**Applied Data Science Capstone Project –**

**on**

**Car accident severity**

**Peer-graded Assignment:**

**Final Report**

**By**

**Ravi Jonnalagadda**

**Senior Research Scholar**

**Osmania University**

**Submitted to**

**Coursera IBM Data Science Professional Certification**

**Table of Contents**

1. **Introduction**
2. **Business understanding**
3. **Objectives**
4. **Finding Data resource**
5. **Data understanding and Data preparation**
6. **Data science methodology**
7. **Data visualization and analysis**
8. **Data modelling and Deployment**
9. **Data analysis and Results**
10. **Conclusion and Recommendations**
11. **References**

**1.Introduction**

There is a huge impact on the society due to traffic accidents where there is a great cost of fatalities and injuries. In recent years, there is an increase in the researches attention to determine the significantly affect the severity of the driver’s injuries which is caused due to the road accidents. Accurate and comprehensive accident records are the basis of accident analysis. the effective use of accident records depends on some factors, like the accuracy of the data, record retention, and data analysis. There are many approaches applied to this scenario to study this problem.

A recent study illustrated that the residential and shopping sites are more hazardous than village areas.as might have been predicted, the frequencies of the casualties were higher near the zones of residence possibly because of the higher exposure. A study revealed that the casualty rates among the residential areas are classified as relatively deprived and significantly higher than those from relatively affluent areas. Road accidents have become very common these days. Nearly 1.25**million**people die in road crashes each year, on average,**3,287** deaths a day. Moreover, **20–50 million** people are injured or disabled annually. Road traffic crashes rank as the **9th** leading cause of death and accounts for **2.2%** of all deaths globally. In this contest to better servity of the accidents machine learning and neural techniques has been used analysis. These techniques helpful to better accident servity. Car accidents are one of the common types of collision occurring everywhere globally every day. By analysing the different factors which cause the collision. In this section we are discussing about data Capstone project topic Car accident servity. where you discuss the business problem and who would be interested in this project.

**2.Business Understanding**

Car collision or car accident one of the collision type in Road accidents**.** According to Corrigan [1], despite collecting large quantities of traffic data, Transportation Departments of all levels are unable to use such data to good effect. Founded in 2015, a start-up called ODN could predict when and where accidents are most likely to happen. Officials could use such information to direct safety efforts at the stretches of road where the impacts could be the biggest. In the context of this research, some of the developed countries like US, UK governments could use the information generated from a prediction system with a Neural Network predicting the accident severity and use this information to enhance the laws to build safer roads for the future. In this project we are dealing with the all the possible ways to reach the destination by overcoming car accident servity with the different critical traffic conditions on the way to journey. By prediction car accident servity improve the traffic safety measures. And implements the traffic rules accordingly by governments Better servity conditions.

**Motivation of the study**

The goal of the study is to encourage the improvement of road safety by analysing several factors that can affect safety in Built up roads. The project aims to utilise related literature along with analyzation of roads to discern any continuously trending factors, which may influence Road Safety and accident rates within the United Kingdom. The project carries the car accident servity.

**3.Objectives**

The objectives of this Capstone project 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 UK-2019 accident set.

4.Suggest appropriate measures for the factors and the road class determined the most dangerous for improve car accident servity.

**4.Data resource**

For the current project I have used from a pen source from google data search engine **data** data.gov.uk Find open data. The data details are as followsPublished by: Leeds City Council Last updated: 22 July 2020 Topic: Not added Licence: Open Government Licence.<URL:https://datasetsearch.research.google.com/search?query=car%20accidents&docid=yKyJqNCmNypGNKe7AAAAAA%3D%3D>**.** In which I have selected downloaded accidents dataset 2019 in csv format

**5.Data understanding and Data preparation methods**

**Data understanding**

The project data UK accident year 2019 set consists following attributes. The data which I have downloaded already assigned to attributes. In order make data comfort and easy to use number are assigned for each column of object. Because The columns road surface, road class, Lighting 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.

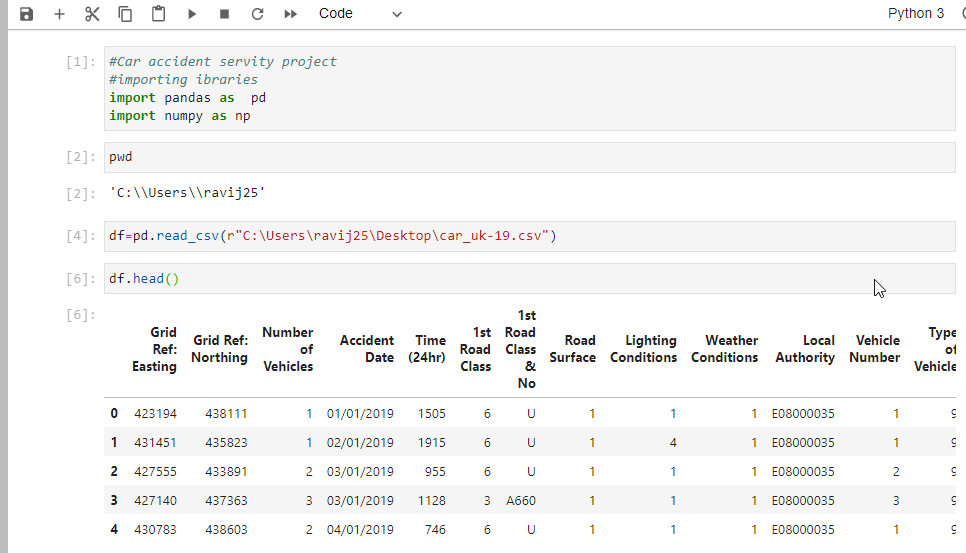
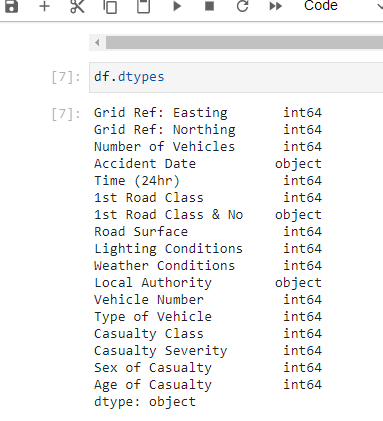
fie

