**AI Report**

**rtificial Intelligence and Machine Learning**

**The Factors affecting Accident Severity in the 2016 Traffic Report**



**BSc (Hons) Computer Science:**

**Callum Melrose**

**Evi Nikolaidou**

**Marios Limnatis**

Word count: 3835 (not including the title page, titles, references, tables or captions)

# 1.0 - Abstract

This report explores which machine learning method is best suited to predict the road traffic accident severity based on a number of factors. The dataset consisted of all road traffic accidents from 2016. Three methods of machine learning were used: Naïve-Bayes, K-Nearest Neighbors and Regression. The results found that all the factors described in the questions affected the accident severity and that Naïve-Bayes was the best machine learning algorithm to use for the selected dataset, though more research could be done into other machine learning algorithms.

Table of Contents

[1.0 - Abstract 2](#_Toc504361861)

[2.0 - Introduction 5](#_Toc504361862)

[2.1 - Background 5](#_Toc504361863)

[2.2 - Aim 5](#_Toc504361864)

[2.3 - Objectives 5](#_Toc504361865)

[3.0 - Dataset Description 6](#_Toc504361866)

[4.0 - Questions 7](#_Toc504361867)

[5.0 - K-Nearest Neighbors by Callum 8](#_Toc504361868)

[5.1 - Summary of the approach 8](#_Toc504361869)

[5.2 - Data pre-processing, visualisation, feature selection 8](#_Toc504361870)

[5.3 - Model training, evaluation and testing 8](#_Toc504361871)

[5.4 - Results and discussion 9](#_Toc504361872)

[5.5 - Conclusion and recommendations 10](#_Toc504361873)

[6.0 - Naïve Bayes by Evi 11](#_Toc504361874)

[6.1 - Summary of the approach 11](#_Toc504361875)

[6.2 - Data pre-processing, visualisation, feature selection 12](#_Toc504361876)

[6.3 - Model training, evaluation and testing 13](#_Toc504361877)

[6.4 - Results and discussion 13](#_Toc504361878)

[6.5 - Conclusion and recommendations 15](#_Toc504361879)

[7.0 – Regression analysis by Marios 16](#_Toc504361880)

[7.1 - Summary of the approach 16](#_Toc504361881)

[7.2 - Data pre-processing, visualisation, feature selection 16](#_Toc504361882)

[7.3 - Model training, evaluation and testing 16](#_Toc504361883)

[7.4 - Results and discussion 17](#_Toc504361884)

[7.5 - Conclusion and recommendations 19](#_Toc504361885)

[8.0 - Results, comparison and discussion 20](#_Toc504361886)

[9.0 - Conclusion and recommendations 21](#_Toc504361887)

[10.0 - References 22](#_Toc504361888)

[11.0 – Appendices 23](#_Toc504361889)

[Appendix A 23](#_Toc504361890)

[Appendix B 23](#_Toc504361891)

[Appendix C 23](#_Toc504361892)

[Appendix D 23](#_Toc504361893)

[Appendix E 23](#_Toc504361894)

[Appendix F 24](#_Toc504361895)

[Appendix G 24](#_Toc504361896)

# 2.0 - Introduction

## 2.1 - Background

Machine learning is unlike other forms of programming, where a program is made to do exactly what the programmer wants. Instead, machine learning applies different methods and principles that give the machine the ability to learn and thus make predictions. Again, unlike other types of programming, machine learning algorithms can be trained to improve their accuracy. How this training occurs varies dependant on the method of machine learning employed.

Though machine learning was originally primarily used for data mining, there are many uses. One way in which machine learning can be used is to notice the trends and relationships in large quantities of data. This is particularly useful as it would require a large amount of time with the potential for human errors if a person was to do this (Kotsiantis, 2007), thus machine learning can be used to save time and produce error free results.

In this report, three methods of machine learning were used to process data on road traffic accidents.

## 2.2 - Aim

The aim of the report is to evaluate the factors which affect road traffic accidents using machine learning algorithms.

## 2.3 - Objectives

The objectives are to:

* Create three different machine learning algorithms to explore the factors affecting accident severity.
* Compare and contrast the results of the machine learning algorithms
* Evaluate the best machine learning algorithm to use for each question
* Create a presentation giving a brief explanation of the group’s findings

# 3.0 - Dataset Description

The dataset used was selected from the Department for Transport, detailing all road traffic accidents in the UK in 2016. The description for the dataset itself states that it lists all ‘personal injury accidents on public roads that are reported to the police’ (Department for Transport, 2017), so accidents that are not reported to the police or occur on private roads will not be included in the dataset.

This dataset was selected for two major reasons. The first reason was due to the size, with it consisting of 136622 rows and 32 columns. This allowed for numerous pieces of data which could be sampled without the worry of using the same part of the data numerous times. The second reason this dataset was chosen was because it gave the group a choice in the selection of the different factors that could be used to train artificial intelligence.

The group chose to explore the factors affecting the accident severity rating, using the machine learning algorithms developed to find potential links between these factors and the rating.

