**Software Requirement Specification**

**1.1Purpose:**

To analyse traffic accidents in New York City and identify key trends and patterns in the data. The project aims to provide insights that can help policymakers, transportation planners, and other stakeholders improve road safety and reduce the number of accidents and injuries in the city. By analysing factors such as population density, car ownership, and time of day, the project seeks to shed light on the causes of accidents and suggest ways to prevent them. Ultimately, the goal of the project is to contribute to a safer and more sustainable transportation system in New York City.

**1.2 Definitions, Acronyms, and abbreviations Acronyms**

Definitions:

* Population density: the number of people living in a given area, usually measured as the number of people per square mile or square kilometre.
* Crash density: the number of crashes per unit of area, usually measured as the number of crashes per square mile or square kilometre.
* Car ownership: the percentage of households or individuals who own a car or other motor vehicle.
* Commute: the act of travelling to and from work or school.
* Peak travel hours: the times of day when the most people are commuting, typically the morning and evening rush hours.
* Driver inattention/distraction: a factor that contributes to many traffic accidents, often caused by using electronic devices, eating, or other distractions while driving.
* Fatigued/drowsy driving: a common cause of accidents, particularly during long commutes or driving late at night.

Acronyms and abbreviations:

* NYC: New York City
* SUV: Sports Utility Vehicle
* km/h: kilometers per hour
* mph: miles per hour
* %: percent (or percentage)
* EDA: Exploratory Data Analysis
* LR: Logistic Regression
* SVC: Support Vector Classifier
* GB: Gradient Boosting

**1.3 References for Requirement Analysis and Design:**

* "Software Requirements" by Karl Wiegers and Joy Beatty
* "Applying UML and Patterns" by Craig Larman
* "Design Patterns: Elements of Reusable Object-Oriented Software" by Erich Gamma
* "Clean Code: A Handbook of Agile Software Craftsmanship" by Robert C. Martin

**1.4 Overview**

The project is aimed at analyzing and visualizing data related to road accidents in New York City. The analysis includes examining the factors that contribute to accidents, the types of vehicles involved, the time and location of accidents, and the severity of injuries and fatalities. The project aims to provide insights into the causes of accidents and to identify areas for improvement in road safety measures. The project will involve collecting and cleaning data from various sources, including the New York City Open Data Portal The data will then be analysed and visualized using various tools and techniques such as Python, Pandas, and Matplotlib. The project will also include the development of a web page that will provide interactive visualizations and allow to explore the data and gain insights into the factors contributing to road accidents in New York City. Overall, the project aims to provide a comprehensive analysis of road accidents in New York City and to use this analysis to inform policy decisions and improve road safety measures in the city.

List Of Functions:

| S.no | Name | Function |
| --- | --- | --- |
| 1 | Extract Weather Data | Web Scraping Weather Data |
| 2 | Extract Crash Data | Loading Crash Data |
| 3 | EDA of Weather Data | Analysing and finding trends in data |
| 4 | EDA of Crash Data | Analysing and finding trends in data |
| 5 | Merge data | Merge on basis of date and timestamp |
| 6 | Insertion in database | Pushing data into MongoDB |
| 7 | Machine Learning | Applying LR, GB classifiers |
| 8 | Visualisation | Creating html page with Power BI |

**2 Functional Description**

* Extract Weather Data: This function involves web scraping from the weather underground website using python and extracting and formatting data from JSON to CSV for the year 2016 to 2023 using BeautifulSoup. Beautiful Soup is a Python package for parsing HTML and XML documents (including having malformed markup, i.e. non-closed tags, so named after tag soup). It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping.
* Extract Crash Data: This function involves the collection of data related to traffic accidents in the New York City NYC Open Data website from the year 2012 to 2023.
* EDA of Weather Data: The data needs to be analysed to identify patterns, trends, and insights related to weather attributes such as wind speed , visibility, precipitation rate, condition, temperature etc.
* EDA of Crash Data: The data needs to be analysed to identify patterns, trends, and insights related to traffic accidents in New York City with respect to injuries, kills, type of vehicle ,region, time of the day etc
* Merge data: This function deals with creating a single column by combining date and timestamp to create a unique key in both the data set and performing a join by using PySpark. PySpark is the Python API for Apache Spark, an open source, distributed computing framework and set of libraries for real-time, large-scale data processing
* Insertion in database: This function deals with creating the clusters of MongoDB Atlas, the Cloud-Native Document Database as a Service for Easiest Way to Deploy, Operate, and Scale MongoDB in the Cloud .Atlas simplifies deploying and managing your databases while offering the versatility you need to build resilient and performant global applications on the cloud providers of your choice
* Machine Learning: This function deals with creating the target variable “severity level” , applying machine learning technique to do classification tasks based on created target variable. The formula for severity level is calculated on the basis of “['NUMBER OF MOTORIST KILLED']\*3 +['NUMBER OF MOTORIST INJURED']\*1” resulting in severity levels of low, medium, high, very high
* Visualisation: The analysed data needs to be presented in a visually appealing way using graphs, charts, and maps to make it easier for users to understand and draw insights from the data using a HTML page with PowerBI.

**3 Specific Requirements**

**3.1 Software interfaces:**

* Python
* matplotlib
* scikit-learn
* pandas
* jupyterlab
* pymongo
* chrome
* HTML
* PySpark
* Power BI
* MongoDB

**3.2 Hardware Interfaces**

* 8GB RAM
* 512GB HDD
* Intel core i5 processor
* Intel Graphics card