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**

For the current project I have used KNN Machine learning model. k-Nearest Neighbour Method a machine learning technique is implemented in python programming language. 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.

Then the model is evolved is using Jacob and F-1 square methods.

**Data preparation methods**

In this step of project, we are loading the csv data file from which we have collected.

To visualize data, I used the Python programming language. All necessary libraries loaded, and the script required the following libraries: 1) pandas for data frame management, 2) Marplot for visuals, and 4) NumPy to perform scientific computing. For making requests, Google authentication is mandatory. The python code must set the google application credentials environment variable.

This stage focuses on the quality of the data, specifically on

the quality of the attributes of the dataset. Only the attributes

that are selected will be used as input for the machine learning

model. The attributes that provide the lowest value, that is, the

useless attributes, are discarded, using the following quality

indicators for each attribute of the dataset:

• Bad quality attributes. These attributes must be

eliminated from the dataset, applying the following

rules: More than 70% of all values in these attributes

are missing; the attributes are identifiers, where for

each row of the dataset they generate a different value;

the attribute is constant with more than 90% of all

values are equal.

• Attribute that has a very low or very high correlation

with respect to the categorical attribute used to predict

whether an accident is serious or not: Low correlation,

a correlation of less than 0.01%; High correlation, aThis stage focuses on the quality of the data, specifically on the quality of the attributes of the dataset. Only the attributes that are selected will be used as input for the machine learning model. The attributes that provide the lowest value, that is, the useless attributes, are discarded, using the following quality indicators for each attribute of the dataset: • Bad quality attributes. These attributes must be eliminated from the dataset, applying the following rules: More than 70% of all values in these attributes are missing; the attributes are identifiers, where for each row of the dataset they generate a different value; the attribute is constant with more than 90% of all values are equal. • Attribute that has a very low or very high correlation with respect to the categorical attribute used to predict whether an accident is serious or not: Low correlation, a correlation of less than 0.01%; High correlation, a

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

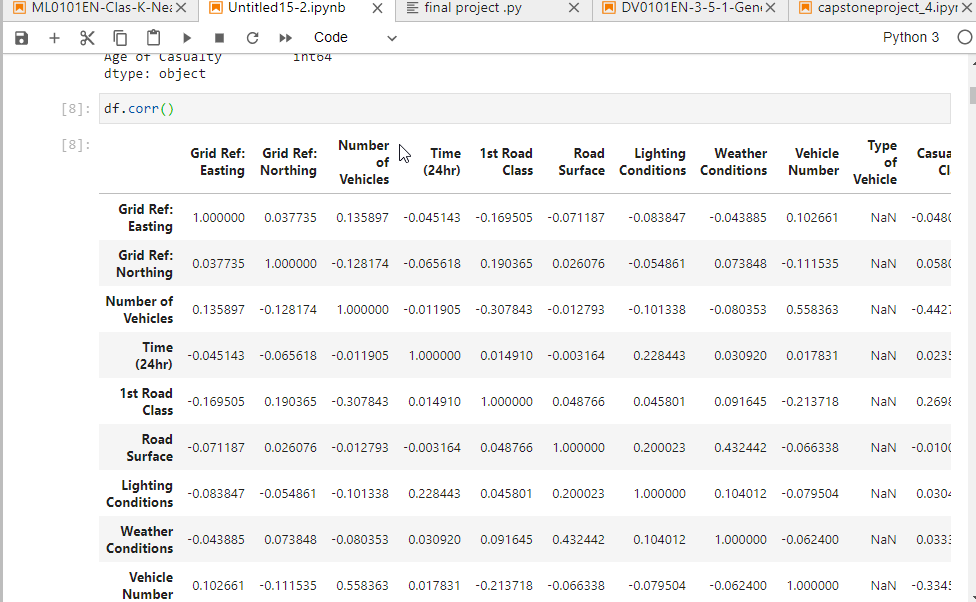


Fig.3 I have applied corer () function for to eliminate null values and no values.

