MAKERERE UNIVERSITY

COLLEGE OF COMPUTING AND

INFORMATION SCIENCES

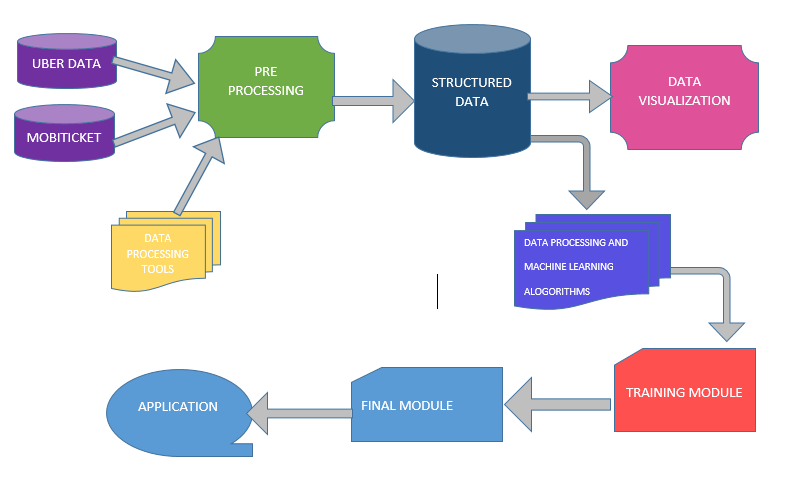
**BSE 2301 PROFESSIONAL SOFTWARE ENGINEERING MINI PRACTICAL**

**PROJECT II**

**SOFTWARE REQUIREMENTS SPECIFICATION DOCUMENT**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **NAMES** | **Reg No** | **Stud No** |
| 1 | LWANGA DARLINE | 17/U/5836/EVE | 217017397 |
| 2 | KYADONDO SOLOMON | 17/U/5538/EVE | 217017395 |
| 3 | GYAGENDA ARAFAT | 17/U/4117/EVE | 217003258 |
| 4 | KISA JEREMIAH | 17/U/5250/EVE | 217017394 |

**DATA PIPELINE**



## Data sources

## The data is obtained from csv files(datasets) from zindi africa and we will be using mainly two datasets namely;

## nairobi\_traffic.csv

## The dataset of tickets purchased from Mobiticket for the 14 routes

## test\_questions.csv

## The dataset on which we will use to estimate number of tickets sold by Mobiticket per unique ride

## Libraries to be used

## The various libraries imported will enable the carrying out of various functions like dataset importing, data visualization.

## The libraries to be imported include

## NumPy

We are using this library due to its provision of high-performance manipulation of sequences of homogenous data items. The library will be represented as np

* Pandas

This package contains multiple methods for convenient data filtering. Pandas has a variety of utilities to perform Input/output operations in a seamless manner.

* Sea-born

This library will be used in carry out statistical data visualization

* Matplotlib

The library will be used because it more efficient in terms of data visualization and easy to use. A sub package of the library named pyplot will be used and is represented as plt

### Data Preprocessing

Libraries used here are numpy, matplotlib, pandas, sea-born, scipy and warnings. We changed the date to the date time format using the pandas library, we also used some visualization under here and variable analysis and scaling using the distribution of the ticket sales and then plotted a QQ-plot.Under data processing we used two types of algorithms;

Linear regression

Logistic regression.

1. **Missing information**

We have to check for the rows with missing data and delete those rows because the existence of rows with missing data can cause errors.

1. **Visualization**

Columns that have been checked can be used to plot a bar graph to show the relationship between these columns, Travel to and car-type can be used to plot a bar graph to find out the correlation between them.

Visualization componentS

Scatter plot

We use scatter plot to visualize nonlinear data

Bar graph

We a going to use this to summarize large amounts of data

Histograms

We choose to use a histogram because it allows the visualization of distribution of data.

1. **Data modules**

In data modules, we used a training module and deployed the selected module to get the final module of the data.

1. **Application**

This is the last step of the system which clearly shows the predicted number of tickets sold by different companies and the different car modules used.