforwardly so far. The increasing use of real-time data not only makes it easier to identify the safety impact of traffic and weather characteristics but most importantly makes possible the identification of their combined effect. The more systematic use of these real-time data may address several of the research gaps identified in this research.

1st Published Paper

The research paper "**Accident Analysis on National Highway 3 between Indore to Dhamnod**" by K Meshram and S.H Goliya was published in the International Journal of Application or Innovation in Engineering & Management (IJAIEM) in 2013[2]. The paper analyzes the causes of accidents on a 100-kilometer stretch of National Highway 3 (NH-3) between Indore and Dhamnod in the Indian state of Madhya Pradesh.

The authors collected data on accidents that occurred on NH-3 from 2009 to September 2011. They found that there were a total of 1,200 accidents during this period, resulting in 1,000 injuries and 100 fatalities. The most common type of accident was a head-on

collision, accounting for 40% of all accidents. Other common types of accidents included rear-end collisions, sideswipes, and single-vehicle crashes.

The authors identified a number of factors that contribute to accidents on NH-3. These factors include:

- Driver error: The most common cause of accidents on NH-3 was driver error. This included speeding, driving under the influence of alcohol or drugs, and not paying attention to the road.
- Road conditions: The road conditions on NH-3 were also a factor in many accidents. The road was narrow and winding in some areas, and there were a number of potholes and other hazards.
- Traffic volume: The traffic volume on NH-3 is high, especially during peak hours. This can lead to congestion and driver frustration, which can increase the risk of accidents.

The authors recommend a number of measures to reduce the number of accidents on NH-3. These measures include:

- Improving driver education and training
- Increasing enforcement of traffic laws
- Improving road conditions
- Reducing traffic congestion

The research paper by Meshram and Goliya provides valuable insights into the causes of accidents on NH-3. The authors' recommendations can be used to develop effective strategies to reduce the number of accidents on this important highway.

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.

2nd Published paper:

The paper "**Road Accident Analysis using Machine Learning**" by Jayesh Patil, Mandar Prabhu, Dhaval Walavalkar, and Vivian Brian Lobo[3]. 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 authors used a machine learning algorithm to classify the accidents into different categories. The algorithm was trained on a dataset of previously classified accidents. The authors evaluated the performance of the algorithm on a test set of accidents. The algorithm was able to correctly classify the accidents with an accuracy of 90%. The authors used the machine learning model to identify the factors that contribute to road accidents.

The factors that were found to be most important were:

- Driver error: The most common cause of accidents was driver error. This included speeding, driving under the influence of alcohol or drugs, and not paying attention to the road.
- Weather conditions: The weather conditions were also a factor in many accidents. Rain and fog were found to be the most dangerous weather conditions.
- Road conditions: The road conditions were also a factor in many accidents. Potholes and poor visibility were found to be the most dangerous road conditions.

The authors recommend a number of measures to reduce the number of road accidents. These measures include:

- Improving driver education and training
- Increasing enforcement of traffic laws
- Improving road conditions
- Providing better weather information to drivers

The research paper by Patil et al. provides a valuable contribution to the field of road safety. The authors' machine learning approach can be used to identify the factors that contribute to road accidents. This information can be used to develop effective strategies to reduce the number of accidents.

In addition to the factors identified by Patil et al., other potential causes of road accidents include:

- Poor visibility due to fog or rain
- Lack of roadside safety barriers
- Presence of animals on the road
- Construction zones

The use of machine learning to analyze road accidents is a promising new approach. Machine learning algorithms can be used to identify patterns in data that would be difficult or impossible to identify manually. This information can be used to develop effective strategies to prevent road accidents.

3rd Published paper:

The research paper "**Road Accident Analysis**" by Dr. Anitha Patila, Prithvish Kumbleb, Naresh Kc, and Sriharid, investigates the causes of road accidents in India[4]. 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.

The paper presents a study of road accidents in India. The authors collected data on road accidents from the Ministry of Road Transport and Highways of India. The data included information on the date, time, location, and severity of each accident. The authors also collected weather data and information on the road conditions.
The authors used a variety of methods to analyze the data. They used descriptive statistics to describe the characteristics of the accidents. They also used inferential statistics to test for relationships between the variables.

The authors found that the most common type of accident in India is a head-on collision. This type of accident is often caused by driver error, such as speeding or driving under the influence of alcohol or drugs. The authors also found that weather conditions, such as rain and fog, can contribute to road accidents. The research paper by Patila et al. provides a valuable contribution to the field of road safety in India. The authors' findings can be used to develop effective strategies to reduce the number of road accidents in the country.

It is important to note that these are just some of the potential causes of road accidents in India. The actual causes of any particular accident may vary depending on the specific circumstances. The use of data analysis to study road accidents is a promising new approach. Data analysis can be used to identify patterns in data that would be difficult or impossible to identify manually. This information can be used to develop effective strategies to prevent road accidents.

CHAPTER 3

SOFTWARE REQUIREMENTS AND SPECIFICATION

3.1 ASSUMPTIONS AND DEPENDENCIES

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. This means that the data should include all of the factors that can contribute to an accident, such as driver behavior, road conditions, weather and etc. The data should also be accurate and up to date.

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. We will need to find a balance between model complexity and accuracy. We want the model to be accurate, but we also want it to be easy 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 behavior, road conditions, weather, etc. should be included in the data. We will need to collect a large dataset of accident data. This data will be used to train the model to predict future accidents.

Evaluation data: The evaluation data should be separate from the training data. This data is used to evaluate the accuracy of the model. We will need to collect a separate dataset of accident data that was not used to train the model. This data will be used to test 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. We will need to develop a web app that allows users to input data about the current conditions and receive a prediction of the severity of an accident.

Bias: The model should be free of bias. The model should not be biased against any particular group of people. We will need to be careful to avoid bias in the data collection and training process. We will also need to monitor the model for bias once it is deployed.

Interpretability: The model should be interpretable. The model should be able to explain its predictions. We will need to be able to explain to users why the model made a particular prediction. This will help users to trust the model and use it to make informed decisions.

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

3.2 Project Scheduling

A comprehensive process that outlines the project phases, tasks under each stage, and dependencies is known as project scheduling. It also considers skills and the number of resources required for each task, their order of occurrence, milestones, interdependencies, and timeline. Furthermore, it involves analyzing the resource availability and implementing the scheduling technique to ascertain timely delivery while maintaining the resource health index. Many project managers successfully generate the right schedule, yet most of them find it challenging to manage the resources intelligently. It can cause delays and discrepancies in the deliverables as their talent pool is responsible for executing these tasks. Thus, they must master each aspect of project planning and scheduling.

Internal team conflicts are minimized when the entire team is on the same page. Resources are aware of the task dependencies and work diligently to ensure that the overall delivery is not affected. When teams opt for sophisticated project scheduling software, they get real-time updates on every project metric, which promotes proactive planning, monitoring, and coherent risk management.

3.3 Technology Used

3.3.1 Python

Python is a general-purpose programming language that is known for its simplicity, readability, and versatility. It is a powerful tool that can be used to build a wide variety of projects, from simple scripts to complex applications.

One of the things that makes Python so powerful for building projects is its large library of third-party modules. These modules provide access to a wide range of functionalities, such as data manipulation, web development, and machine learning. This means that you can start building projects without having to write all of the code from scratch.

As we are looking for a powerful and versatile programming language that can be used to build a wide variety of projects, Python is a great choice. With its simple syntax, large library of modules, and active community of developers, Python is a great tool for anyone who wants to build projects.

3.3.2 NumPy

NumPy is a Python library for scientific computing. It provides a high-performance multidimensional array object and tools for working with these arrays. NumPy is the fundamental package for scientific computing in Python.

NumPy can be used as a drop-in replacement for Numeric. It provides more powerful data structures and functions and is faster due to its use of BLAS and LAPACK. NumPy is also the core library for many other packages, such as SciPy and Pandas.

As we are working with numerical data in Python, then NumPy is an essential tool. It is a powerful and versatile library that can help you speed up your code and improve your results.

3.3.3 Google collab

Google Colaboratory[5], or Colab for short, is a free Jupyter Notebook environment that runs in the cloud. Colab allows you to write and execute Python code, and to access a wide range of pre-trained machine-learning models. Colab is a great tool for learning about machine learning, and for developing and testing machine learning models.

To use Colab, you first need to create a Google account. Once you have a Google account, you can open a Colab notebook by going to the Colab website. Colab notebooks are stored in your Google Drive account. You can easily share your Colab notebooks with co-workers or friends, allowing them to comment on your notebooks or even edit them.

Colab notebooks are a great way to learn about machine learning. Colab notebooks provide a simple and easy-to-use interface for writing and executing Python code. Colab notebooks also provide access to a wide range of pre-trained machine learning models. This makes it easy to get started with machine learning, and to develop and test machine learning models.

3.3.4 Scikit-learn

Scikit-learn is a free software machine-learning library for the Python programming language. It features various classification, regression, and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means, and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. Scikit-learn is a NumFOCUS fiscally sponsored project.

Scikit-learn is a popular machine-learning library in the Python community. It is used by a wide range of people, from students and researchers to data scientists and engineers. Scikit-learn is easy to use and has a wide range of features. It is also well-documented and has a large community of users and contributors.

Scikit-learn is a powerful tool for machine learning. It can be used to solve a wide range of problems, including classification, regression, clustering, and dimensionality reduction. Scikit-learn is also scalable and can be used to work with large datasets.

3.3.5 Streamlit

Streamlit is an open-source Python library that makes it easy to create web apps for data science[6]. With Streamlit, we can create beautiful, interactive apps in minutes, without having to write any JavaScript or HTML.