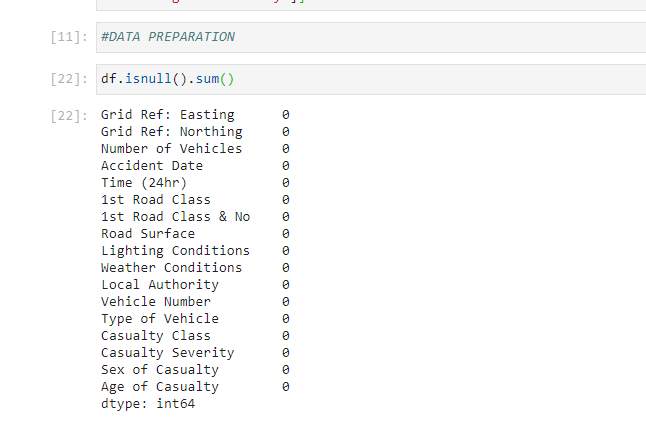


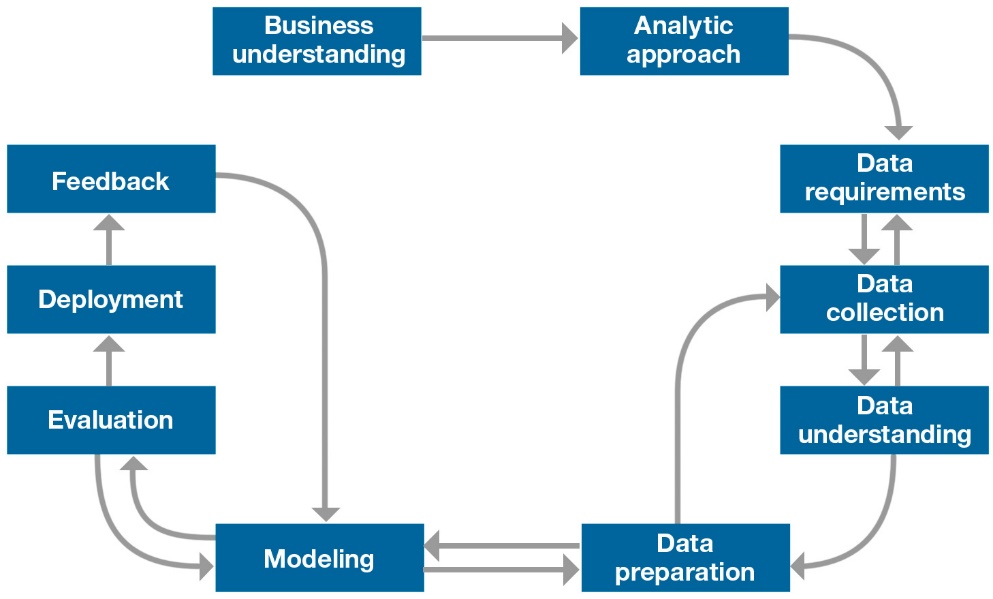
Fig.4 function df.isnull() to delete null values

**6.Data science methodology**

The project car accident involves the following steps of data science methodology they are as follows,1. Business understanding 2. Data understanding 3. Data preparation 4. Modelling 5. Deployment 6. Evaluation.

For the current project I have used KNN Machine learning model. k-Nearest Neighbour Method a machine learning technique is implemented in python programming language. k k-Nearest Neighbour Method which contained a number of algorithms used in the process of prediction and the relationship between both dependent and independent variables.

Then the model is evolved is using Jacob and F-1 square methods.



**Data science methodology**

**Programming language used for the project**

In a recent worldwide survey, it was found that 83% of 24,000 data professionals used Python. Data Scientists consider Python because it is a general-purpose and dynamic programming language. It contains ideal packages for Machine Learning tasks and is inherently an Object-Oriented Programming Language. In contrast, millions of analysts and data researchers use R Programming Language to handle their most difficult issues in the fields running from Computer Science to extensive marketing.

I have used Anaconda NAVIGAOR, Jupyter notebook, and I have my capstone project on published GitHub Repository.

The following area the libraries from used.

#importing libraries

import pandas as pd

import numpy as np

Seaborn as sns

import itertools

import numpy as np

import matplotlib.pyplot as plt

from matplotlib.ticker import NullFormatter

import pandas as pd

import numpy as np

import matplotlib.ticker as ticker

from sklearn import preprocessing

%matplotlib inline

**7.Data visualization and analysis**

Using different charts using matploib.pyplot libraries functions

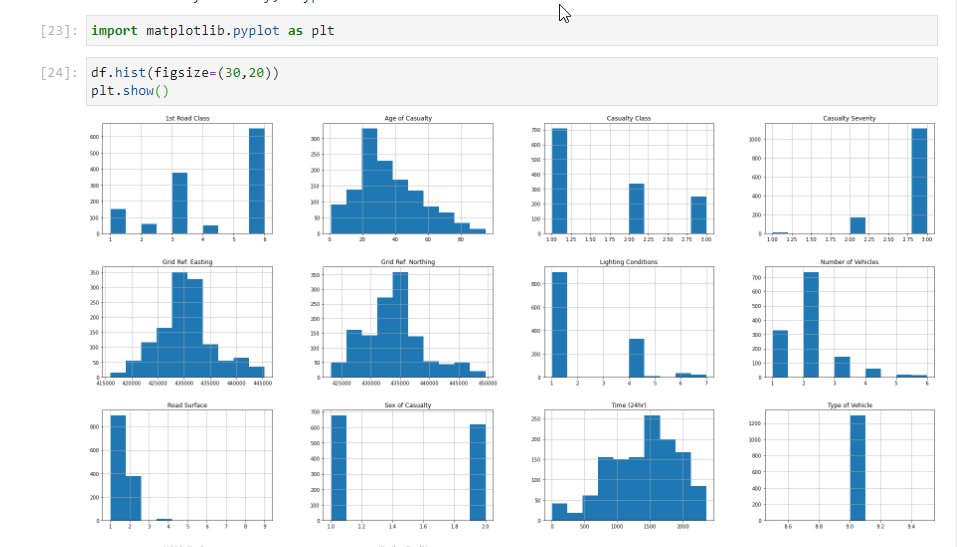
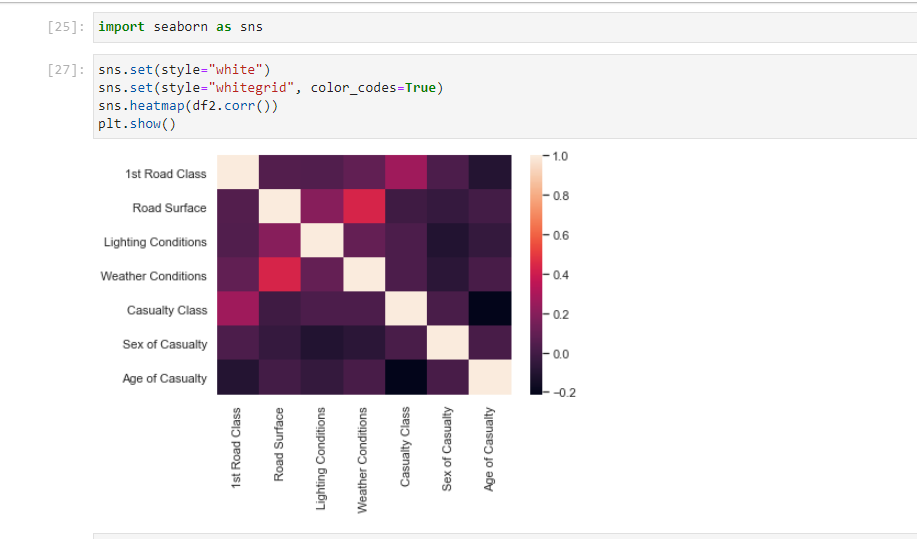