Due to the sheer number of columns on the dataset, a number of factors had no impact on the outcome of the severity rating (e.g. accident index, the local highway authority, did a police officer attend scene of the crime). These factors were ignored by dropping them from the sample set (as seen in **Appendix A**). There were also factors which could have an effect on the severity rating but weren’t relevant to the group’s questions which were excluded from the sample set (as seen in **Appendix B**).

# 4.0 - Questions

After analysing the dataset as a group, we decided to use the severity rating as the classifier for the questions we would pose.

Three questions were raised, these were:

1. Do the number of vehicles and casualties in a traffic accident affect accident severity rating?
2. Do the light and weather conditions affect the severity rating in a traffic accident?
3. How do numerous factors affect the traffic accident severity rating?

The ‘numerous factors’ mentioned in the third question refer to the following conditions:

* Road Type
* Speed Limit
* Road Surface Conditions
* Light Conditions
* Weather Conditions
* Number of Vehicles
* Number of Casualties

# 5.0 - K-Nearest Neighbors by Callum

## 5.1 - Summary of the approach

The K-Nearest Neighbors (often shortened to k-NN) is a machine learning algorithm that finds K closest ‘neighbors’, where K is a positive integer, in the feature space and predicts which group the data group belongs (Keller et. al, 1985). The algorithm splits the data it is given into ‘training’ and ‘testing’ data, using the training data to make predictions with the testing data. K-Nearest Neighbors is also a non-parametric method, meaning that no assumptions are made about the probability distribution of the input.

This method was chosen due a number of reasons, one such reason being that it is not sensitive to extreme outliers. This is due to the fact that the algorithm checks the data points that have the least Euclidean distance between them, reducing the impact of extreme outliers as they will rarely be the closest to the chosen data point. It should also be noted that having a high value for K also reduces the impact of extreme outliers (e.g. if K = 100, 1 outlier would only impact the prediction by 1%). Another reason K-Nearest Neighbors was chosen was due to the fact that it makes no assumptions about the dataset, making it an interesting comparison to Naïve-Bayes, which was chosen by another group member and assumes that attributes are independent. Finally, K-Nearest Neighbors is simple in comparison to other machine learning algorithms (Weinberger and Saul, 2009), making it an ideal algorithm to learn the fundamentals and gain understanding of machine learning.

To answer our research questions the K-Nearest Neighbors algorithm was used to predict the accident severity of traffic accidents.

## 5.2 - Data pre-processing, visualisation, feature selection

Only one change was made to the dataset prior to it being read by the program, this was moving the accident severity to the last column. This was done primarily to increase the clarity of the dataset, allowing the group to see classifier clearly on the last column.

As mentioned in section 3.0, many columns were dropped from the dataset in the code (as seen in **Appendix A** and **Appendix B**). Any entries in the dataset that had data left blank were ignored so as not to cause any outlandish results or anomalies when training and testing the sample data (see **Appendix C**).

## 5.3 - Model training, evaluation and testing

The test sizes for all three questions were set to 40% to keep computational time reasonable.

One of the most important parts of machine learning is calculating the accuracy of the predictions. This was done initialising two variables named ‘correct’ and ‘total’ as 0 (see **Appendix D**). For every correct prediction ‘correct’ is incremented by 1 and for every prediction ‘total’ is also incremented by 1 (see **Appendix E**). The accuracy is then calculated by dividing ‘correct’ by ‘total’ and rounding it to the nearest two decimal places before it is printed (see **Appendix F**).

The confidence was also calculated to give insight on the how the algorithm made the decision for each prediction (see **Appendix G**).

One way in which the algorithm was fine-tuned was by finding the best value for K. Naturally this varied from question to question so the value for K differed for each situation (with K = 600 for the first and second question and K = 200 for the third question).

A sample size of 4000 was selected at random from the dataset for each question (see **Appendix H**), alleviating computational time by not using the full dataset, which could take hours to run. The random row from which the data was read in the dataset was determined by the variable named ‘randRow’, which creates a random integer between 1 and the number of rows in the dataset minus the sample size (to prevent the last entry being loaded and returning an error).

Table 1 – Finding K for the second question

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| K | 5 | 10 | 20 | 50 | 100 | 200 | 400 | 600 | 1000 |
| Average accuracy | 6% | 10% | 16% | 28% | 34% | 39% | 36% | 84% | 79% |

Table 1 shows an attempt at finding the optimum value for K for the second question.

## 5.4 - Results and discussion

Three tables were created for each question to show the results of 10 runs, with an average included so the results can be compared.