Streamlit is used by data scientists, engineers, and analysts to build and deploy data apps. Streamlit apps are used for a variety of purposes, including:

- Data exploration: Streamlit apps can be used to explore data and discover insights.
- Machine learning model building: Streamlit apps can be used to build and deploy machine learning models.
- Data visualization: Streamlit apps can be used to visualize data in a variety of ways.
- Dashboards: Streamlit apps can be used to create dashboards that provide a single view of data.

Streamlit is a powerful tool that can help you build data apps quickly and easily. If you are a data scientist, engineer, or analyst, then Streamlit is a tool that you should check out.

3.3.6 Jupyter Notebook

Jupyter Notebook is a web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. Jupyter Notebooks are a popular tool for data science, machine learning, and scientific computing.

Jupyter Notebooks are made up of cells. Each cell can contain either code or text. Code

cells can be run to execute the code they contain. The output of the code is displayed below the cell. Text cells can be used to add narrative text, equations, or visualizations to the notebook.

Jupyter Notebooks can be run in a web browser. This makes them very portable and easy to share. Jupyter Notebooks can also be run on a local machine. This can be useful for working with large datasets or for developing applications that require more computational power.

Jupyter Notebooks are a powerful tool for data science, machine learning, and scientific computing. They are easy to use and can be run in a web browser. This makes them a popular choice for a wide range of users.

3.4 Software and Hardware Requirements

3.4.1 Software Requirements:

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

3.4.2 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)

CHAPTER 4
SYSTEM DESIGN

4.1 The software development life cycle (SDLC)

The software development life cycle (SDLC) is a process that software developers use to plan, design, develop, test, and deploy software. The SDLC is divided into several phases, each of which has its own specific goals and tasks.

Planning: The planning phase is the first step in the SDLC. In this phase, you will define the scope of your project and create a project plan.

Requirements gathering: The requirement's gathering phase is where you will collect the requirements. This includes understanding the needs, the technical requirements, and the user requirements.

Design: The design phase is where you will create the blueprint for your software. This includes creating the system architecture, the user interface design, and the database design.

Development: The development phase is where you will build the software according to the design. This includes coding and debugging.

Testing: The testing phase is where you will test the software to ensure that it meets the requirements. This includes unit testing, integration testing, and system testing.

Deployment: The deployment phase is where you will make the software available to the users. This includes installing the software and providing support.

Maintenance: The maintenance phase is where you will fix bugs, add new features, and improve the software. This includes bug fixing, feature development, and performance tuning.

The SDLC is a cyclical process, and you may need to go back to previous phases if necessary. For example, if you discover that the requirements are not accurate, you will need to go back to the requirements gathering phase.

The SDLC is a valuable tool for managing software development projects. By following the SDLC, you can ensure that your project is successful.

4.2 Data Flow Diagram

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 modelling tools include the dataflow diagram.

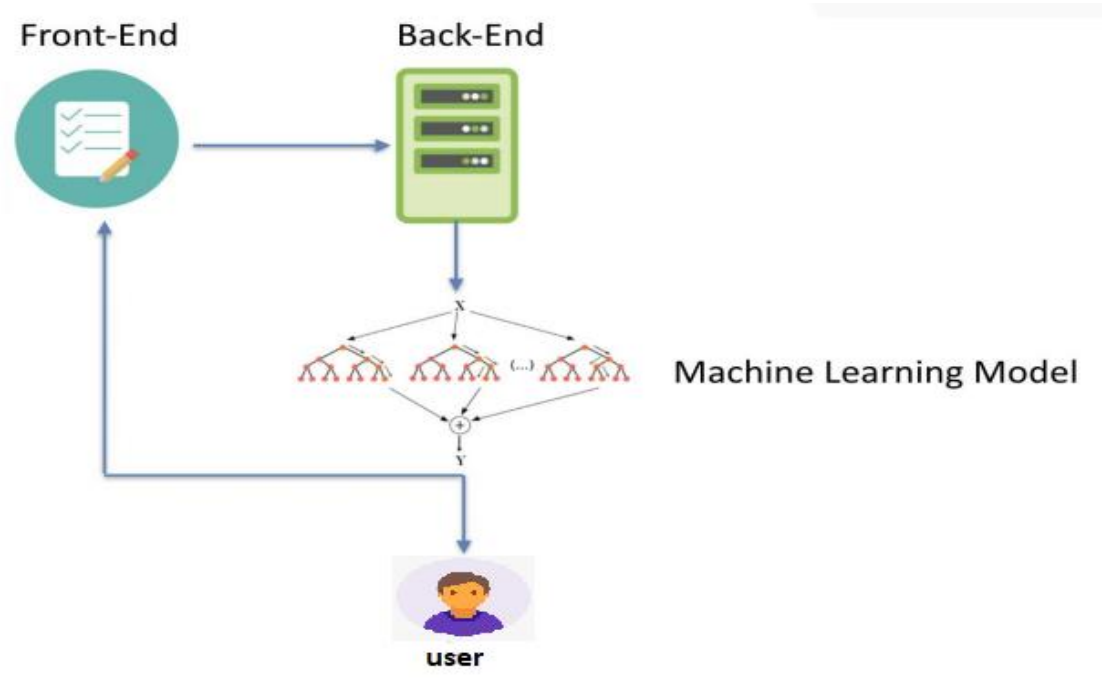

Figure 4.1: System model

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.

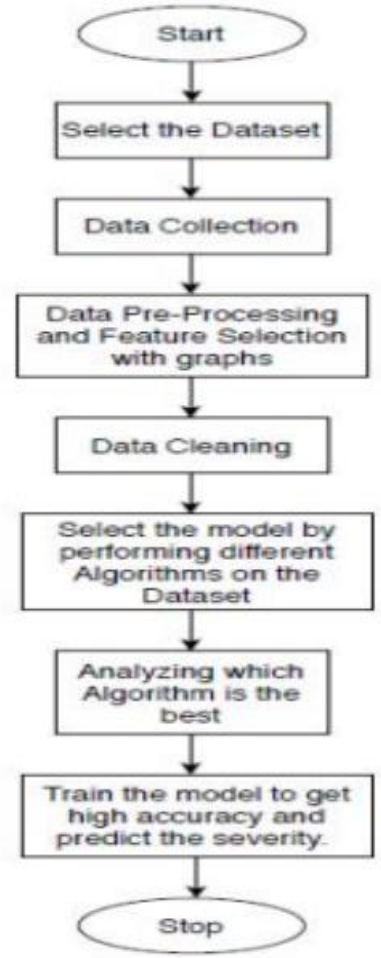

Figure 4.2: Data Flow Diagram

4.3 System Architecture

System architecture is the process of defining the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system. A system architecture can consist of system components and the sub-systems developed, that will work together to implement the overall system. There have been efforts to formalize languages to describe system architecture, collectively these are called architecture description languages (ADLs).

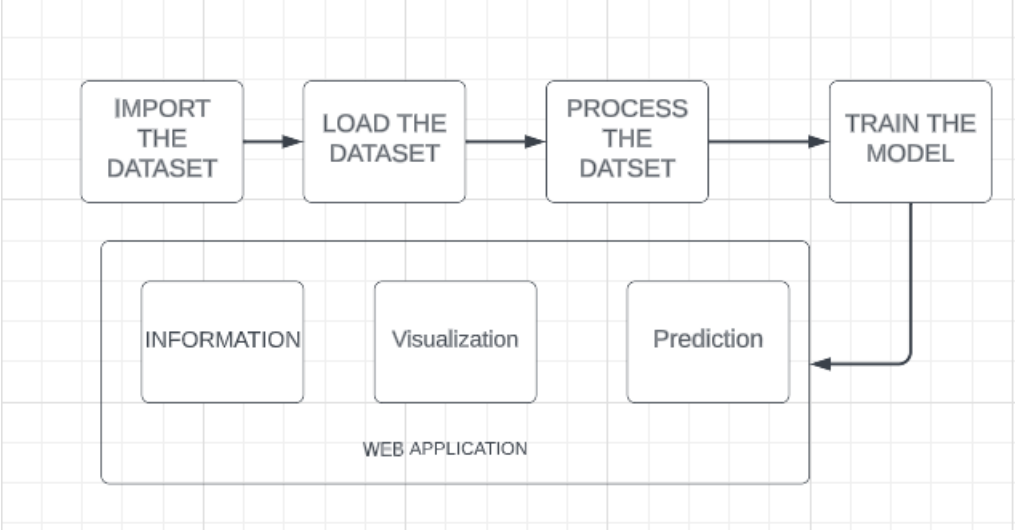

Figure 4.3: System Architecture

4.4 Use Case Diagram

A use case diagram is a graphical depiction of a user's possible interactions with a system. A use case diagram shows various use cases and different types of users the system has and will often be accompanied by other types of diagrams as well.