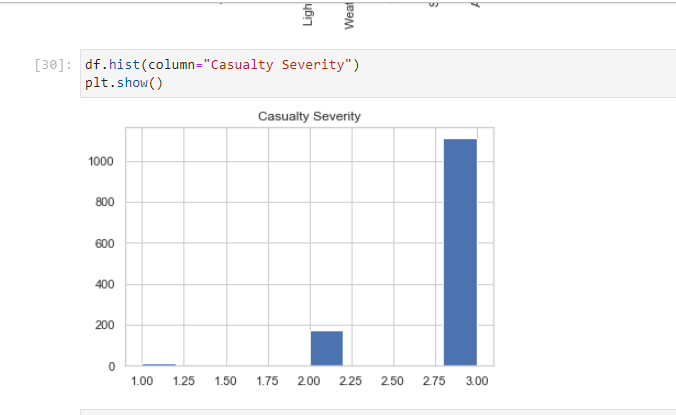
Fig.5 hitogram for data df

**Correlation matrix**

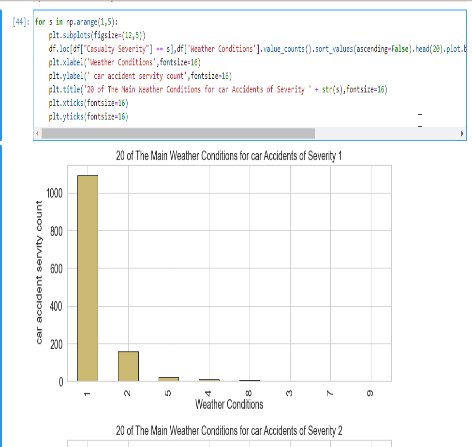
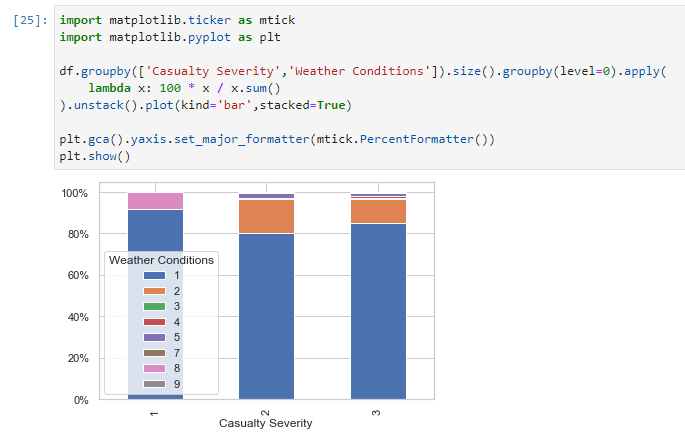
**Using seaborn library correlation matrix**



