

car accident severity

Applied Data Science Capstone



lORENZO bRESCIA

FINAL REPORT - IBM DATA SCIENCE PROFESSIONA CERTIFICATE

OCTOBER 10th 2020

**Table of Contents**

[Introduction 2](#_Toc52985395)

[Business Understanding 3](#_Toc52985396)

[Objectives 5](#_Toc52985397)

[Finding Data Source 6](#_Toc52985398)

[Data Understanding and Data Preparation 7](#_Toc52985399)

[Data Visualization and Analysis 14](#_Toc52985400)

[Data Modelling and Deployment 20](#_Toc52985401)

[Data Analysis and Results 23](#_Toc52985402)

[Conclusion and Recommendations 25](#_Toc52985403)

[References 26](#_Toc52985404)

# Introduction

This assignment is the final project of the IBM Professional Data Science Certificate. It is meant for us to explore the skills we have learnt during the entire course. Furthermore, it will be a medium to test the functionalities of python programming language and get a grasp of data scientists’ everyday life.

The assignment will be centered around one theme: car accident severity. This topic may seem a little outdated, however, with the rising importance and extensity of the technologies that cars and other types of vehicles carry onboard, it is more relevant now than ever. Thanks to the technologies that we have today, we have access to data that we could not trace back then. That is why it is of the outmost importance to take advantages of these means and help solve certain issues or at least try to mitigate the risks involved in our everyday commutes.

# Business Understanding

Accidents occur every day. Specially, in roads and affect the lives of every member of the society... These events occur usually because of human behavior or machine failure. This represents, of course, an undesired event as accidents happen unexpectedly and unintentionally and can damage road infrastructure, people's health and people's properties. More than 38,000 people die every year in crashes on U.S. roadways. The U.S. traffic fatality rate is 12.4 deaths per 100,000 inhabitants. An additional 4.4 million are injured seriously enough to require medical attention. Road crashes are the leading cause of death in the U.S. for people aged 1-54. Internationally, nearly 1.3 million people die every year from car accidents. In fact, road traffic crashes rank as the 9th leading cause of death and accounts for 2.2% of all deaths globally.

Therefore, such problems, must be mitigated. They can be prevented if proper data is gathered and crucial variables are considered to analyze, recognize and classify why an accident occurs. Of course, this must be done through meticulous data analysis and modelling. For a proper analysis, first we make assumptions. We will need to gather all the information and knowledge available on the subject at hand and ask ourselves:

1. at what time in the day an accident is likely to occur?
2. accidents happen usually during weekdays or weekends?
3. in which area car accidents are more likely to occur?
4. on which kind of roads are accidents more likely to occur? are accidents more likely to occur at cross-roads or on highways?
5. do accidents occur more frequently among young/teenagers or adults/elderly? can we get any useful data from demographics analysis?
6. how does the variable "weather conditions" affect car accidents on average?

And so on...

So, we've described the scenario and the variables to look at are countless. Location, weather conditions, Road conditions, Light conditions, Day of the week, junction type and so on. The objective of this project is to propose a model to better handle this issue and try to solve or improve certain services such as:

1. safe route planning
2. emergency vehicle allocation
3. roadway design
4. where to place additional signage

After having evaluated the scenario/environment of car accidents let's talk about data in the next section

# Objectives

This assignment aims at exploring a dataset comprised of car accident data and creating a machine learning model around it.

Specifically, this assignment aims at:

* explore the dataset in detail and gather insights on car accidents and road safety
* analyze certain variables of interest in order to better understand their impact on car accidents
* suggest best-practices to act upon the discoveries made from the analysis

# Finding Data Source

The source of data is the first step of the way. It is crucial to find proper sources of data for your analysis. The assignment will focus on data regarding car accidents in the city of Leeds, UK. The dataset was retrieved directly from the UK government’s website:

<https://data.gov.uk/dataset/6efe5505-941f-45bf-b576-4c1e09b579a1/road-traffic-accidents/datafile/c885f604-ee28-418d-a988-160cde514756/preview>

This dataset includes a set of information regarding car accidents in Leeds in the year 2019. Pretty good, since it is up to date.