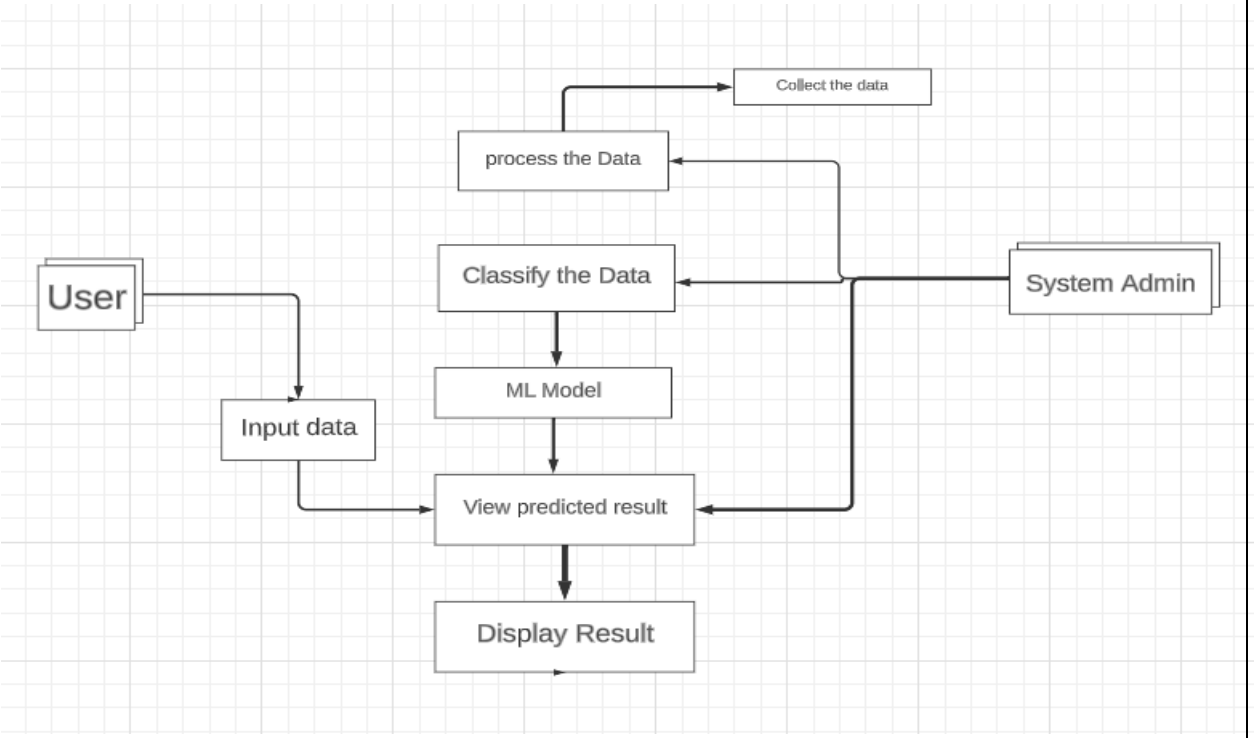

Figure 4.4: Use Case Diagram

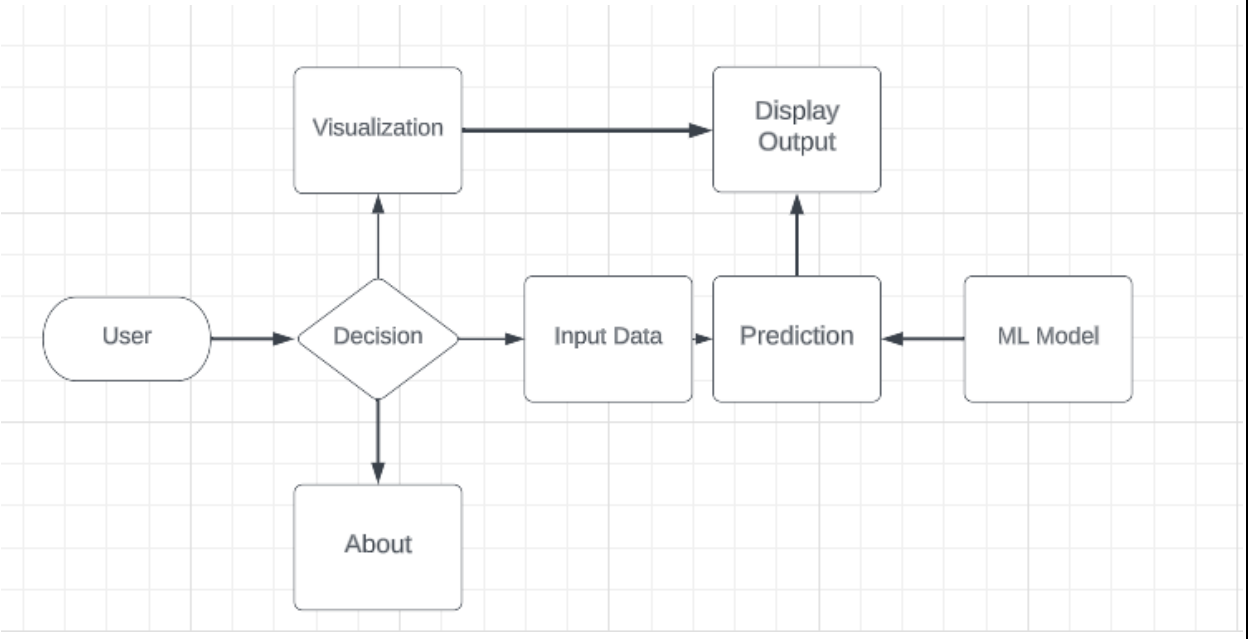

Figure 4.5: Web Application Interaction

CHAPTER 5
IMPLEMENTATION

5.1 Implementation

We are now in the implementation phase of our project. This is where we take the ideas that we have developed and put them into action. We will be working closely with our team to make sure that the project is implemented smoothly and that all of the requirements are met. We are excited to see this project come to life and we are confident that it will be a success.

Here are some of the things that we will be doing in the implementation phase:

- Developing a detailed plan. We will need to develop a detailed plan that outlines the steps that need to be taken in order to implement the project. This plan will be essential for ensuring that the project stays on track.
- Communicating with Teammates. We will need to communicate regularly with our teammates to keep them updated on the progress of the project. This will help to ensure that everyone is on the same page and that there are no surprises.
- Monitoring progress. We will need to monitor the progress of the project closely in order to identify any potential problems early on. This will help us to take corrective action before the problems become serious.
- Making adjustments as needed. As the project progresses, we may need to adjust the plan. This is perfectly normal and should be expected. The important thing is to be flexible and adaptable.

We have divided the implementation process into four phases:

- **Phase 1:** In this phase, we will gather all of the necessary information and data for our project. We will then clean the data to remove any errors or inconsistencies.
- **Phase 2:** (Data visualization, findings, and feature selection and model selection) In this phase, we will visualize the data to gain insights and identify patterns. We will then select the features that are most important for our model and select the best model for our data.
- **Phase 3: (**Training and evaluation) In this phase, we will train the model on the data and evaluate its performance. We will make adjustments to the model as

needed until we are satisfied with its performance.

- **Phase 4:** (Web application deployment and maintenance) In this phase, we will deploy the model to a web application and maintain it. We will also monitor the performance of the model and make adjustments as needed.

5.2 Phase 1

In this phase, we will gather all of the necessary information and data for our project. We will then clean the data to remove any errors or inconsistencies. This phase contains planning and analysis of information Gathering, Data Gathering, and Data Preprocessing

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 [7]. 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 [8]. 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.

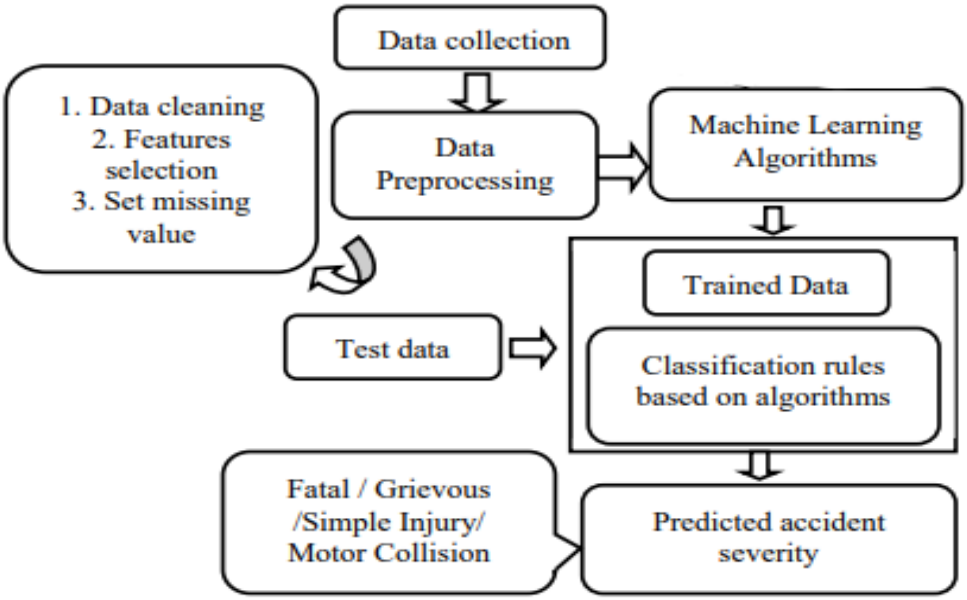

Figure 5.1: Model Pipeline Process

5.3 Phase 2

In this phase, we will visualize the data to gain insights and identify patterns. We will then select the features that are most important for our model and select the best model for our data.