Table 2 - Do the number of vehicles and casualties in a traffic accident affect accident severity rating? (k = 600)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 | Average |
| Runtime | 46.36 | 46.3 | 45.98 | 46.21 | 46.07 | 46.1 | 45.93 | 46.02 | 46 | 45.98 | 46.1 |
| Confidence | 79.17 | 84.67 | 60.5 | 73.83 | 71.5 | 76.67 | 68.83 | 71.67 | 71.83 | 81.67 | 74.03 |
| Accuracy | 83.62 | 91.31 | 85 | 80.69 | 79.56 | 82.81 | 85.69 | 82.06 | 81.06 | 86.25 | 83.81 |

Table 3 - Do the light and weather conditions affect the severity rating in a traffic accident? (k = 600)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 | Average |
| Runtime | 47.77 | 46.28 | 46.24 | 46.24 | 45.99 | 46.13 | 45.68 | 46.03 | 46 | 46.16 | 46.25 |
| Confidence | 53.17 | 81.33 | 54 | 53.17 | 49.17 | 89.13 | 54.17 | 48.5 | 55.83 | 55.83 | 59.43 |
| Accuracy | 83 | 88.88 | 78.56 | 81.25 | 69.62 | 92.81 | 73.94 | 31.81 | 80.44 | 83.5 | 76.38 |

Table 4 - How numerous factors affect the traffic accident severity rating (k = 200)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 | Average |
| Runtime | 48.36 | 48.22 | 48.1 | 48.07 | 48 | 47.39 | 47.94 | 47.59 | 47.9 | 48.14 | 47.97 |
| Confidence | 77 | 84.5 | 82.5 | 53.5 | 71 | 85.5 | 77 | 69.5 | 81 | 76 | 75.75 |
| Accuracy | 79.5 | 78.5 | 85.24 | 70.25 | 81.19 | 83.38 | 80 | 80.62 | 85.81 | 80.19 | 80.47 |

Table 2 shows that the K-Nearest Neighbors algorithm had the highest accuracy for the first question. All three tables show a similar average runtime, with the third question taking an extra 1.72 seconds to complete than the second question. Table 3 highlights that the results for the second question had a significantly lower accuracy in comparison to the other two questions, though this appears to have been caused by an extreme result in the 8th run.

## 5.5 - Conclusion and recommendations

In conclusion to the results in the previous section, all of the factors in each of the questions seem to affect the accident severity rating. Table 2 suggests that due to having the highest accuracy, the number of vehicles and casualties in an accident have the most impact on the severity rating. Whereas the light and weather conditions seem to have the smallest impact on severity rating, it should be noted that these results have a low confidence rating and more testing may need to be done.

The abnormally high optimum values for K could also suggest that normalisation is occurring with the results and thus, it may be possible that this is having a bigger impact on the accuracy than the factors aforementioned in the questions. More research could be carried out to determine whether this is the case.

# 6.0 - Naïve Bayes by Evi

Naïve Bayes classifier represents a supervised learning method that is based on applying Bayes Theorem. This classifier is named after Thomas Bayes (1702-1761) who proposed the Bayes Theorem, which is built on conditional probability. It assumes that there are no dependencies among attributes, therefore it calculates the probabilities of every factor. It is names “Naïve” because it simplifies the computations that are involved. Naïve Bayes classification can help us calculate the probability of something that will happen in the future based on something else that already happen.

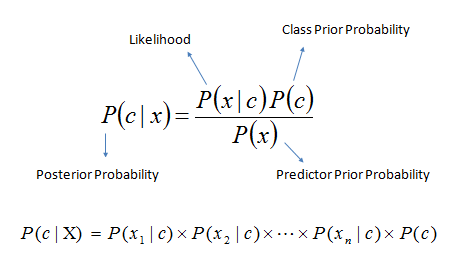


Figure 1: Bayes Theorem

Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). In the above equation P(c|x) is the probability of c given that x happens, P(x|c) is the probability of x given that c happens and P(c) is the prior probability of class and P(x) is the prior probability of predictor.

## 6.1 - Summary of the approach

Alike all classifiers Naïve Bayes has both advantages and disadvantages.

Advantages:

* It’s easy and fast when making decisions.
* It performs well in multi-class prediction
* There is no need for large amount of data before the learning process begin
* Performs better in case of categorical input variables than numerical input variables

Disadvantages:

* Naïve Bayes assumes that all predictors are independent which in real life isn’t always true
* Zero Frequency: If data set has a category with categorical variable that was not observed in training dataset then the model will automatically assign zero probability and therefore will be unable to make prediction