# Data Understanding and Data Preparation[¶](https://render.githubusercontent.com/view/ipynb?commit=d3785a91f23e9d434f437d00e1d0a0743ecd9aa7&enc_url=68747470733a2f2f7261772e67697468756275736572636f6e74656e742e636f6d2f6c6f72656e7a6f37382f6769746875622d6578616d706c652f643337383561393166323365396434333466343337643030653164306130373433656364396161372f412532306465736372697074696f6e2532306f6625323074686525323064617461253230616e64253230686f77253230697425323077696c6c2532306265253230757365642e6970796e62&nwo=lorenzo78%2Fgithub-example&path=A+description+of+the+data+and+how+it+will+be+used.ipynb&repository_id=238065045&repository_type=Repository#A-description-of-the-data-and-how-it-will-be-used)

The dataset at hand consists of many different attributes. Below the list of attributes and their data types and description is outlined.

|  |  |  |
| --- | --- | --- |
| Attribute Name | Data Type | Description (if any) |
| Reference Number | Object | Identification number of casualty |
| Grid Ref: Easting | Int64 | UTM coordinates |
| Grid Ref: Northing | Int64 | UTM coordinates |
| Number of Vehicles | Int64 | Number of vehicles involved |
| Accident Date | Object | Date of the accident |
| Time (24hr) | Int64 | Time of the accident |
| 1st Road Class | Int64 | Type of roads:  1 Motorway  2 A(M)  3 A  4 B  5 C  6 Unclassified |
| 1st Road Class & No | Object | Name of road |
| Road Surface | Int64 | Road conditions during accident:  1 Dry 2 Wet / Damp 3 Snow 4 Frost / Ice 5 Flood (surface water over 3cm deep) |
| Lighting Conditions | Int64 | Light conditions during accident:  1 Daylight: streetlights present  2 Daylight: no street lighting  3 Daylight: street lighting unknown  4 Darkness: streetlights present and lit  5 Darkness: streetlights present and unlit  6 Darkness: no street lighting  7 Darkness: street lighting unknown |
| Weather Conditions | Int64 | Weather conditions during accident:  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 |
| Local Authority | Object | Leeds local authority code |
| Vehicle Number | Int64 | Number of vehicles involved |
| Type of Vehicle | Int64 | Type of vehicle involved:  1. Pedal cycle  2. M/cycle 50cc and under  3. Motorcycle over 50cc and up to 125cc  4. Motorcycle over 125cc and up to 500cc  5. Motorcycle over 500cc  6. [Not used]  7. [Not used]  8. Taxi/Private hire car  9. Car  10. Minibus (8 – 16 passenger seats)  11. Bus or coach (17 or more passenger seats)  12. [Not used]  13. [Not used]  14. Other motor vehicle  15. Other non-motor vehicle  16. Ridden horse  17. Agricultural vehicle (includes diggers etc.)  18. Tram / Light rail  19. Goods vehicle 3.5 tonnes mgw and under  20. Goods vehicle over 3.5 tonnes and under 7.5 tonnes mgw  21. Goods vehicle 7.5 tonnes mgw and over  22. Mobility Scooter  90. Other Vehicle  97. Motorcycle - Unknown CC |
| Casualty Class | Int64 | Driver information: 1. Driver or rider  2. Vehicle or pillion passenger  3. Pedestrian |
| Casualty Severity | Int64 | Severity of the accident:  1. Fatal  2. Serious  3. Slight |
| Sex of Casualty | Int64 | Sex of casualty:  1. Male  2. Female |
| Age of Casualty | Int64 | Age of casualty given in years |

For the purpose of the project I have imported a series of libraries that allow to make certain data manipulations. In Figure 1 a first summarized view of the data set is shown

Graphical user interface, application, table

Description automatically generatedFig.1 Data frame after uploading csv file.

In Figure 2 I run the describe function which aims at giving general information about the dataset. It computes a summary of statistics pertaining to the data frame columns. This function gives the count, mean, standard deviation and Interquartile Range values.

A picture containing graphical user interface

Description automatically generatedFig. 2 Output after running describe() function.

In Figure 3 I run a specific function to derive and evaluate all attributes’ data types.  
  
Table

Description automatically generated

Fig. 3 Output after running dtypes function.

In figures 4 and 5 I run other ordinary function to be able to get the full picture of the dataset. Checking for null values (the sum of null values for each attribute) data types and row count for each variable. After having prepared the data and evaluated the nature of the data frame I have set the basis for the next phase.