Data Visualization:

The visual representation of data is known as data visualization [9]. 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:

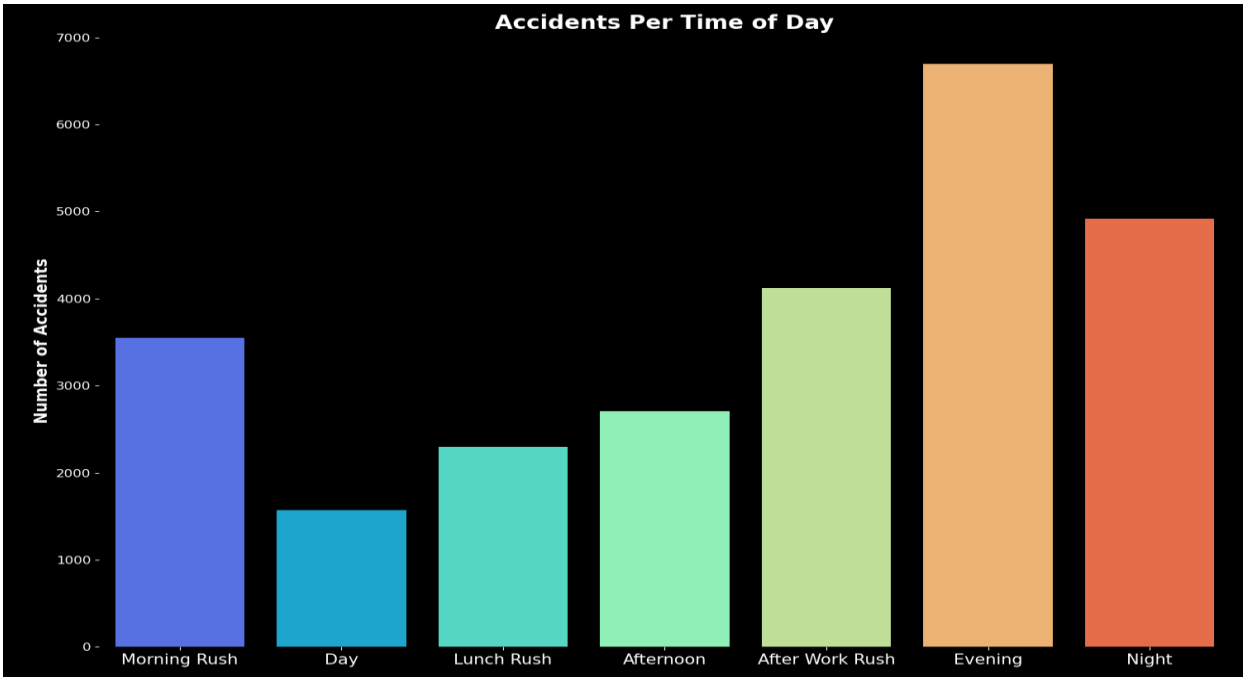

FIGURE 5.2: Accidents at Russ hours of the day

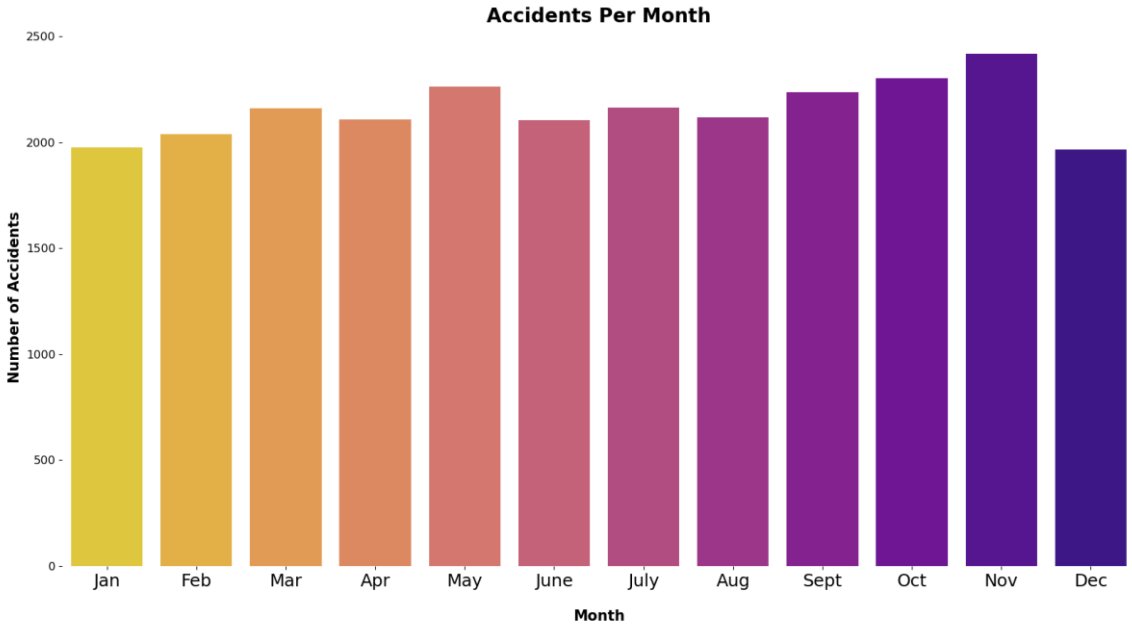

Figure 5.3: Accidents per month

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.

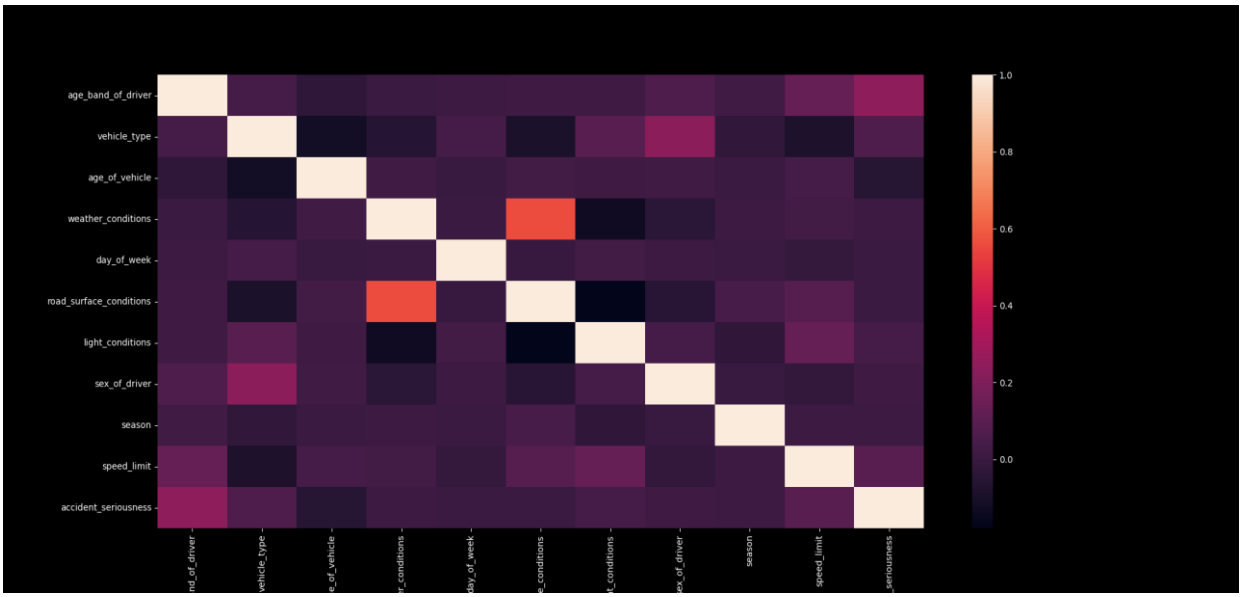

Figure 5.4: Heat Map od Selected Features

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** is an ensemble model that combines multiple decision trees to make predictions. It is a powerful model that can be used for a variety of tasks, including classification, regression, and clustering.

- **Support vector machine (SVM)** is a supervised learning model that can be used for classification and regression tasks. It works by finding a hyperplane that separates the data points into two classes.

- **XGBoost** is a gradient boosting algorithm that is known for its speed and accuracy. It is a popular choice for large-scale machine learning problems.

- **Gradient boosting classifier** is a supervised learning algorithm that can be used for classification tasks. It works by building a series of weak learners and then combining them to create a strong learner.

- **K-nearest neighbors (KNN)** is a non-parametric algorithm that can be used for both classification and regression tasks. It works by finding the K most similar data points to a new data point and then using those points to make a prediction.

- **Decision tree** classifier is a supervised learning algorithm that can be used for classification tasks. It works by creating a tree-like structure that represents the relationships between the features and the target variable.

- **Logistic regression** is a supervised learning algorithm that can be used for classification tasks. It works by fitting a line or curve to the data that separates the two classes.

These are just a few of the many models that can be used for large datasets. The best model for a particular problem will depend on the specific characteristics of the data and the desired outcome.

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.

5.4 Phase 3

In this phase, we will train the model on the data and evaluate its performance. We will make adjustments to the model as needed until we are satisfied with its performance. We will also select the best model from the models that we have selected and try to evaluate the model.

