2. A description of the data and how it will be used to solve the problem.

Data understanding:

Data analysis is important for implementing essential countermeasures for road traffic accidents. It allows analyzation of historical data and allows the topics to be addressed from a different point of view. Although the external validity of data analysis can be questioned, they are useful since the data itself is of the real world.

Since the number of casualties and reported accidents have decreased since, the issue most frequent and challenging within the transportation industry is safety with built up roads and urban environments a constant problem.

By observing the data, it some feature objects Date, Road class Are object type remaining are float. And the attributes are assigned integer values already

By analysing the data visually, we can identity the major and minor cause es of different attributes. Weather, road surface, light condition and type of sex car driving female or male, age group, date and time of the important attributes for the car accident servity UK data. Casualty Severity. The Fatal, Serious, Slight area the Casualty Severity conditions.Description of data to solve car accident servity

Many strategies can be deployed to reduce deaths during traffic accidents, and one of them, consists in speeding post- accidents attention. In this sense, predicting accidents severity could be a key for quick response to such accidents [3]. In this sense, data mining approach combined with other techniques for knowledge discovery is an encouraging way for understanding, classifying and even for predicting injury severity Many strategies can be deployed to reduce deaths during traffic accidents, and one of them, consists in speeding post-accidents attention. In this sense, predicting accidents severity could be a key for quick response to such accidents [3]. In this sense, data mining approach combined with other techniques for knowledge discovery is an encouraging way for understanding, classifying and even for predicting injury severity Many strategies can be deployed to reduce deaths during traffic accidents, and one of them, consists in speeding post-accidents attention. In this sense, predicting accidents severity could be a key for quick response to such accidents [3]. In this sense, data mining approach combined with other techniques for knowledge discovery is an encouraging way for understanding, classifying and even for predicting injury severity Many strategies can be deployed to reduce deaths during traffic accidents, and one of them, consists in speeding post-accidents attention. In this sense, predicting accidents severity could be a key for quick response to such accidents [3]. In this sense, data mining approach combined with other techniques for knowledge discovery is an encouraging way for understanding, classifying and even for predicting injury severity

For the car accident servity weather, road, light conditions,

k k-Nearest Neighbor Method which contained a number of algorithms used in the process of prediction and the relationship between both dependent and independent variables. The location effectiveness and timeliness features of Twitter can be proved in a recent accident detection study that uses the GPS-enabled smartphones and travel behaviour study which has been validated by the household travel survey.

By applying the machine learning techniques, we find the solution for it. For the applied data science car accident servity project, I have used KNN and Decision classifier techniques. For the current project source file is collision types csv from capstone project. There different 17 type attributes which columns and 100 rows. To make possible ways to destination without car accident, I have ignored some attributes. Decision tree classification is best suited for the current understanding of the project. The following shows the columns of attributes descript1ion.

Conditions etc. attributes area object type.

|  |  |  |  |
| --- | --- | --- | --- |
| S.No | Attribute | Description | Data type |
| 1 | Grid Ref: Easting | UTM co-ordinates | int64 |
| 2 | Grid Ref: Northing | UTM co-ordinates | int64 |
| 3 | Number of Vehicles' |  | int64 |
| 4 | Accident Date' | Date | object |
| 5 | Time (24hr)' |  | int64 |
| 6 | '1st Road Class' | Motorway, A(M), A, B, C , Unclassified | object |
| 7 | '1st Road Class & No' |  | int64 |
| 8 | 'Road Surface' | 1 Dry  2 Wet / Damp  3 Snow  4 Frost / Ice  5 Flood (surface water over 3cm deep) | int64 |
| 9 | 'Lighting Conditions', | 1 Daylight: street lights present  12 Daylight: no street lighting  3 Daylight: street lighting unknown  4 Darkness: street lights present and lit | int64 |
| 10 | 'Weather Conditions', | 1 Fine without high winds  2 Raining without high winds  3 Snowing without high winds  4 Fine with high winds  5 Raining with high winds  6 Snowing with high winds  7 Fog or mist – if hazard  8 Other  9 Unknown | int64 |
| 11 | 'Local Authority', |  | object |
| 12 | 'Vehicle Number', |  | int64 |
| 13 | 'Type of Vehicle', | 9.Car | int64 |
| 14 |  |  | int64 |
| 15 | 'Casualty Class', ' | 1 Driver or rider  2 Vehicle or pillion passenger  3 Pedestrian | int64 |
| 16 | Casualty Severity', ' | 1 Fatal  2 Serious  3 Slight | int64 |
| 17 | Sex of Casualty', | 1 Male  2 Female | int64 |
| 18 | 'Age of Casualty' | Age group | int64 |

The current project about car accident servity, then I drop the all other rows accept car accident data. For this, in column name type of vehicle I have selected 9(9 value which corresponds to car) values by applying filter in excel data and convert to csv format.