Graphical user interface, table

Description automatically generated

Fig. 4 Output after running info() function.

A picture containing graphical user interface, text

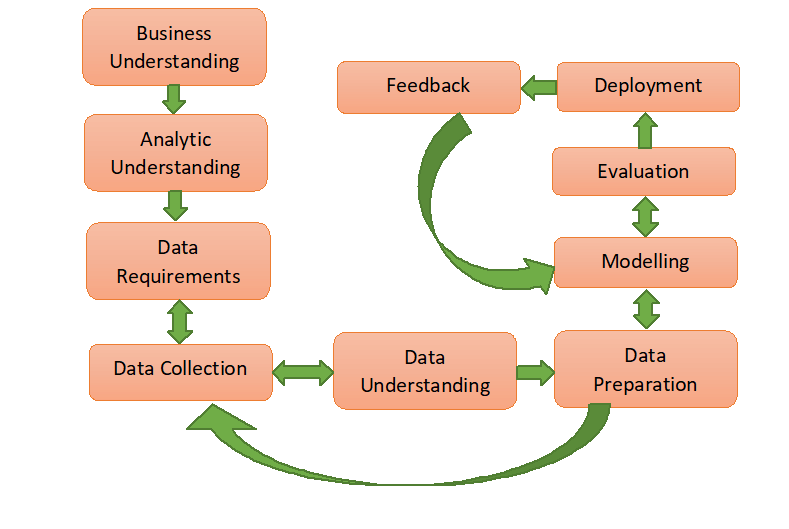
Description automatically generated

Fig. 5 Output after running isnull().sum() function.

Data Science Methodology

When it comes to Data analysis there are many steps to take into consideration such as:

1. **data collection**
2. **data preparation/exploration**
3. **data modelling**
4. **data evaluation**
5. **data deployment**

Fig. 6 Data Science Methodology

#### Data Collection

As mentioned in the previous section the data has been searched in google. Eventually, I landed on the UK government website and found a useful dataset. I downloaded the datasets with attached a guidance word document that explained what each variable meant. This allowed for proper data understanding.

Once imported all the necessary libraries (pandas, matplotlib, NumPy, seaborn and others) and the data needed (in this case we will use a csv file imported directly in the notebook) data was ready to be observed and, if necessary, cleaned from raw data. One step more is to look for data types in your datasets. After, the data will be manipulated, filtered and prepared in such a way that will favor the data modelling stage. After tweaking a little bit, the dataset, it was finally ready for analysis.

The data will be used for the purpose of suggesting proper methods to mitigate the risk of car accidents based on machine learning algorithms built with the variables that most impact road safety and car accidents occurrence.

#### Data Preparation/Exploration

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data (standardizing data formats, enriching source data, and/or removing outliers). In the dataset null values or missing values will be removed, existing data correlation will be analyzed and more manipulation on data will be performed.

#### Data Modelling

At this stage we will implement machine learning algorithms to create and evaluate models. Here is the list of commonly used machine learning algorithms. These algorithms can be applied to almost any data problem:

1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. SVM
5. Naive Bayes
6. kNN
7. K-Means
8. Random Forest
9. Dimensionality Reduction Algorithms
10. Gradient Boosting algorithms (GBM, XGBoost, LightGBM, CatBoost)

Some of these algorithms will be tested and accuracy of each will be evaluated.

#### Data 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.

#### Data Deployment

The concept of deployment in data science refers to the application of a model for prediction using a new data. Building a model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data science process. For this project I have used the data science programming language named Python. According to recent studies, Python is the preferred programming language for data scientists. They need an easy-to-use language that has decent library availability and great community participation. Python features lots of machine learning libraries to choose from and easy to access.

# Data Visualization and Analysis

This section aims at showing the results obtained from the initial data analysis. Visualization techniques have been used to address the objectives of this research. During the evaluation stage I decided to focus on a specific set of variables. These variables are:

* 1st Road Class
* Weather Conditions
* Lighting Conditions
* Road Surface
* Type of Vehicle
* Number of Vehicles
* Age of Casualty
* Sex of Casualty
* Casualty Class
* Casualty Severity

These attributes were chosen to be best for predicting casualty severity. In fact, as one may think, type of road, weather conditions, lighting conditions, road surface etc. during the casualty could be good indicators for casualty severity.