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	82.65%
Gradient Boosting Classifier	84.50%
SVC (Support Vector Machine)	82.17%
K-Neighbors Classifier	78.81%
Logistic Regression	82.90%
Decision Tree Classifier	84.26%
XGBoost	83.64%

Table 5.1: Accuracy of Different Models

We found that the random forest model [10] 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.

```
In [ ]: from sklearn.model_selection import GridSearchCV
        # Define the parameter grid to search over
        param_grid = {
            'n_estimators': [100, 500, 1000,2000],
            'max_features': ['auto', 'sqrt', 'log2', None],
            'max_depth': [10, 20, 30, 40, None],
            'min_samples_split': [2, 5, 10, 15],
            'min_samples_leaf': [1, 2, 4, 8],
            'bootstrap': [True, False]
        }
        rf = RandomForestClassifier(random_state=50) # Create a random forest classifier object
        grid_search = GridSearchCV(rf, param_grid, cv=5, n_jobs=-1)  # Create a grid search object
        grid_search.fit(X_train, y_train)   # Fit the grid search object to the training data
        best_model = grid_search.best_estimator_ # Get the best model from the grid search
        filename = 'random_forest_model.pkl'                  # Save the best model to a file
        with open(filename, 'wb') as file:
            pickle.dump(best_model, file)

        # Use the best model to make predictions on the test set
        y_pred = best_model.predict(X_test)
        acc1=accuracy_score(y_test, y_pred) # Evaluate the model performance
        print('Accuracy:', acc1)
        print('Confusion Matrix:\n', confusion_matrix(y_test, y_pred))
        print('Classification Report:\n', classification_report(y_test, y_pred))
        print(type(rf).__name__, "accuracy:", acc1, "with best hyperparameters:", grid_search.best_params_)
        cm = confusion_matrix(y_test, y_pred)
```

Figure 5.5: Random Forest Model

The above code is used to train a random forest classifier and evaluate its performance. The code first defines a parameter grid that specifies the values of the hyperparameters that will be searched over. The code then creates a random forest classifier object and a grid search object. The grid search object is then fit to the training data. The best model from the grid search is then saved to a file. The best model is then used to make predictions on the test set. The performance of the model is then evaluated using accuracy, confusion matrix, and classification report.

Hyperparameter tuning [11] is the process of finding the best values for the hyperparameters of a machine learning model. Hyperparameters are the settings of a model that are not learned from the data. They can have a significant impact on the performance of the model, so it is important to choose them carefully.

Grid search is a brute-force method for hyperparameter tuning. It involves creating a grid of all possible combinations of hyperparameter values and then evaluating each combination on a holdout dataset. The combination with the best performance is then chosen as the best hyperparameters for the model.

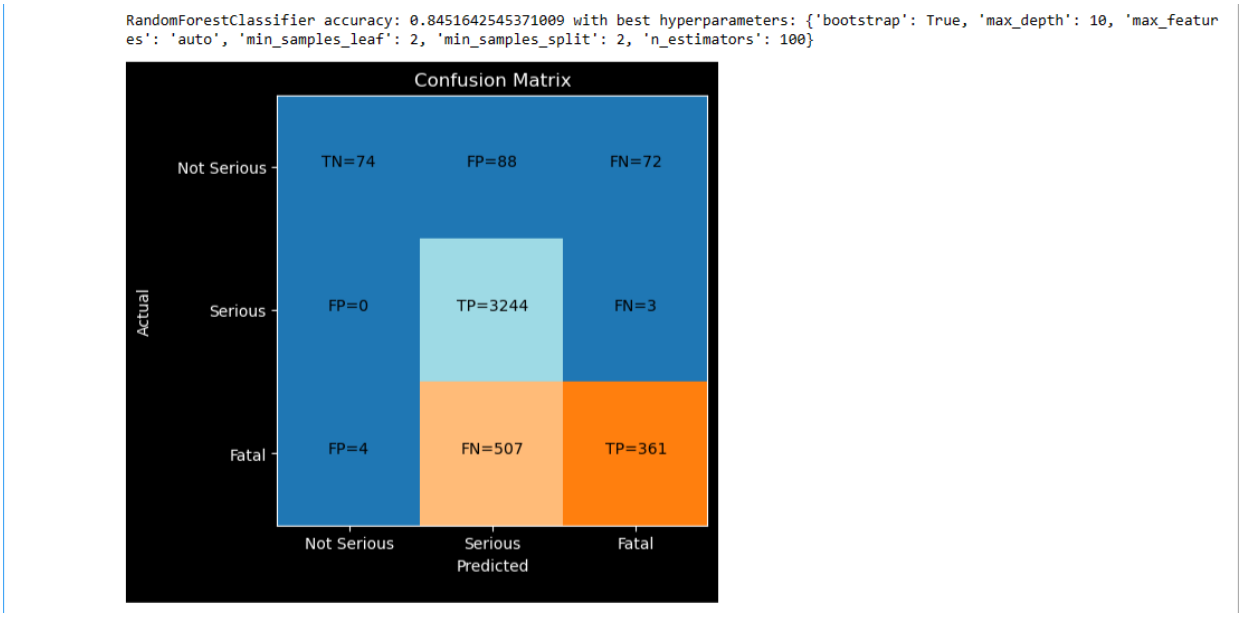

Figure 5.6: Confusion Matrix

The figure above shows the best parameters and confusion matrix for a random forest classifier. The best parameters are n_estimators=100, max_features='auto', max_depth=10, min_samples_split=2, min_samples_leaf=2, and bootstrap=True. The confusion matrix shows that the model correctly classified 84.51% of the test set samples. The best parameters were found using a grid search. The grid search searched over a range of hyperparameter values and found the values that resulted in the best model performance. The confusion matrix was created by using the best model to make predictions on the test set and then comparing the predictions to the actual labels.

The results of the grid search and the confusion matrix show that the random forest classifier is a good model for this task. The model has high accuracy and low false positive and false negative rates. This means that the model is good at correctly classifying both positive and negative samples.

5.5 Phase 4

In this phase, we will deploy the model to a web application and maintain it. We will also monitor the performance of the model and make adjustments as needed.

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. Multi-page Streamlit applications are created by organizing the code into separate files, each of which corresponds to a different page. When a user clicks on a page in the navigation bar, the corresponding code is executed, and the page is displayed. The pages are then displayed in a navigation bar on the left side of the 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. The main page accepts the input from the user data and displays the predicted result. The main page also contains a navigation bar for the Home, Visual, and About pages.

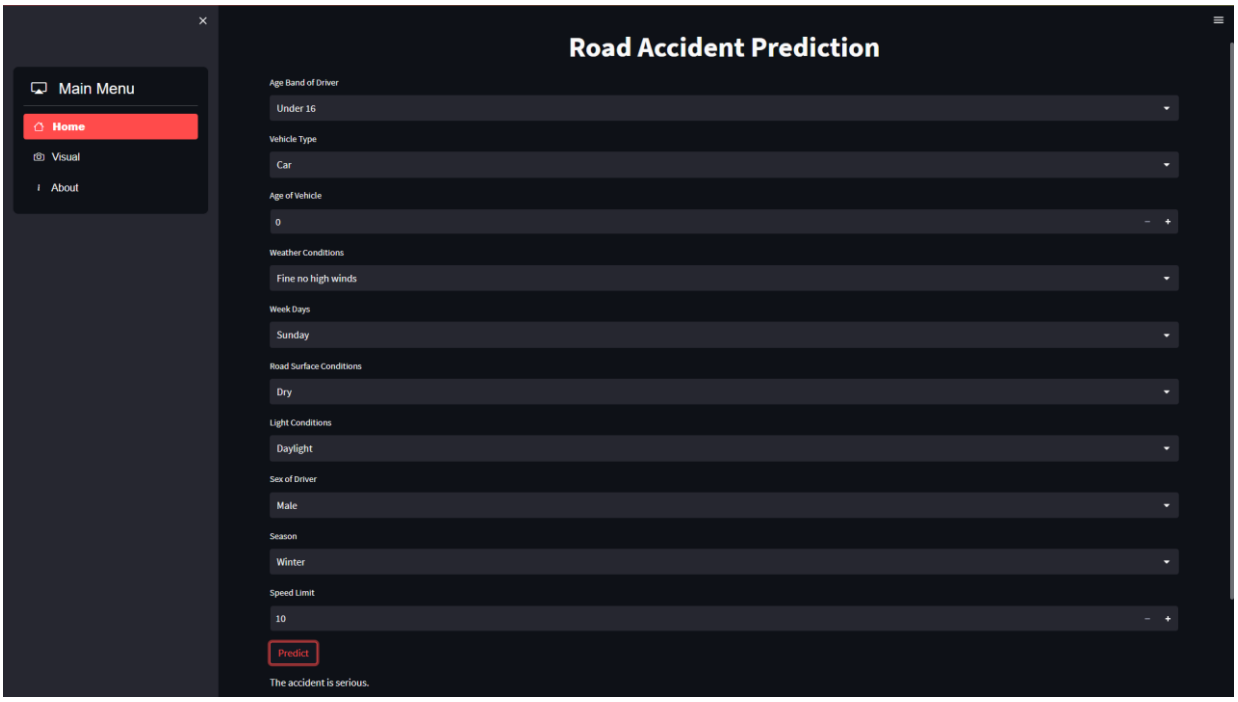

Figure 5.7: A Screenshot Of The Home Web-Page

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.

