

**SOFTWARE REQUIREMENTS SPECIFICATION**

**Data analysis of road traffic accidents and to minimize and prevent the rate of accidents**

**Project Guide:** Prof. Chetan Shetty

**Team Leader:**Prashant Krishnan V

**Team Details:**

|  |  |  |  |
| --- | --- | --- | --- |
| **SL No.** | **USN** | **NAME** | **SIGNATURE** |
| 1 | 1MS15CS074 | M V S VISWANADH |  |
| 2 | 1MS15CS091 | PRASHANT KRISHNAN V |  |
| 3 | 1MS15CS143 | VIVEK CHANDRA SHEEL |  |

1. **Product Overview**

A dataset containing data based on accidents in a particular area has been obtained and studied. The dataset has close to 4.8 million data points and consists of attributes such as the longitude and latitude; showing the date, time, day of the week, location of the accident, the type of accident, the condition of light, severity, speed zone, whether it was caused due to the consumption of alcohol, whether it is a hit and run case, the type of pedestrians involved in the accident and so on. The rate at which such accidents are occurring and the patterns in the attributes that govern these occurrences are being studied in order to prevent such circumstances to avoid accidents.

This is study that involves using Machine Learning algorithms in order to analyse the given dataset and derive a time series analysis and forecast certain conditions of accidents that will help prevent them in the future. It focuses on finding out the reason underlying these frequent accidents. This project involves making a model in order to carry out this task to make predictions and graphs that will help us to do so.

1. **Hardware Requirements**

* 32- or 64-bit Computer
* Operating System: Windows / Linux / MacOS
* Minimum 3GB of disk space for download and install
* RAM: Minimum of 4GB
* Processors: Intel® Core™ i3 processor onwards
* Python 3.6.x
* Pandas, Numpy, Scikit-Learn and Matplotlib libraries

1. **Software Requirements**

|  |  |
| --- | --- |
| **Software Used** | **Description** |
| 1. Anaconda | Anaconda is a freemium open source distribution of the Python and R programming languages for large-scale data processing, predictive analytics, and scientific computing, that aims to simplify package management and deployment. |
| 1. Jupyter Notebook | The Jupyter Notebook is an open-source web application that allows to create and share documents that contain live code, equations, visualizations and narrative text.  Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more. |
| 1. Spyder | Spyder is an open source cross-platform integrated development environment for scientific programming in the Python language |
| 1. RStudio | RStudio is a free and open-source integrated development environment for R, a programming language for statistical computing and graphics. |
| 1. Scikit-learn Library | 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. |

1. **Functional Requirements**
   1. **Alcohol consumption on a day-wise basis**

We’re using data analytics tools and various graphs to find out the if there are relations with accidents in which alcohol consumption is involved with respect to the day. This would help the traffic police to probably use extra safety measures on specific days of the week.

* 1. **Severity of accident with respect to light conditions and alcohol consumption**

We use modules and a decision tree to find out how light conditions and alcohol consumption as independent factors affect the severity of accident. It is often seen that light conditions in area and with cases in which alcohol consumption is involved, the accidents can be more severe. This analysis can help us reduce the amount of accidents that are highly severe.

* 1. **Severity of accident with respect to day of the week, light conditions and speed zone**

A decision tree helps us to tell us how the day of the week, light conditions and speed zone as independent factors affect the severity of accident. The speed zones of areas often attribute to how severe accidents can be. When factors like light conditions and day of the week are considered, we see that drivers can be reckless in certain areas and/or on certain days. The information extracted from this analysis would help in taking proactive measures.

* 1. **Number of deaths due to accidents with respect to city**

Clustering is used to relate accidents in which deaths are involved to the city in which it happened. This study can extend from this dataset as to why there are high rates of fatalities in accidents involving certain cities. There can be stricter traffic laws in these cities to ensure higher degree of safety.

* 1. **Time Series Analysis and Forecasting**

Time series data are data that are indexed by a sequence of dates or times. The data formatted in such a way helps us analyze accidents on the basis of the dates on which they’ve occurred. Time series also helps us to categorize these accidents into them occurring in different times of the day. This lets us know at how the police should be dealing with situations in different times of the day. This concept is also extended to predict when there is a high possibility of accidents occurring**.**

1. **Non-Functional Requirements**

* **Handling Data**

The data requirements are very simple and direct. The data is to be supplied to the model in the form of a dataset which will be worked upon using the Pandas library. A database may be used to retrieve the data in which case we may using an SQL based framework, however since this is a purely machine learning-based project we may use data available directly for the model to work upon.

* **Scalability**

The scalability of the model depends on the amount of data provided to be studied. The model is only as accurate in prediction as the quality of the data set provided. With more data a better working model can demonstrate more accurate predictions.

* **Availability**

The model will be available for testing and by direct access to the source code and data set or via an online web app.

* **Usability**

The model will allow the time forecasting of accidents that may or may not happen in a particular area and the accuracy will completely depend on the training data provided.

* **Correctness**

The model will be using various algorithms to process the data and obtain the most accurate predictions for accidents occurring in a given locality.

* **Maintainability**

The frameworks and technology used is up to date and can be easily maintained to satisfy needs that may arise in the future.

* **Portability**

The model will be portable across systems that support the required software specifications and consist of the tools used to develop the model.

1. **Performance Requirements**

* The data regarding the accidents required for testing and drawing the required inferences must be collected and fed to the model.
* The data provided must be legitimate and must be collected by trustworthy procedures.
* The data must be in the form of a table saved as a file in the .csv format.
* The inferences drawn are to be studied from the point of view of analytical information which may help in drawing further conclusions.
* Of the various attributes the date, day and time of the accidents that occurred must be mentioned for the best time series forecasting.

1. **Design Constraints**

* The primary design constraint is the availability and collection of legitimate and qualitative Data to be used for training the model to obtain satisfactory results. For the predictions to be accurate we must obtain the right data and the right amount of data as well.
* The data set must not have too many missing values.
* If categorical data is involved all the data obtained must not be too biased.
* We must get accurate data that well describes the complete conditions of the accidents that take place in the mentioned locations.
* The inferences drawn are done exclusive of the other target variables in the data set.

1. **Other Requirements**

* Computer with proper processing power.
* Numpy, Pandas and Scikit Libraries are to be available.
* Libraries for plotting the graphs.