The libraries used for data analysis are:

1. import pandas as pd
2. import numpy as np
3. import os
4. import matplotlib.pyplot as plt
5. import seaborn as sns
6. import itertools
7. import Null Formatter from matplotlib.Ticker
8. import matplotlib.ticker as ticker
9. import preprocessing from sklearn

**Chart, waterfall chart

Description automatically generated**

Fig. 7 Number of accidents by: lighting, weather and road surface conditions

Chart

Description automatically generatedFig. 8 Number of accidents by: vehicle type, road and casualty class

Chart, bar chart

Description automatically generated

Fig. 9 Number of accidents by age of casualty

Chart, bar chart

Description automatically generatedFig. 10 Number of accidents by sex of casualty

Chart, bar chart

Description automatically generated

Fig. 11 Number of accidents by number of vehicles

Fig. 12 Number of accidents grouped by Casualty Severity and: weather, lighting and road surface conditions, vehicle type, road class and number of vehicles

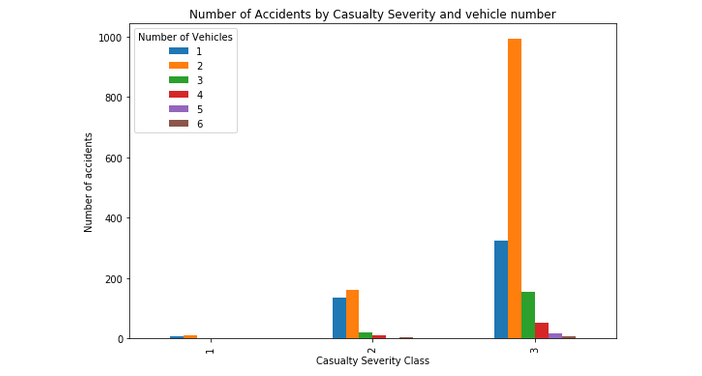
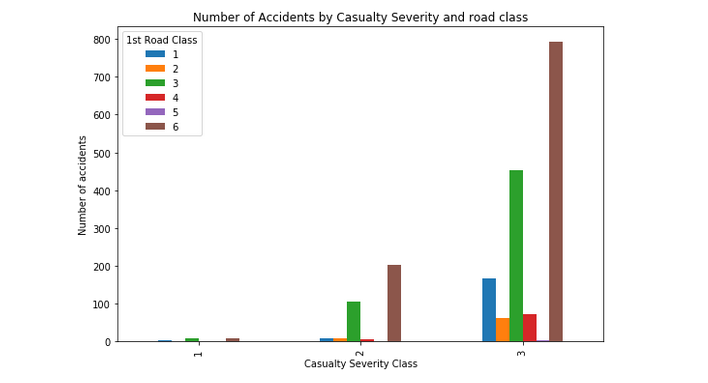
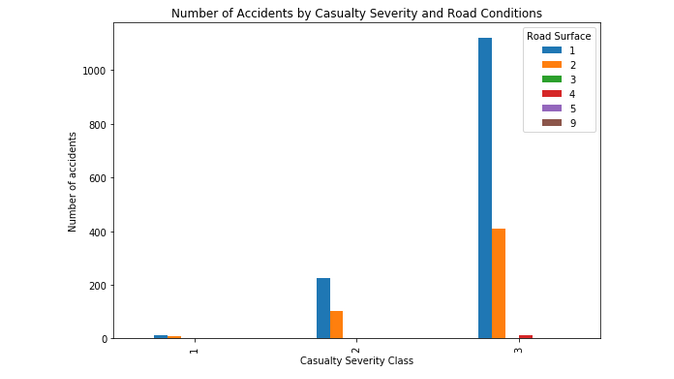
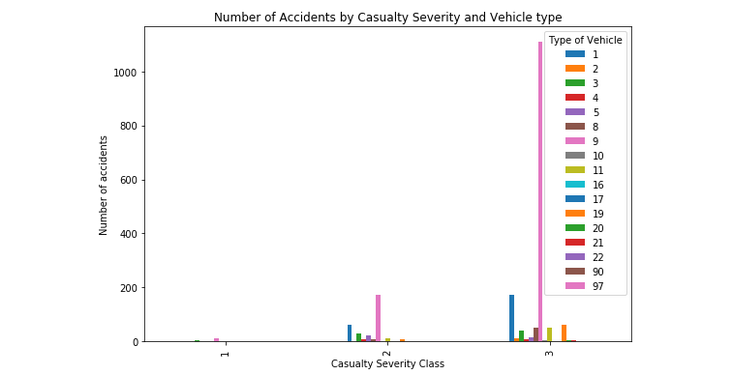
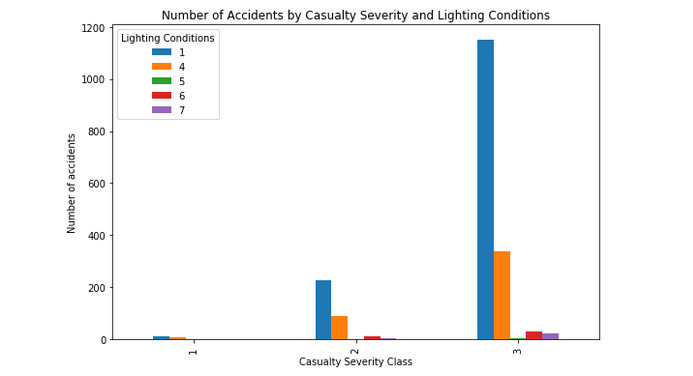
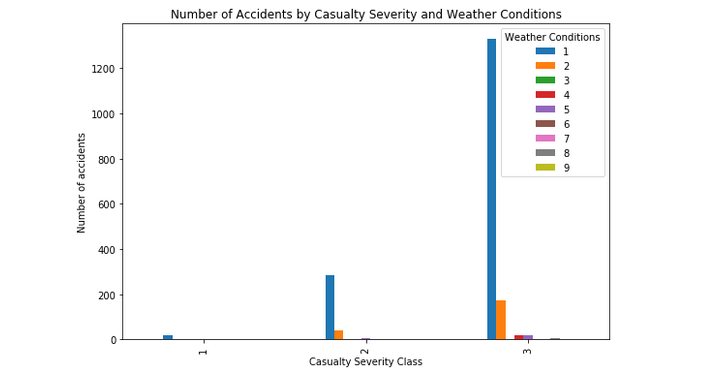
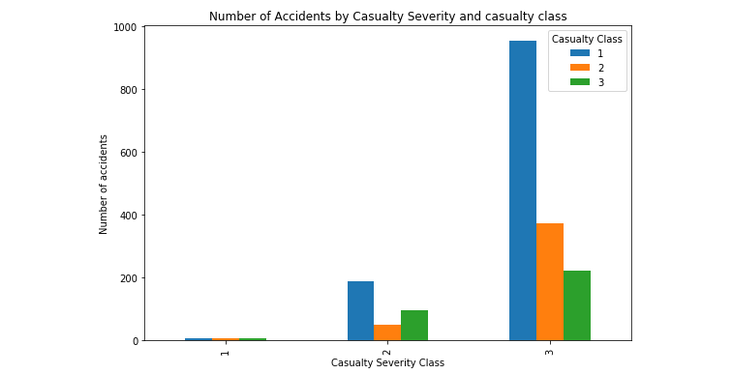
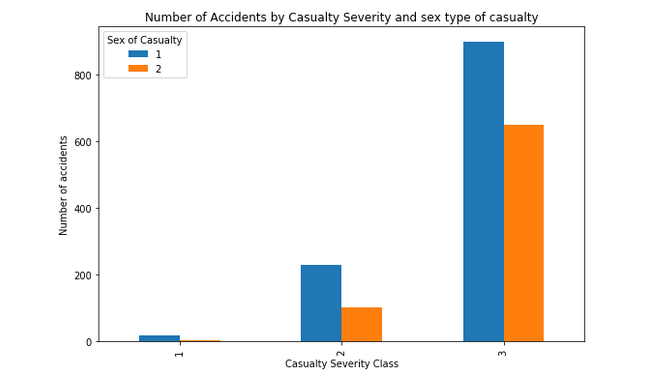


Fig. 13 Number of accidents grouped by Casualty severity and: sex type and casualty class



# Data Modelling and Deployment

For this part of the project the dataset has been split to train and test samples. Train/Test Split involves splitting the dataset into training and testing sets respectively, which are mutually exclusive. After this operation, the model is trained with the training set and tested with the testing set. This will provide a more accurate evaluation on out-of-sample accuracy because the testing dataset is not part of the dataset that have been used to train the data. It is more realistic for real world problems.