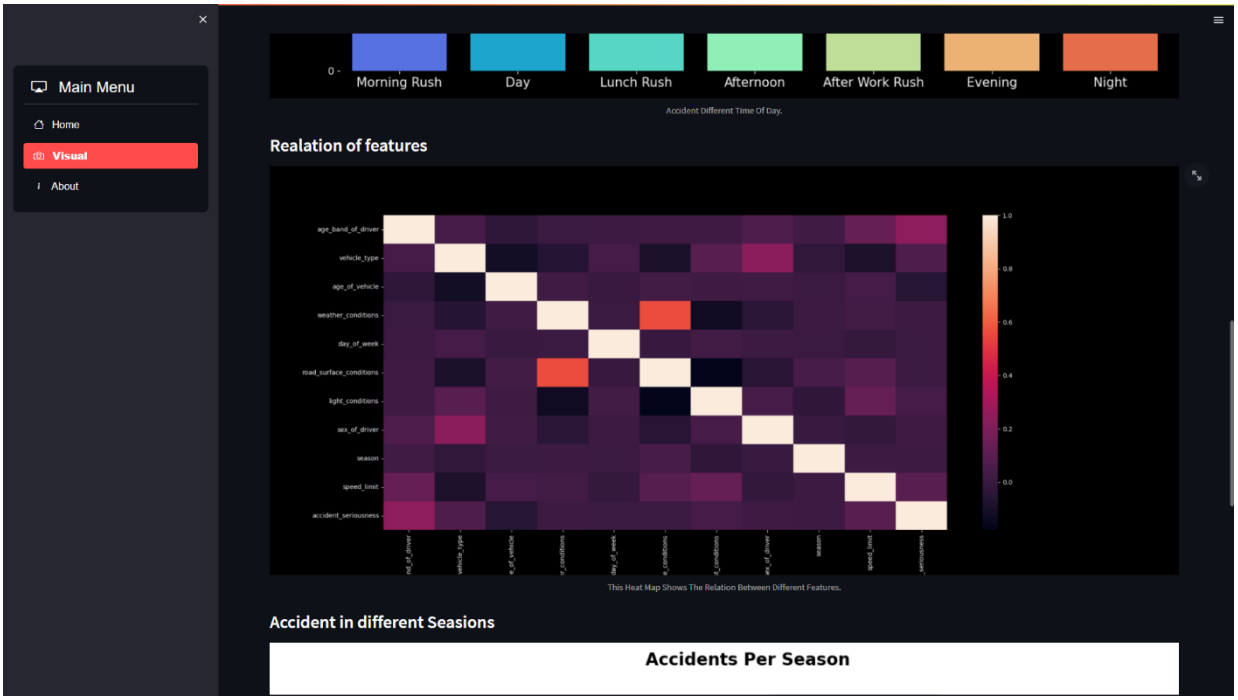

Figure 5.8: A Screenshot of The Visual Web-Page

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.

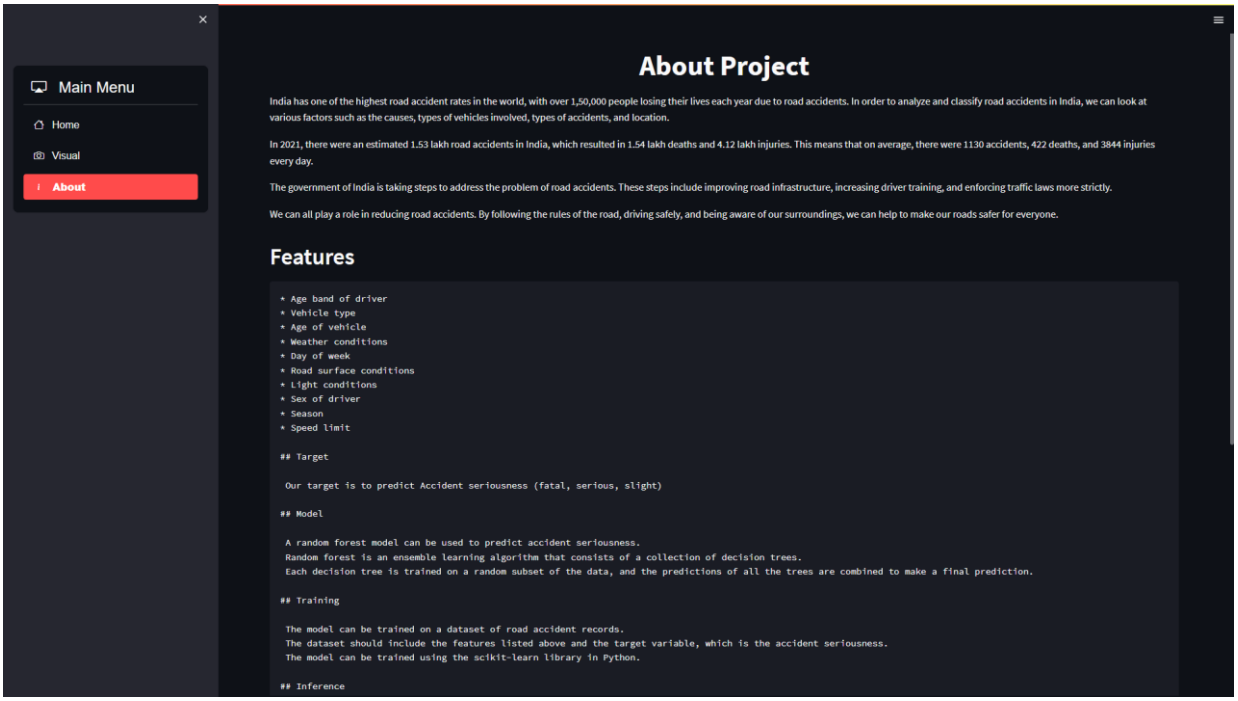

Figure 5.9: A Screenshot of The About Web-Page

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.

CHAPTER 6
RISK ANALYSIS

6.1 Risk Analysis Introduction

Risk analysis is the process of identifying and analyzing potential issues that could negatively impact the algorithm or project [12]. This process is done in order to help organizations avoid or mitigate those risks. Performing a risk analysis includes considering the possibility of adverse events caused by malicious or inadvertent human activities. An important part of risk analysis is identifying the potential for harm from these events, as well as the likelihood that they will occur.

Organizations must understand the risks associated with the use of their information systems to protect their information assets effectively and efficiently. Risk analysis can help an organization improve its security in a number of ways. Depending on the type and extent of the risk analysis, organizations can use the results to help.

6.2 Steps of Risk Analysis

1 Conduct a risk assessment survey: This first step, getting input from management and department heads, is critical to the risk assessment process. The risk assessment survey is a way to begin documenting specific risks or threats within each department.

2 Identify the risks: The reason for performing a risk assessment is to evaluate an IT system or other aspect of the organization and then ask: What are the risks to the software, hardware, data, and IT employees? What are the possible adverse events that could occur, such as human error, fire, flooding, or earthquakes? What is the potential that the integrity of the system will be compromised or that it won't be available?

3 Analyze the risks: Once the risks are identified, the risk analysis process should determine the likelihood that each risk will occur, as well as the consequences linked to each risk and how they might affect the objectives of a project.

4 Analysis a risk management plan: Based on an analysis of which assets are valuable and which threats will probably affect those assets negatively, the risk analysis should produce control recommendations that can be used to mitigate, transfer, accept or avoid the risk.

5 Implement the risk management plan: The ultimate goal of risk assessment is to implement measures to remove or reduce the risks. Starting with the highest-priority risk, resolve or at least mitigate each risk so it's no longer a threat.

6 Monitor the risks: The ongoing process of identifying, treating and managing risks should be an important part of any risk analysis process. The focus of the analysis, as well as the format of the results, will vary depending on the type of risk analysis being carried out.

6.3 Technical Feasibility

The technical feasibility of a project is performing a check on whether the development of the project is possible with the available technological resources. Technical feasibility is a very important aspect to be considered before the official commencement of the project by the organization. The technical feasibility is checked by pondering over the functional requirements of the user. To determine whether the proposed system is technically feasible, the technical issues involved in the system should be taken into consideration. The proposed system uses Python technology. Python is an open-source technology, it is available for free of cost and convenience. As far as the platform for the project is concerned, it is decided to perform the project on the window OS.

6.4 Real-World Risks

Real-world Risks are those risks that can affect property or human life. In many fields, Artificial intelligence and machine learning algorithms are making a decision for many organizations for different purposes. When this decision is made wrong the organization suffers financial/physical losses and this can happen due to faulty data or human mistake or coding error or glitches. Because of the inaccurate data, machine learning algorithms could generate inaccurate results which could lead to faulty decision-making in the program.

In road accident analysis if our program gives a faulty output or inaccurate output it may cause property or human life damage. Therefore, risk analysis and risk management are one of the most important factors to consider.