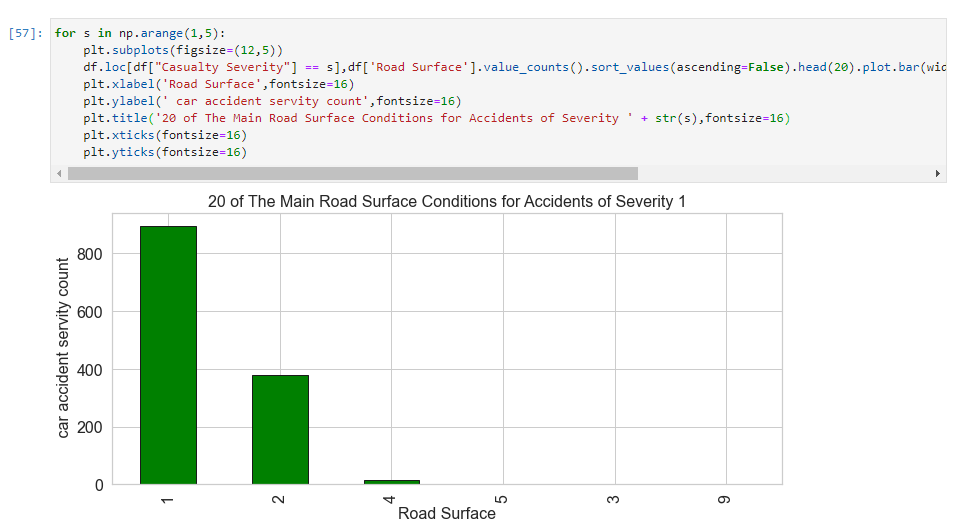
**Fig 6. Correlation matrix**



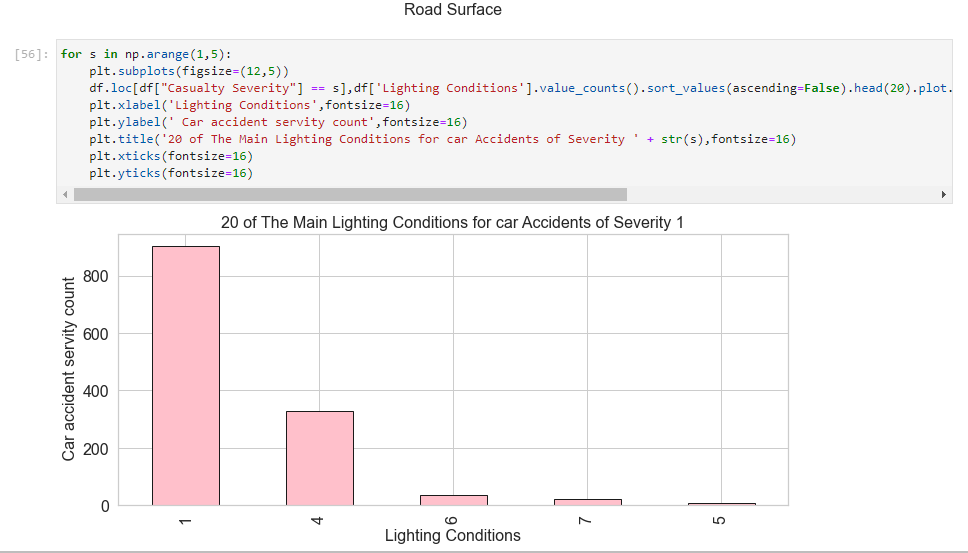
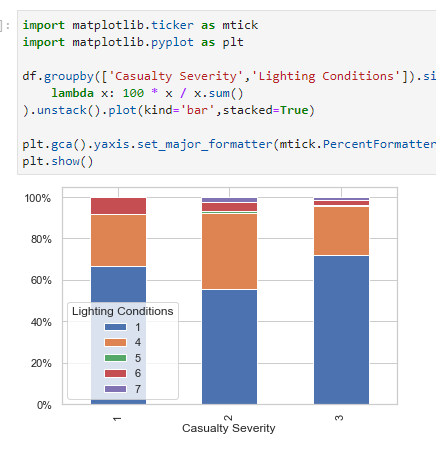
**Fig.7 Car Causality of servity histogram**

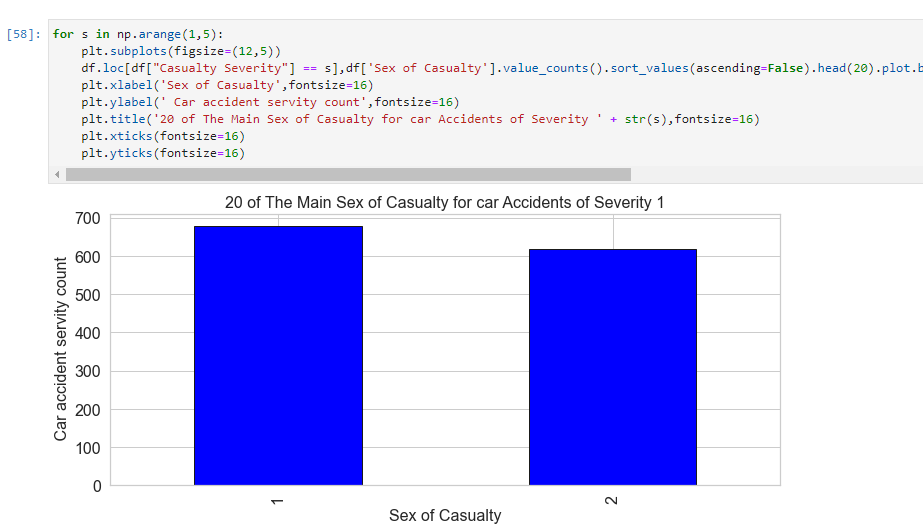
**Fig.8 Car accident servity with Different weather conditions using bar chart**

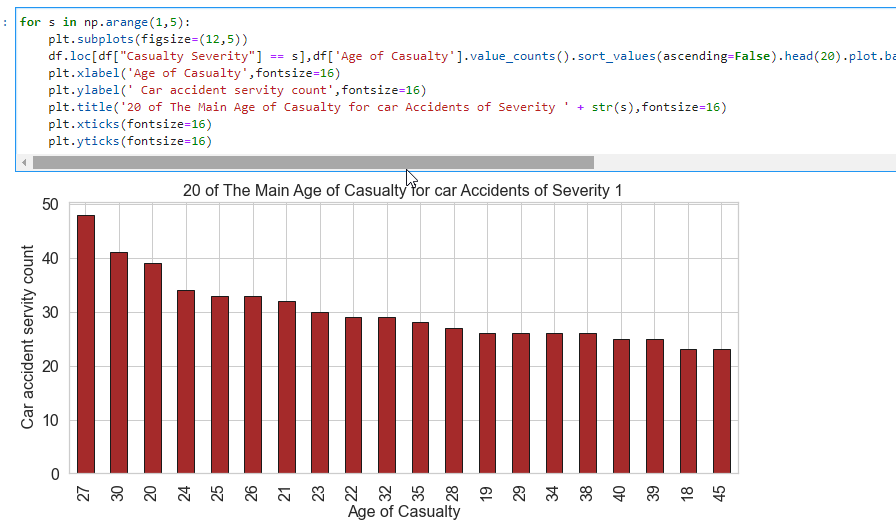
**Fig.9 Car accident servity with surface conditions using bar chart**

