**MAKERERE  UNIVERSITY**

**COLLEGE OF COMPUTING AND INFORMATION SCIENCES**

**(YEAR II) RECESS TERM**

**CONCEPT PAPER FOR PYTHON PROJECT 2019**

**PROJECT MEMBERS**

|  |  |  |
| --- | --- | --- |
| **NAME** | **REGISTRATION NUMBER** | **STUDENT NUMBER** |
| Nagaba Angel | 17/U/726 | 217000189 |
| Okello Marvin Kevin Ochira | 17/U/9569/PS | 217017015 |
| Karungi Lydia | 17/U/4676/PS | 217002012 |
| Wepukhulu Bruno | 17/U/10891/PS | 217012574 |

**PROJECT COORDINATOR**: MR. KAMULEGEYA GRACE

**CONCEPT NOTE FOR ROAD ACCIDENTS PROBLEM**

**INTRODUCTION.**

Our proposed area of interest is road traffic accidents. The major inspiration for this project is the increased death rate due to road traffic accidents to analyse data in order to come up with insights to help us answer various questions concerning accidents e.g the likelihood of death when an accident occurs, the likelihood of accidents for different age groups e.t.c. we also hope to come up with policy recommendations to help reduce road traffic accidents.

**BACKGROUND**

We got our data from <https://www.kaggle.com/silicon99/dft-accident-data>

The data was collected

The dataset was collected by the UK police forces on every vehicle collision for the period of 2005-2015. The data is usually collected when an accident happens by filling a Stats19 Report Form from the link

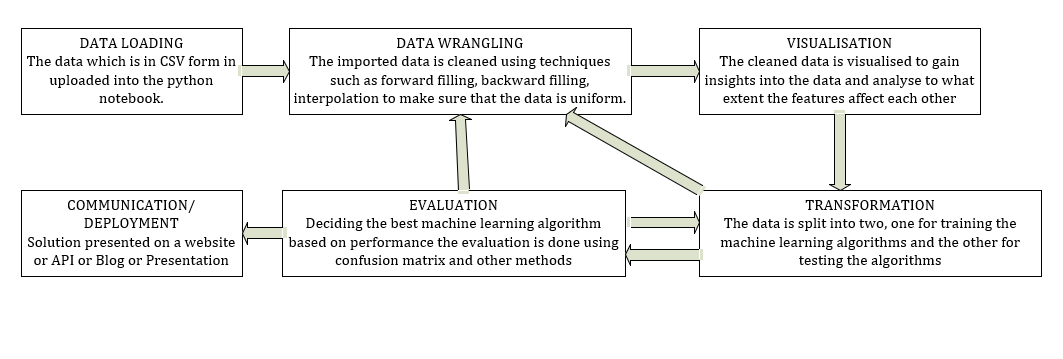
[http://docs.adrn.ac.uk/888043/mrdoc/pdf/888043\_stats19-road-accident-injury-statistics-report-form.pdf](http://docs.adrn.ac.uk/888043/mrdoc/pdf/888043_stats19-road-accident-injury-statistics-report-form.pdf" \t "https://mail.google.com/mail/u/1/" \l "inbox/_blank)

The dataset is multivariate containing of three CSV files i.e accidents, casualities and vehicles. Accidents is the primary dataset which has 32 columns. It has the unique Accident\_index which references the casualities and vehicles tables. The casualities table has 15 columns and the vehicles table has 22 columns. Both the vehicles and casualities table contain the Accident\_index that references to the primary dataset(Accidents)

This dataset will help us to analyse the trend of accidents in the uk and also apply various data visualisation tools to come up with insights.

**DATA PIPELINE**

This illustration of the pipeline was creatively generated by the group and it is explained in the steps below.

****

A data pipeline enables the data scientist to transform data from one representation to another through a series of steps. It involvesthe following steps:

1. Data Loading

At this step, the data is loaded into the python notebook. The three csv files i .e accidents,casualities and vehicles are all loaded and transformed into dataframes using the pandas python library. The three datasets are then merged into a single dataframe for further analysis. This is easier because all the three datasets have a common Unique id Accident\_Index.

1. Data Wrangling

Data Wrangling involves is the process by which data is cleaned so that it becomes easy to analyse and visualise it. The data wrangling process is as follows:

Checking missing values The dataset is checked to find out if there are any missing values and int this case we have two types of missing values which is -1 and ‘Nan’. various methods such as Backward fill, Foward fill, Interpolation, Dropping all missing values, Filling in the mean or median e.t.c can be used to remove the missing values.

Label encoding

This is the process where the categorical features of the data are transformed into numeric values usually 0’s and 1’s. for this case the dataset features are all numerical already so there is no need to label encode.

Feature scaling

This is a step of data pre-processing which is applied to independent variables or features of data. It helps to normalise the data within a particular range. It also helps in speeding up the calculations in an algorithm. The data set will be feature scaled in order to bring the outliers closer to other values in the dataset. Two main libraries are used i .e RobustScaler and StandardScaler.

1. Visualisation

Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. Data visualisation will help us to determine the spread of the dataset through measures like range,quartiles and the interquartile range,variance and standard deviation, determine the correlation between features. Various visualisation tools such as heat maps, histograms, scatter plot, line plots, box plots, regression analysis and they will be applied to the following key features of the dataset: Accident severity, Number of vehicles involved in the accident, day of the week the accident happened, time of the day when the accident happened, Road type on which the accident happened, speed limit, light and weather conditions, vehicle type, vehicle manouvre, age of the driver,sex of the driver, point of impact, age band, age of the vehicle, casuality class, casuality severity, casuality type, age of the casuality e.t.c. Visualisation tools like hiatograms, maps, scatter plot, line plots e.t.c will be used to analyse the spread of data which will in turn help us come up with better policy recommendations.

1. Transformation

At this step of the pipeline, machine learning algorithms are applied onto the data and they will help us to predict our final outcome which is the possibility of death when an accident happens. Machine learning algorithms include: logistic regression, linear regression, Support Vector Machine, Decision trees, K- Nearest Neighbours e.t.c

1. Evaluation

Theaccuracy of the above algorithms is then evaluated based on the scores in order to find out the best algorithm. This evaluation can be carried out using the precison, Recall, confusion matrix and cross validation performance measures.

1. Deployment/Communication

After the data has been analysed and visualised, the solution has to be presented or communicated. This can be done with the help of either of the following tools: A blog, a website, an API, a dashboard or a presentation.

References

<https://www.kaggle.com/silicon99/dft-accident-data>