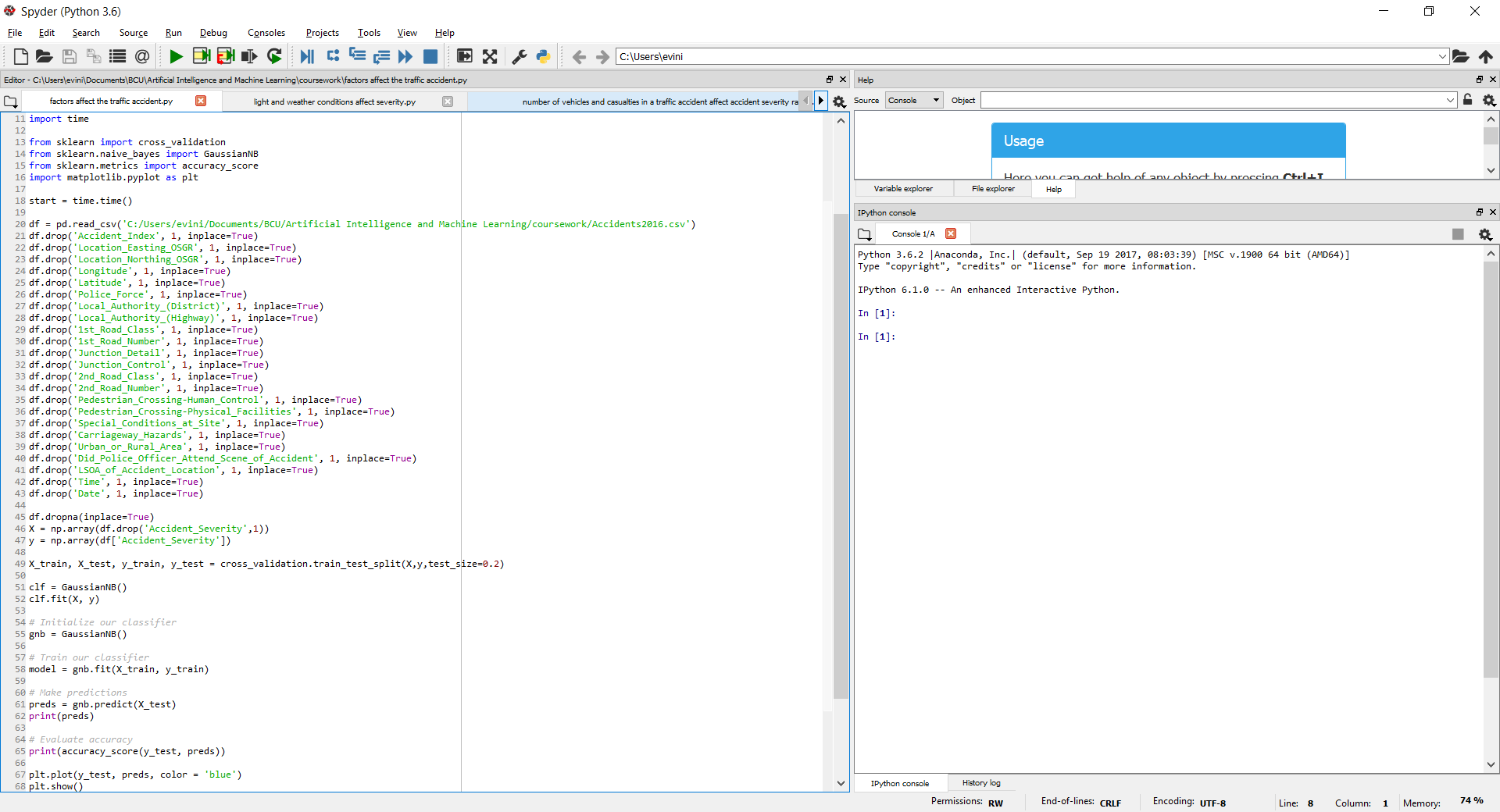
There are three types of Naïve Bayes Algorithm:

1. Guassian Naïve Bayes
2. MultiNomial Naïve Bayes
3. Bernoulli Naïve Bayes

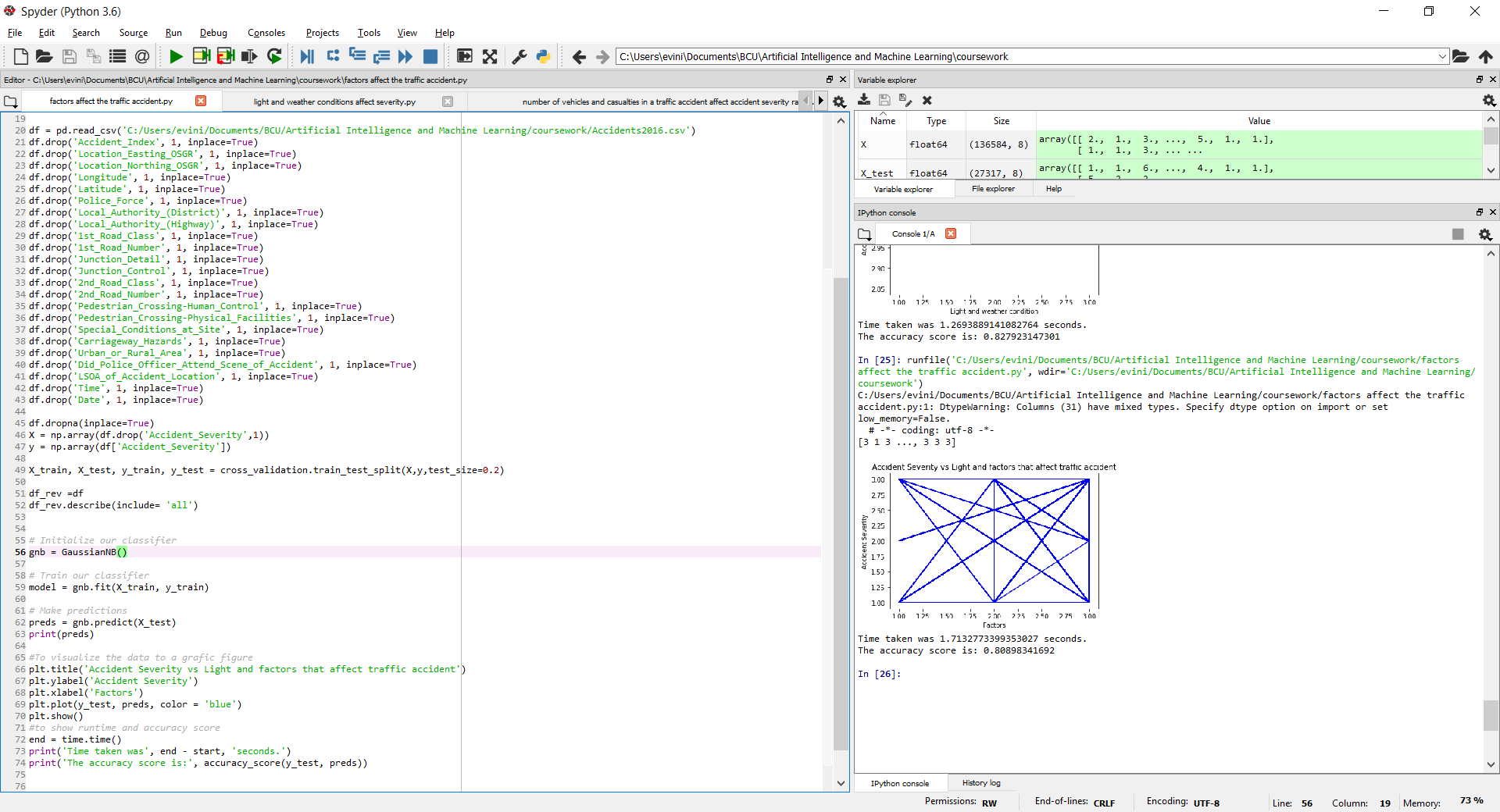
I choose the Guassian Naïve Bayes classifier because I believe that it’s the most appropriate method to answer the questions my group assignment. It provides straightforward probabilistic prediction and it’s very fast. The chosen dataset has very well separate categories and data are not complex. Last, keeping in mind that the dataset has many classes and that Naïve Bayes algorithm is well known for its ability to predict the probability of multiple classes of a target variable, Naïve Bayes classifier should be a suitable method for our problem.

## 6.2 - Data pre-processing, visualisation, feature selection

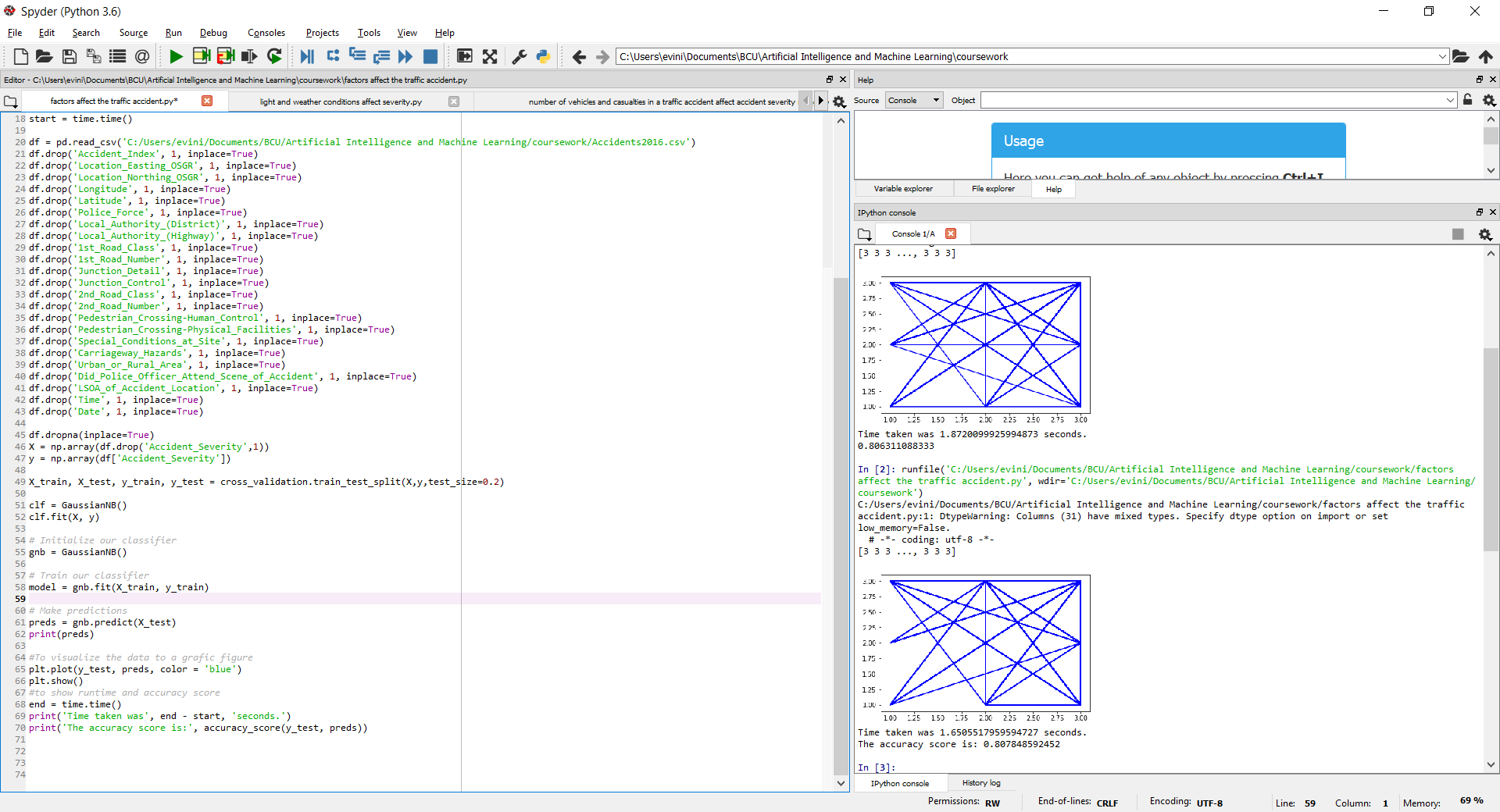
Firstly we import the dataset of our choice, in this case the details of Car Accidents during 2016 to a python file using the pd.read\_csv command. Then after careful consideration we decide to not make any changes to the actual dataset but rather drop all unnecessary columns in our code. Here is an example:



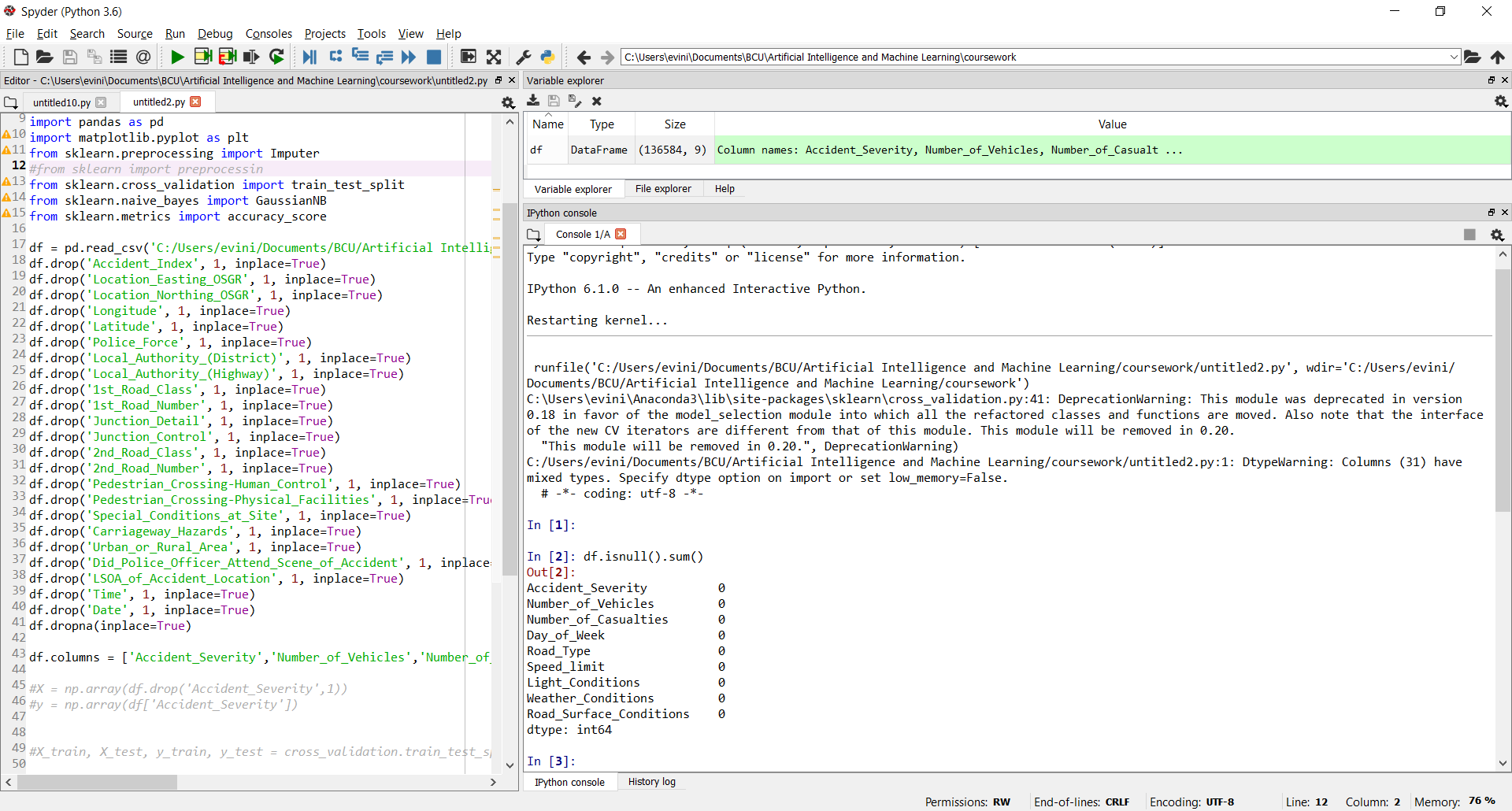
After that we split data into train and test data. To answer our questions we chose to use 20% of data as test and 80% as train.