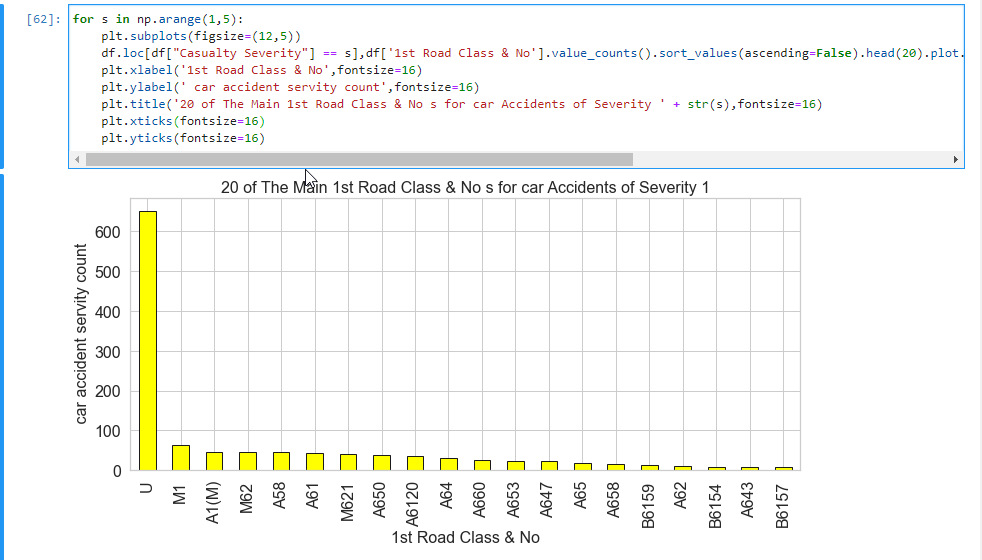
**Fig.10 Car accident servity with light conditions using bar chart**

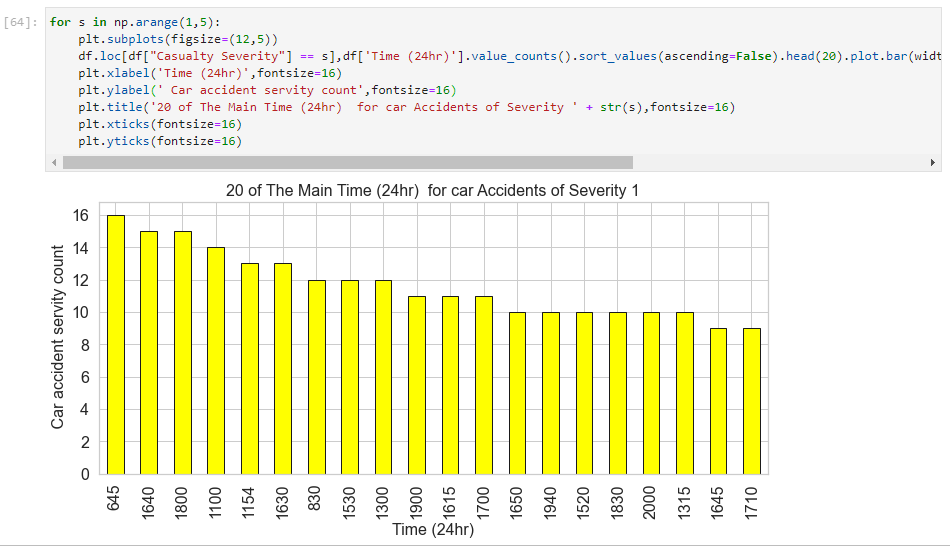
**Fig.11 Car accident servity with sex type using bar chart**



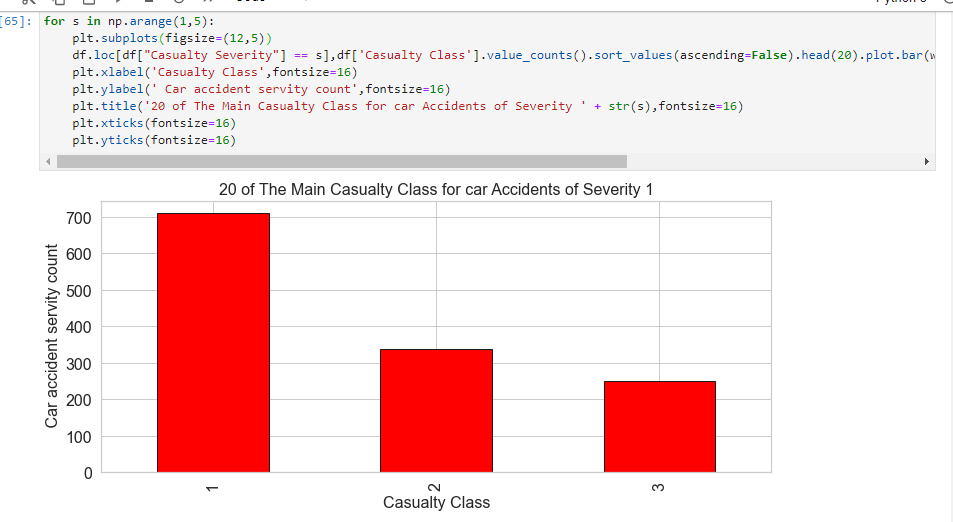
**Fig.12 Car accident servity with age of causality using bar chart**