For the current project python from jupyter lab used.

I have imported csv using pandas, NumPy, so modules from python frame work.

fie

Fig.1 Data frame df to read car\_uk-19 csv data file from my computer.

Fig.2 the function call for read datatypes of each column df.dtypes

Background

Road and traffic accidents are uncertain and unpredictable incidents and their analysis requires the knowledge of the factors affecting them. Road and traffic accidents are defined by a set of variables which are mostly of discrete nature. The major problem in the analysis of accident data is its heterogeneous nature. Thus, heterogeneity must be considered during analysis of the data otherwise, some relationship between the data may remain hidden. Although, researchers used segmentation of the data to reduce this heterogeneity using some measures such as expert knowledge, but there is no guarantee that this will lead to an optimal segmentation which consists of homogeneous groups of road accidents. Therefore, cluster analysis can assist the segmentation of road accidents. The newly emerged data source, social media data, has proved its capability in recent traffic studies including activity pattern identification special traffic-related events traffic flow prediction transport information management travel mode detection destination or route choice, etc. According to Rashidi et al. (2017), as social media data encompasses information that is revealed by users in realistic situations, such data is free from sampling, surveying or laboratory biases. The location effectiveness and timeliness features of Twitter can be proved in a recent accident detection study that uses the GPS-enabled smartphones and travel behaviour study which has been validated by the household travel survey The machine learning methods have thrived in the applications of language and text modelling in recent years, which can potentially counter the challenges in processing and classifying the tweets. In most of the studies, language modelling can be taken as a kind of information extraction from the text messages, which is the process of converting the unstructured text information into a structured database and solving it as a supervised or unsupervised learning task. Accidents Dataset. A file containing annual data of traffic accidents, structured on 38 attributes, such as luminosity, road type, weather conditions, accident type, day of the week, region, total victims, total injuries, etc.

4 Research Objectives

The objectives of this dissertation are mainly the following:

1. Gather a comprehensive database of road accident statistics for built up roads with factors that affect road safety which have been provided by the database.

2. Analyse data for the factors, which can impact accident rates (e.g. light conditions, weather, road surface conditions, types of junctions etc.)

3. Determine type of road classes with highest and lowest amount of accident rates from analysing tables of road accident statistics and charts created from the database (STATS19)

4. Suggest appropriate measures for the factors and the road class determined the most dangerous

Accidents Dataset. A file containing annual data of

traffic accidents, structured on 38 attributes, such as

luminosity, road type, weather conditions, accident

type, day of the week, region, total victims, total

injuries, etc. This file contains more than 83,000

records for 2015 year

Accidents Dataset. A file containing annual data of

traffic accidents, structured on 38 attributes, such as

luminosity, road type, weather conditions, accident

type, day of the week, region, total victims, total

injuries, etc. This file contains more than 83,000

records for 2015 year

Accidents Dataset. A file containing annual data of

traffic accidents, structured on 38 attributes, such as

luminosity, road type, weather conditions, accident

type, day of the week, region, total victims, total

injuries, etc. This file contains more than 83,000

records for 2015 year

Data Preparation:

The data preparation includes all the required activities to construct the final dataset which will be fed into the modelling tools. Data preparation can be performed multiple times and it includes balancing the labelled data, transformation, filling missing data, and cleaning the dataset.

The Pyton 3.6 language is the main tool operation have used for the current project

In order remove data redundancy I have removed null values or no values

I perfumed operation

 For data types

 df. dtypes

 df. describe

 df. dtypes

 df. corr()

Modelling:

In this phase, various algorithms and methods can be selected and applied to build the model including supervised machine learning techniques. You can select SVM, XGBoost, decision tree, or any other techniques. You can select a single or multiple machine learning models for the same data mining problem. At this phase, stepping back to the data preparation phase is often required.

Decision tree model has been adopted from python

Evaluation:

Before proceeding to the deployment stage, the model needs to be evaluated thoroughly to ensure that the business or the applications' objectives are achieved. Certain metrics can be used for the model evaluation such as accuracy, recall, F1-score, precision, and others.

Deployment:

The deployment phase requirements vary from project to project. It can be as simple as creating a report, developing interactive visualization, or making the machine learning model available in the production environment. In this environment, the customers or end-users can utilize the model in different ways such as API, website, or so on.

Participation is optional

References

Mitchell, M. (2019) Programming Languages for Data Scientists. Medium. https://towardsdatascience.com/programming-languages-for-data-scientists-afde2e af5cc5

Thanasis (2017) UK Road Safety: Traffic Accidents and Vehicles. Kaggle.com. https://www.kaggle.com/tsiaras/uk-road-safety-accidents-and-vehicles