We choose to answer three questions, therefore we created three python files, one for each question. To visualize the data, we use graphic figures that compare: (a) the accident severity with the number of vehicles and casualties in a traffic accident, (b) the accident severity with the light and weather conditions of a traffic accident and (c)the accident severity with the factors that affect the traffic accident severity. Therefore, the code in each file give the visualization of graphic figures, the accuracy score and the runtime of each question. To do that we did the following:



Moreover, the dataset should not include any null values, but due to the large amount of entries in the dataset its almost impossible to check that manually. Thankfully using the: isnull().sum() command we made sure that there are no null values.



## 6.3 - Model training, evaluation and testing

To implement Naïve Bayes algorithm we need to import the following line: from sklearn.naive\_bayes import GaussianNB then we create a classifier: gnb = GaussianNB()

and then we “fit” it and give the training data. So, we call the fit function with to arguments X\_train and y\_train: model = gnb.fit(X\_train, y\_train) . This is always going to be truth and last we are asking for predictions: preds = gnb.predict(X\_test)

As you can see in the following section the accuracy of the predictions in all questions is in high levels. Accident severity can should only have value 1,2 or 3 because in the dataset there no other entries. Accident severity has only three levels, which according to all testing most figures show no other level. However Naïve Bayes predict that light and weather conditions can only cause level 3 accident severity, which may be true according to the dataset but that isn’t always truth in real life.

## 6.4 - Results and discussion

Table 5 - Do the number of vehicles and casualties in a traffic accident affect accident severity rating?

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 | Average |
| Runtime | 1.58 | 1.48 | 1.48 | 1.78 | 1.89 | 1.30 | 2.56 | 1.87 | Top of Form  1.44Bottom of Form | 1.67 | 1.70 |
| Accuracy | 0.82 | 0.81 | 0.82 | 0.81 | 0.81 | 0.82 | 0.81 | 0.81 | 0.81 | 0.81 | 0.81 |

The algorithm gave for this question the highest runtime with an average of 1.70seconds. Run time is between 1.30 and 2.56. The difference between the lowest and highest run time is 1.26 seconds, which is the highest difference among the other questions however I tent to believe that it is not an import difference. The data accuracy keep an average high score of 0.81. From the following visualisation we can see that all predictions are in limits 1,2 and 3 as expected.

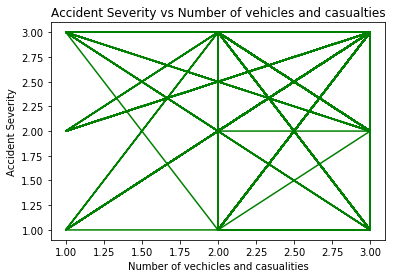


Table 6 - Do the light and weather conditions affect the severity rating in a traffic accident?

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 | Average |
| Runtime | 1.13 | 1.25 | 1.19 | 1.06 | 1.06 | 1.39 | 1.08 | 1.13 | 1.18 | 1.24 | 1.17 |
| Accuracy | 0.83 | 0.83 | 0.83 | 0.83 | 0.83 | 0.83 | Top of Form  0.83Bottom of Form | 0.83 | 0.83 | 0.83 | 0.83 |

The algorithm gave for this question the lowest runtime with an average of 1.17 seconds. Run time is between 1.06 and 1.39. The difference between the lowest and highest run time is only 0.33 second, which I believe it’s normal. The data accuracy keeps an average high score of 0.83. This question has the highest accuracy score of all questions. From the following visualisation we can see that all predictions are in limits 3. This may be correct according to our dataset. For example, if there are no entries that are categorize as level 1 or 2 for accident severity in which the car accident was affected by the light and weather condition then the following figure is correct. Since the dataset is too large to do test this manually I run several times the algorithm however the result is always the same. Furthermore, I believe that there are no accidents with level 1 or 2 accident severity and if they exist there are only a small number that could not show in a figure.