6.5 Conclusion of 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.

CHAPTER 7
RESULT

8.1 Screenshot

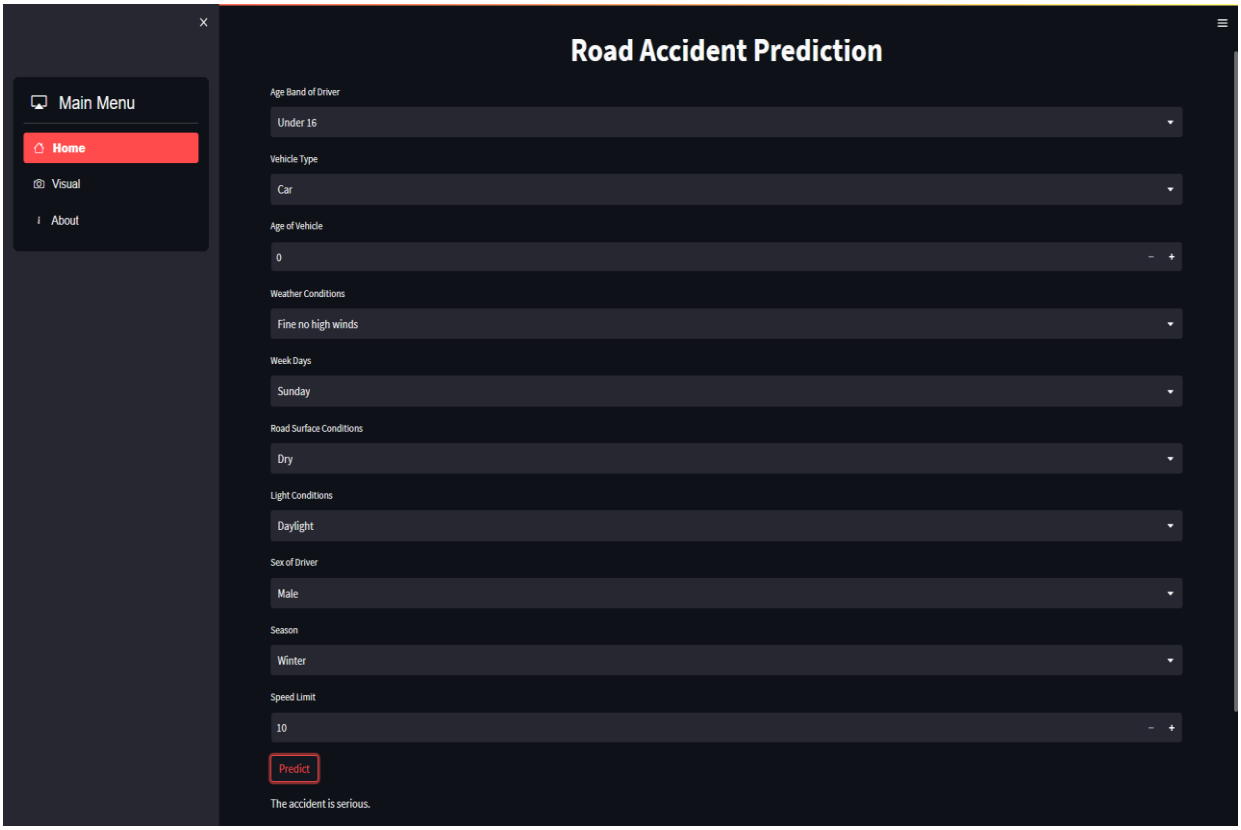

Figure 7.1: Home page with Serious Prediction

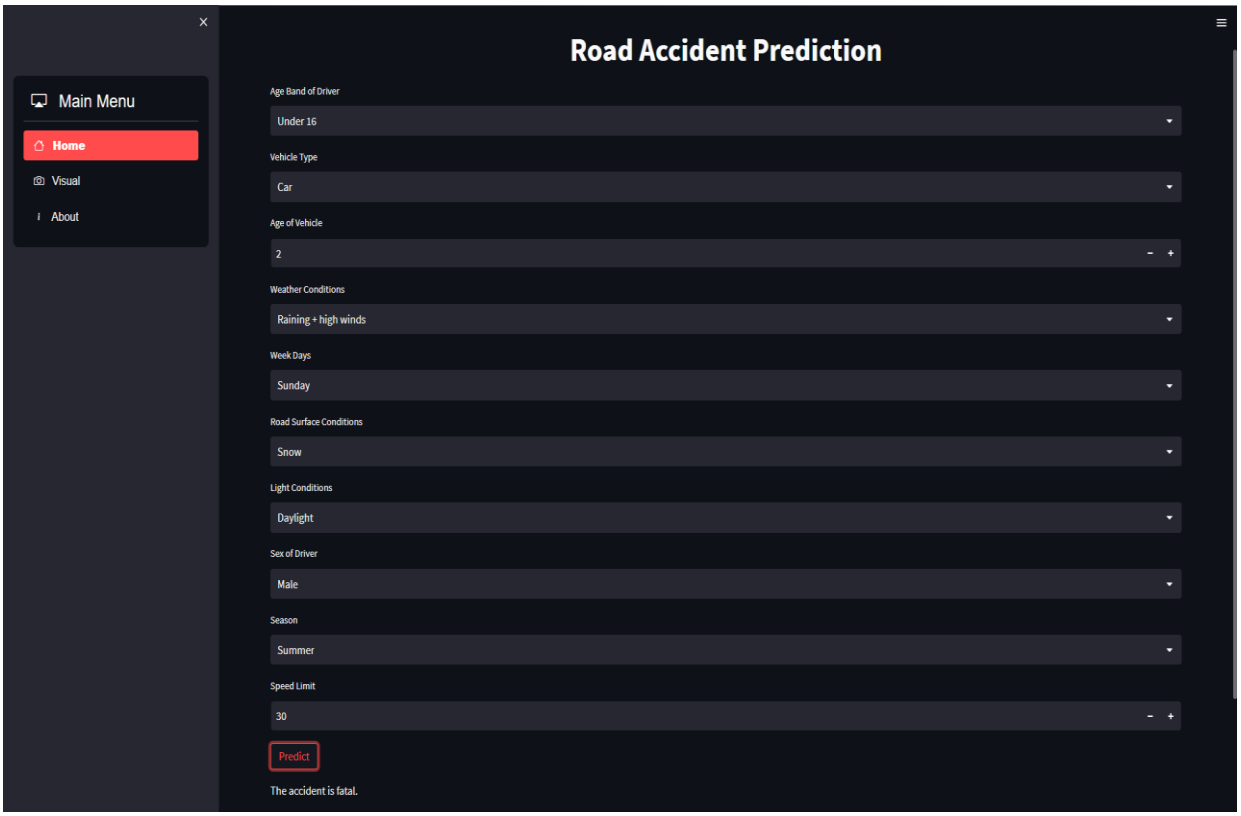

Figure 7.2: Home page with Fatal Prediction

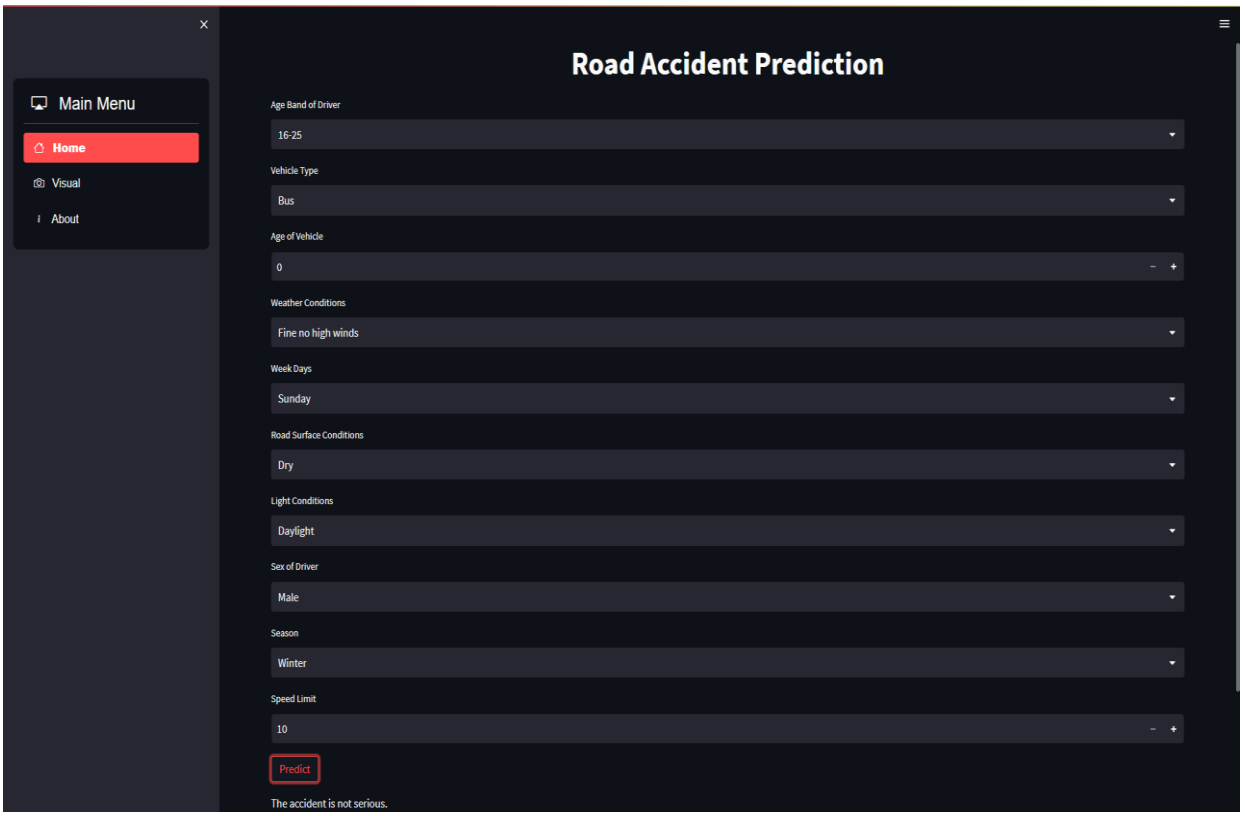

Figure 7.3: Home page with Not-Serious Prediction

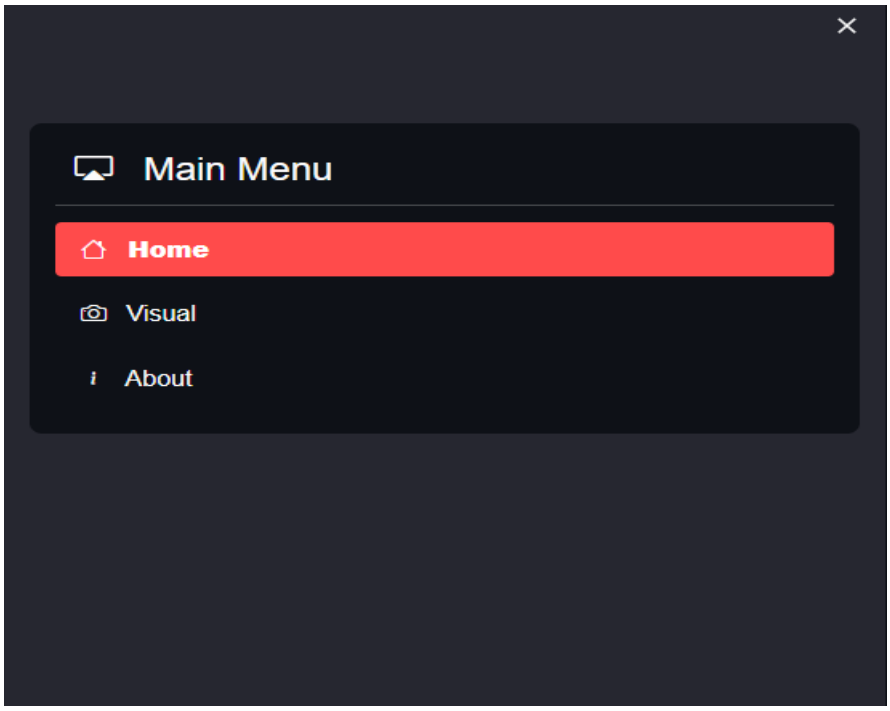

Figure 7.4: Home page Navagation Bar

Figure 7.5: Visual page part-1

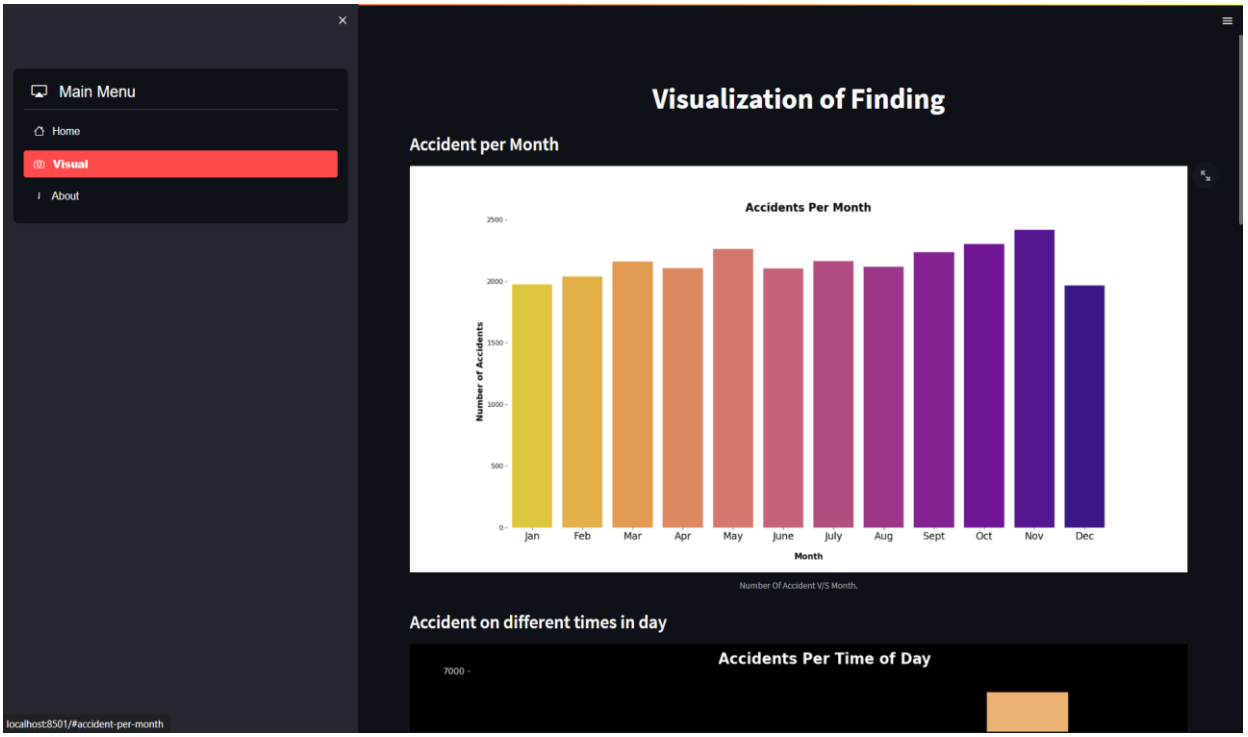

Figure 7.6: Visual page screenshot-2

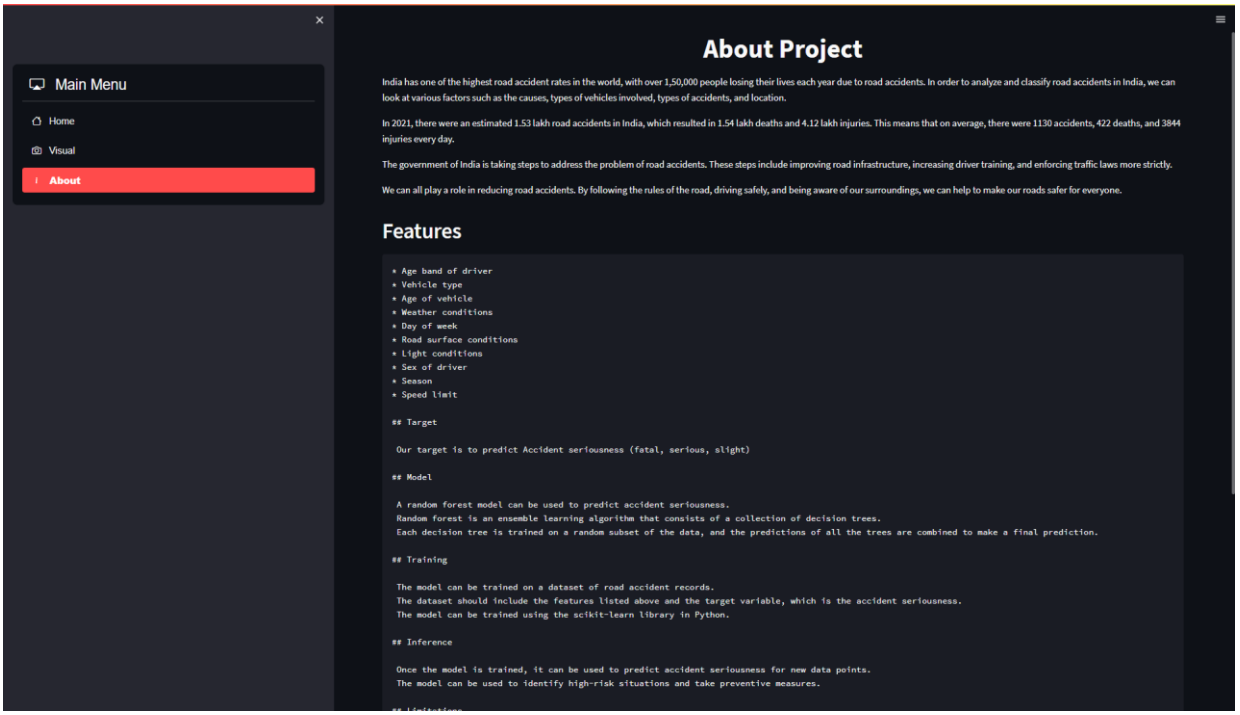

Figure 7.7: About Page

CHAPTER 8
CONCLUSION

8.1 Final 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.

8.2 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 Automobile 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.

8.3 Advantages

- 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.
- 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.
- 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.

- 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.

8.4 Limitations

- 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.
- 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.
- 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.
- 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.

8.5 Challenges

- 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.
- Data cleaning: The data that we collect will be messy and incomplete. We need to clean the data and remove any errors or inconsistencies.
- 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.

- 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.
- 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.

8.6 Applications

- Traffic management: The application can be used to predict traffic accidents and their severity. This information can be used to improve traffic management and reduce the number of accidents.
- Insurance: The application can be used to predict the likelihood of an accident and the severity of the injuries. This information can be used to set insurance premiums and provide better coverage to customers.