In this specific case two specific type of classification models called K nearest neighbor and decision tree. Therefore, following libraries have been imported:

1. from sklearn.model\_selection import train\_test\_split
2. from sklearn.neighbors import KNeighborsClassifier
3. from sklearn import metrics
4. from sklearn.metrics import classification\_report
5. from sklearn.metrics import f1\_score
6. from sklearn.tree import DecisionTreeClassifier

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN in this project has been used in statistical estimation and pattern recognition. Below the code that has been developed for the purpose of the analysis is shown with a summary of the accuracy depending on the differen K values chosen. Accuracy for KNN has been evaluated setting a dynamic K that ranges from 1 to 20. The best accuracy was 0.8141361256544503 with k= 13. After having trained the model is time to test the model. *Train Set Accuracy* and *Test Set Accuracy* have been evaluated.

The same procedure has been utilized with Decision tree modeling. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The result is a tree with decision nodes and leaf nodes.

Fig. 14 KNN classification model

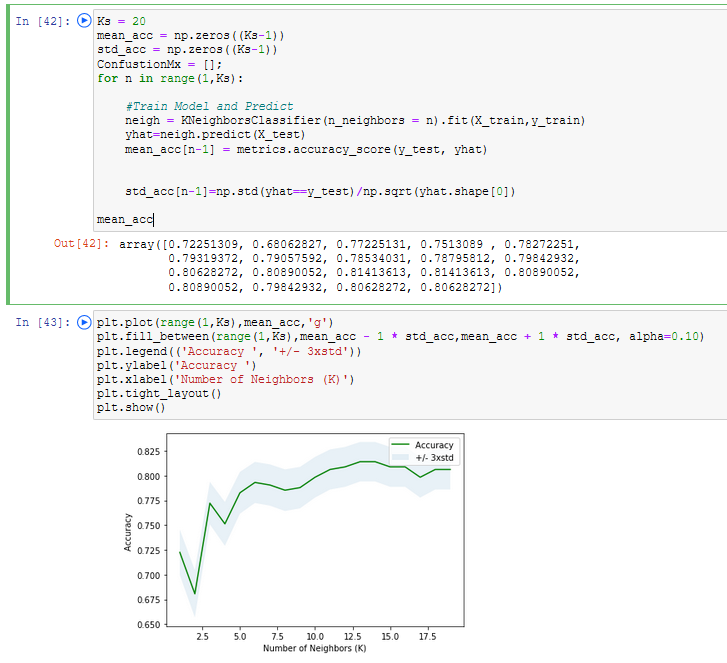
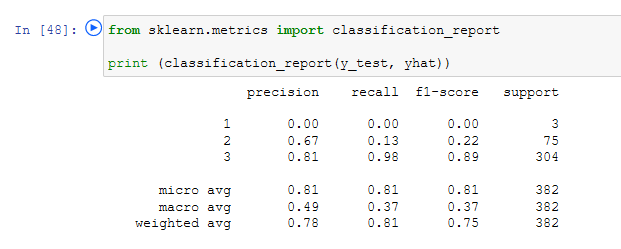
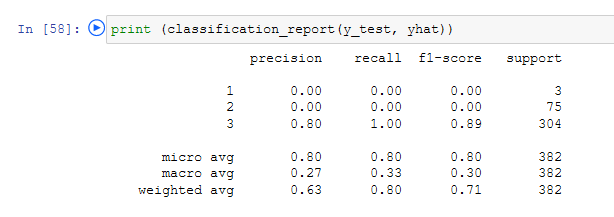


Fig. 15 Decision Tree classification model



# Data Analysis and Results

This section has the objective of taking into consideration all the analysis that has been done and give some main takeaways.

The main findings were:

1. For the variable ‘Lighting Conditions’ we can see that 73% (1389 out of 1907) of car accidents during the year 2019 occurred under ‘Lighting Conditions’ class 1, ‘Daylight, streetlights present’, and of those 73% , 0.9% were fatal, 16.1% were serious and 83% were of light severity. Another crucial 23% (436 out of 1907) of car accidents occurred under class 4, ‘Darkness, streetlights present and lit’, and of those 23%, 2% were fatal,21% were serious and the other 77% were of light severity class.