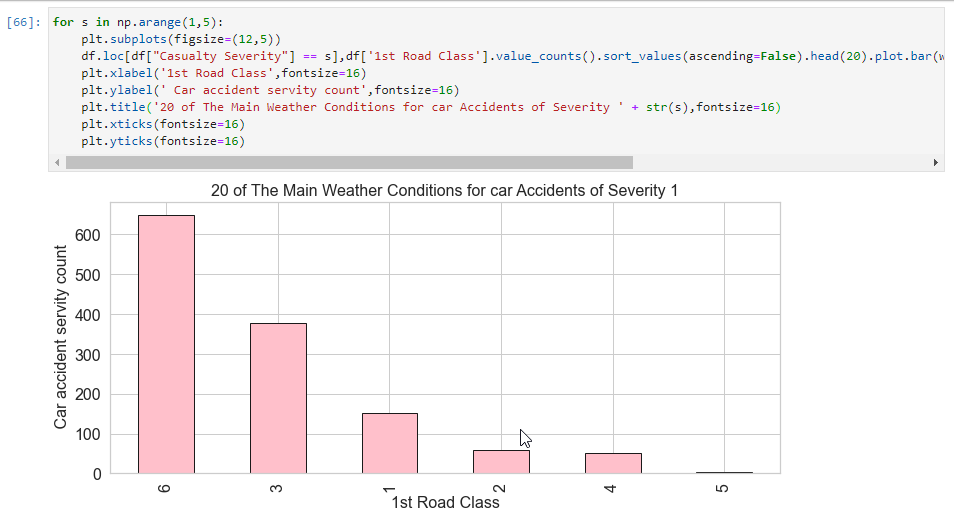
**Fig.13 Car accident servity with the main 1st road class using bar chart**



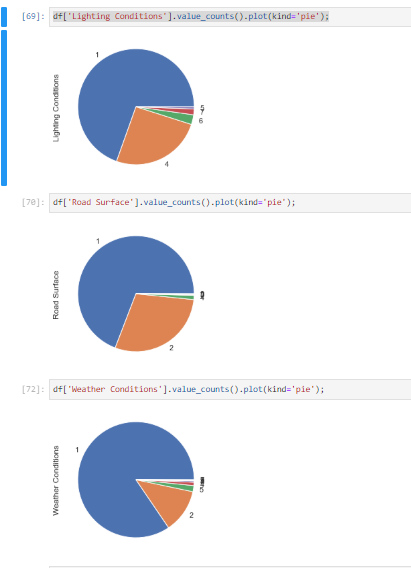
**Fig.14 Car accident servity with in a Day with 24hrs using bar chart**



**Fig. 15 Car accident servity with causality class using bar chart**



**Fig. 15 Car accident servity with 1st road class using bar chart**



**Fig. 16 Car accident servity with Light, road, weather conditions using pie chart**

Pie chart visualisation of different attributes

**8.Data modelling and deployment**

For the current project KNN and Decision tree classifier has been used car accident servity.

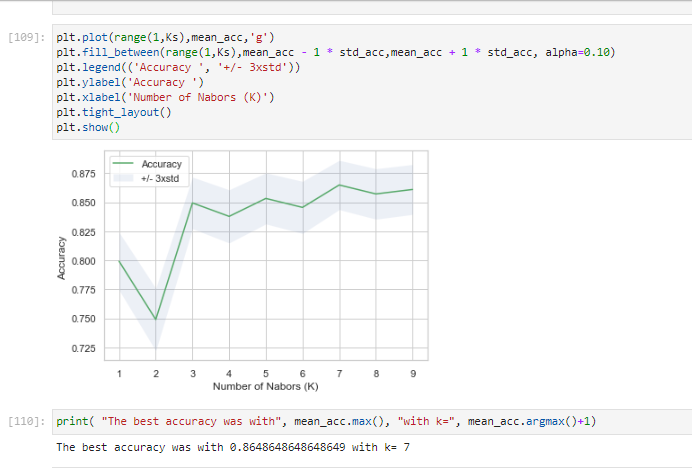
c) KNN: KNN is a classification algorithm which is based on feature similarity. It analyses the data and measure the distance and similarities between data and cluster them based on K values. Distance is calculated in many ways, for this research (Labib 2019), we used Euclidean distance measurement. The class of new input data is classified by calculating the distance between the clusters and assigned it to the closest one. The posterior probability is mainly the probability of “a” being true given that “b” is true.

Decision tree classification

In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand. The logic behind the decision tree can be easily understood because it shows a tree-like structure.

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. The limited word features can be utilized for specific research.

Fig 11. Car accident servity with different weather conditions.