- Public safety: The application can be used to identify areas where accidents are likely to occur. This information can be used to improve public safety by increasing the presence of law enforcement and emergency services in these areas.
- Research: The application can be used to collect data on accidents and their severity. This data can be used to improve the understanding of accidents and develop new ways to prevent them.
- The application can also be used to improve the safety of automatic vehicles. By predicting the likelihood of an accident, the application can help to prevent accidents and injuries. The application can also be used to improve the performance of automatic vehicles by identifying areas where they are more likely to encounter accidents.

REFERENCES:

1. Road accident in India 2021 published by the Ministry of Road Transport and Highways (MoRTH).

2. 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.

3. 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.

4. 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.

5. https://colab.research.google.com/

6. Streamlit: https://docs.streamlit.io/

7. https://www.kaggle.com/datasets/data125661/india-road-accident-analysis-and-classification

8. Data Preprocessing: https://scikit-learn.org/stable/modules/preprocessing.html

9. Visualization: https://seaborn.pydata.org/

10. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

11. https://www.anyscale.com/blog/what-is-hyperparameter-tuning

12. https://www.pmi.org/learning/library/risk-analysis-project-management-7070

DOI: 10.55041/IJSREM21592

ISSN: 2582-3930

Impact Factor: 8.176

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING & MANAGEMENT

An Open Access Scholarly Journal || Index in major Databases & Metadata

CERTIFICATE OF PUBLICATION

International Journal of Scientific Research in Engineering & Management is hereby awarding this certificate to

Shivansh Gautam

in recognizaton to the publication of paper titled

ROAD ACCIDENT ANALYSIS USING MACHINE LEARNING

published in IJSREM Journal on Volume 07 Issue 05 May, 2023

www.ijsrem.com

Editor-in-Chief
IJSREM Journal

ijsremjournal@gmail.com