Class 6 (Darkness: no street lighting) counts 43 accidents out of 1907 (, class 7 (street lighting unknown) counts 30 out of 1907 and class 5 (Darkness: streetlights present and unlit) counts 9 out of 1907.

1. For the variable ‘Weather Conditions’ we unexpectedly observe that a total of 86% (1633 out of 1907) of car accidents occurred under good weather (class 1 🡪 ‘Fine, without high winds’). Of those 86%, 1.2% were fatal, 17.4% were serious and another 81.4% were of light severity. Another 11% (215 out of 1907) of car accidents occurred while raining but without strong winds (class 2). Of those 11%, 1% were fatal, 19% were serious and the other 80% was of light severity. Then 1.3% occurred while there was rain with high winds (of which 20% being serious accidents and the rest light accidents) and another 1% occurred while snowing without high winds (class 4) of which 10% being serious accidents and 90% being light severity accidents. The other class are almost non relevant for this analysis.
2. For ‘Road Surface’, namely, the conditions of the road surface at the time of the accident, we can observe that 99% of accidents fell under classes 1,2 and 4. 71% of accidents fell under class 1 (dry road surface), 27% under class 2 (wet/damp road surface) and another 0,9% under class 4 (frost/Icy road surface). Of the 71% that fell under class 1, 1% was of fatal nature, 16.5% was of serious nature and the rest was of light severity class.
3. Regarding the variable ‘Type of Vehicle’, 80% of the accidents involved cars or pedal cycle. Respectively 68% for cars (class 9) and 12% for pedal cycle (class 1). The rest of the accidents are homogeneous among the other vehicle types.
4. The variable ‘1st Road Class’ sees 92% of the accidents pertaining to classes 6,3 and 1. Respectively most of the accidents occur on unclassified roads, 53% to be specific (0.8 of which fatal, 20% were serious and the rest were light accidents. Another 30% occur on roads classified as A (1.6% were fatal, 18% were serious and the rest were light severity accidents) and only 9% on motorways (1.6% were fatal, 5% were serious and the rest light severity accidents).
5. With respect to the variable ‘Casualty Class’ 60% of the accidents involved a driver. The rest involved either a pillion passenger or a pedestrian.

# Conclusion and Recommendations

For discussion and recommendation purposes, we will refer to the KNN predictive model as it yielded the highest accuracy score for predicting the car accident severity code given the variables chosen. As previously mentioned, the observations showed that most of car accidents occur under normal and optimal conditions for driving. In fact, accidents mainly occurred under good road conditions (Dry Road), nice weather (Fine, without high winds), good lighting conditions (Daylight, streetlights present). Maybe some variables could be added to better predict car accidents probability, since these variables do not seem to be so explanatory for the dataset at hand. **Traffic conditions** in the specific road section where the accident occur could add some accuracy and precision, as well as **attention level** (even if it is not an easy variable to measure) of the driver, speed of the car at the time of crash (but that requires computational systems in place at the specific location), collision and junction type. Also expanding the severity classes could help.

To conclude the research was conducted well and hopefully it will give some guidance to readers, be them institutions, pythoneers or drivers, to grasp better the different variables that come handy in detecting car accidents and determining their severity class.

# References

How to plot a count bar chart with a Pandas DF, g., Boston, S., & RAO, U. (2020). How to plot a count bar chart with a Pandas DF, grouping by one categorical column and colouring by another. Retrieved 6 October 2020, from <https://stackoverflow.com/questions/48939795/how-to-plot-a-count-bar-chart-with-a-pandas-df-grouping-by-one-categorical-colu>

Matplotlib. (2020). Retrieved 6 October 2020, from <https://www.southampton.ac.uk/~feeg1001/notebooks/Matplotlib.html>

matplotlib.pyplot.show — Matplotlib 3.1.2 documentation. (2020). Retrieved 6 October 2020, from <https://matplotlib.org/3.1.1/api/_as_gen/matplotlib.pyplot.show.html>

Python Pandas - Descriptive Statistics - Tutorialspoint. (2020). Retrieved 6 October 2020, from <https://www.tutorialspoint.com/python_pandas/python_pandas_descriptive_statistics.htm>

VanderPlas, J. (2016). *Python Data Science Handbook*. O'Reilly Media.

Hall, T., & Stacey, J. (2013). Python 3 for absolute beginners. New York: Apress.