KNN nearest neighbour model for Car accident servity

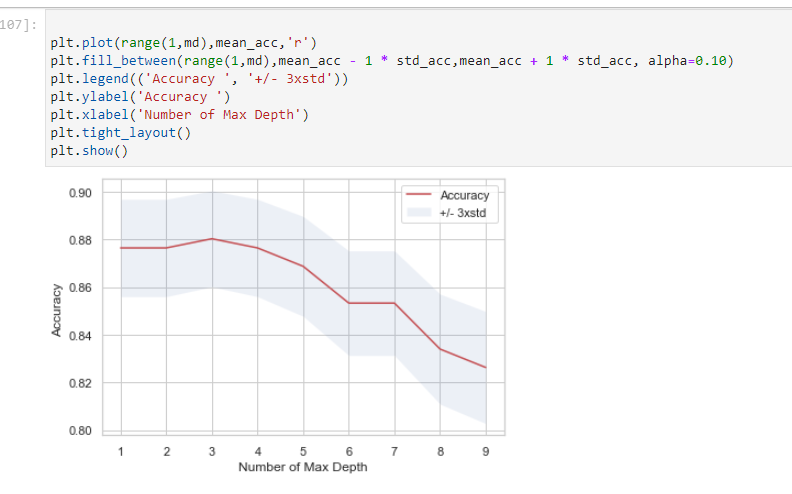
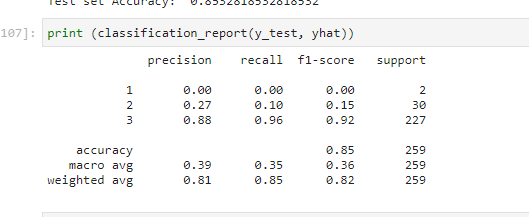


Fig 17 Decision tree model



Decision tree classification report

**9.****Data analysis and Results and conclusions**

For the present project from data UK accidents set the data analysis is follows.

From the bar plot charts, we can observe the Car accident servity with different attributes.

The main attributes we can consider Road class, Road surface, Weather conditions, light conditions,

1. Car accident servity is high for weather condition from fig.8 1,2 Fine without high winds, without high winds and remaining and 3,4,5,6,7 Snowing without high winds, Fine with high winds, raining with high winds, snowing with high winds, Fog or mist – if hazard which are very risky and low servity.
2. Car accident servity is high for Road class from fig.5 6,3 high servity ,1,2,4 ,5 high risk.
3. Car accident servity Road surface 1,2 Dry, Wet / Damp are safe, 4,3,5,6 which are Snow, Frost / Ice, Flood (surface water over 3cm deep) risky.
4. Car accident servity for light conditions from fig.10, 1,4 Daylight: street lights present, Darkness: street lights present and lit are safe.2,3,5 Daylight: no street lighting Darkness: street lights present but unlit, Darkness: no street lighting, Darkness: street lighting unknown are unsafe.
5. Car accident servity for time 24hr in day from fig.14, it is observed that from morning 6:45AM to 17:000PM car accident servity has been decreased.
6. Car accident servity for age group is varying from 20-45.
7. Car accident servity for road no U is high, And safe.
8. Car accident servity for sex 1 slightly high than 2, means male servity high than female.

**10.Conclusion and Recommendations**

**Acknowledgement**

**I express my sincere thanks to all the faculty coursera IBM DATA SCEINCE PROFESSIONAL CERTIFICATE COURSE. I thank Reviewing my assignment.**

**References**

1. 1.Mitchell, M. (2019) Programming Languages for Data Scientists. Medium. [https://towardsdatascience.com/programming-languages-for-data-scientists-afde2e af5cc5](https://towardsdatascience.com/programming-languages-for-data-scientists-afde2e%20af5cc5)
2. 2.Thanasis (2017) UK Road Safety: Traffic Accidents and Vehicles. Kaggle.com. <https://www.kaggle.com/tsiaras/uk-road-safety-accidents-and-vehicles>.
3. Savolainen P, Mannering F, Lord D, Quddus M. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accid Anal Prev. 2011;43:1666–76.
4. Depaire B, Wets G and Vanhoof K. Traffic accident segmentation by means of latent class clustering, accident analysis and prevention, vol. 40. Elsevier; 2008.
5. Labib, Farhan & Rifat, Ahmed & Hossain, Md & Das, Amit & Nawrine, Faria. (2019). Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh. 1-5. 10.1109/ICSCC.2019.8843640.
6. [1] “Nyc vision zero,” 2019, last access November 2019. [Online]. Available: https://www1.nyc.gov/site/visionzero/index.page
7. [2] “Vision zero scorecard and initiatives,” 2019, last access November 2019. [Online]. Available: <https://www1.nyc.gov/site/visionzero/initiatives/initiatives.page>
8. [3] M. M. Chong, A. Abraham, and M. Paprzycki, “Traffic accident analysis using decision trees and neural networks,” 2004.
9. [4] M. Bedard, G. H. Guyatt, M. J. Stones, and J. P. Hirdes, “The independent contribution of driver, crash, and vehicle characteristics to driver fatalities,” Accident Analysis and Prevention, vol. 34, no. 6, p. 717, 2002.
10. [5] M. Singleton, H. Qin, and J. Luan, “Factors associated with higher levels of injury severity in occupants of motor vehicles that were severely damaged in traffic crashes in kentucky, 2000- 2001,” Traffic Injury Prevention, vol. 5, no. 2, pp. 144–150, 2004.
11. [6] R. B. Hanrahan, P. M. Layde, S. Zhu, C. E. Guse, and S. W. Hargarten, “The association of driver age with traffic injury severity in wisconsin.” Traffic Injury Prevention, vol. 10, no. 4, pp. 361 – 367, 2009. [7] NOAA, “Readme file for daily global historical climatology network,” last access November 2019. [Online]. Availa