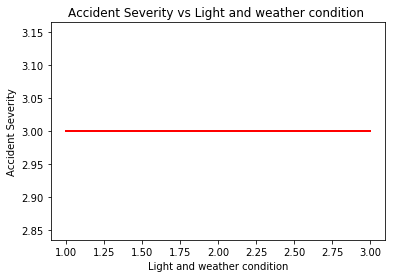
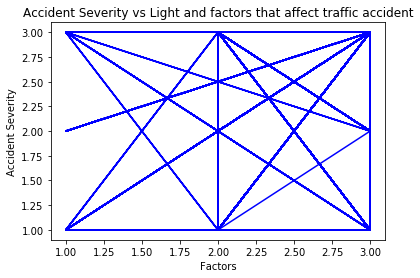


Table 7 - How numerous factors affect the traffic accident severity rating

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 | Average |
| Runtime | 1.42 | 1.52 | 1.49 | 1.42 | 1.52 | 1.32 | 1.26 | 1.34 | 1.31 | 2.02 | 1.46 |
| Accuracy | 0.82 | 0.80 | 0.81 | 0.81 | 0.81 | 0.81 | 0.81 | 0.81 | 0.81 | 0.81 | 0.81 |

The algorithm gave for this question runtime with an average of 1.46 seconds. Run time is between 1.26 and 1.52. The difference between the lowest and highest run time is only 0.26 seconds which I believe it’s normal. The difference between the lowest and the highest run time for this question is the lowest among the other questions. The data accuracy keeps an average high score of 0.81. From the following visualisation we can see that all predictions are in limits 1, 2 and 3 as expected.



## 6.5 - Conclusion and recommendations

In conclusion the accuracy of the result is always high (between 0.81-0.83) which is the reason why most accident severity is in limits. I believe Naïve Bayes algorithm shows with a high accuracy the result of each question but fail to show in question 2, the level 1 and 2 due to small amount of data with the right entries. As mention before we use for testing 20% of the data but a higher percentage may give different outcome.

# 7.0 – Regression analysis by Marios

Regression is a statistical model that used for estimating the relationships among variables. Many techniques are included for modelling and analysing variables. Regression is useful in many sectors like finance, investing or as in our case, analyse a specific dataset in accidents. With regression analysis, we can analyse one or more comparisons about the dataset. Regression can give many statistics like prediction or accuracy of prediction.

## 7.1 - Summary of the approach

Analyse data with regression model has some advantages and disadvantages. One is the ability to determine the relative influence of one or more predictor variables to the criterion value. Another advantage is the ability to identify outliers, or anomalies. The disadvantages of using a regression model usually come down to the data being used. Two examples of this are using incomplete data and falsely concluding that a correlation is a causation (Weedmark,2017) In some datasets that we looked up, regression may had some difficulties. As in our case, from the first point of view, it looks that would be perfect as can analyse one or more variables but at the end, regression had some difficulties on the prediction and accuracy. (Ray, 2015)

## 7.2 - Data pre-processing, visualisation, feature selection

Our dataset has many categories and variables. Algorithm did not included entries that would create any problem to analyse of the data, df.dropna(inplace=True) , is the part that is responsible for this. Also algorithm remove categories of variables that will not select for the analyse, df.drop('Accident\_Index', 1, inplace=True) and similar commands for other categories is for this job. In each of these questions that we chose, we use the necessary commands to remove each category that we will not use. Visualisation of the dataset is well as it show the comparison between accident severity in comparison with other categories like number of vehicles and casualties or light and weather conditions. Algorithm gives three visualisation graphs for each question, one for the current data, one for the predictions and one that is common of current data and predictions, plt.plot(x, y, color = 'red') is the command that give variables for the show of the graph, plt.show() is responsible for the show of the graph.

## 7.3 - Model training, evaluation and testing

Evaluation and comparison of (x, y) of data, x is the Accident Severity (1,2,3) that shows how important was the accident and it compare with y (in each question is different). Linear Regression is the evaluate the dataset, lin\_regression = LinearRegression() ,lin\_regression.fit(x,y) , is evaluate the data. Algorithm give predictions on base on dataset that we use, predictions = lin\_regression.predict(x) is the command that predict reasons of accidents and how important are. As we can see from the accuracy of the predictions, is in low levels. Linear regression predicts some data that are out of the limits. For example, It predict an accident severity that is more than 3 or minus one. Practically, this can be effectively as an accident may is extremely serious or unimportant. In our case, this cannot be possible as the only values that we have is 1, 2, and 3.

## 7.4 - Results and discussion

Table 8 - Do the number of vehicles and casualties in a traffic accident affect accident severity rating?

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 | Average |
| Runtime | 17.07 | 16.68 | 16.32 | 16.56 | 17.89 | 16.90 | 16.89 | 16.30 | 17.54 | 16.25 | 16.84 |
| Accuracy | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |

From the results that algorithm gave, runtime is the lowest from the three questions and comparisons. Runtime is between 16.2 and 17.8 with average time 16.84. According to the data that algorithm had to evaluate, I tend to believe that the different of 1.6 seconds from the lowest to the quickest is a not a huge different. From the visualisation we can see that some of the predictions (green) are out of limits (1-3) and that the most of these are over 2, in comparison with current data (red) that are from 1 to 3.

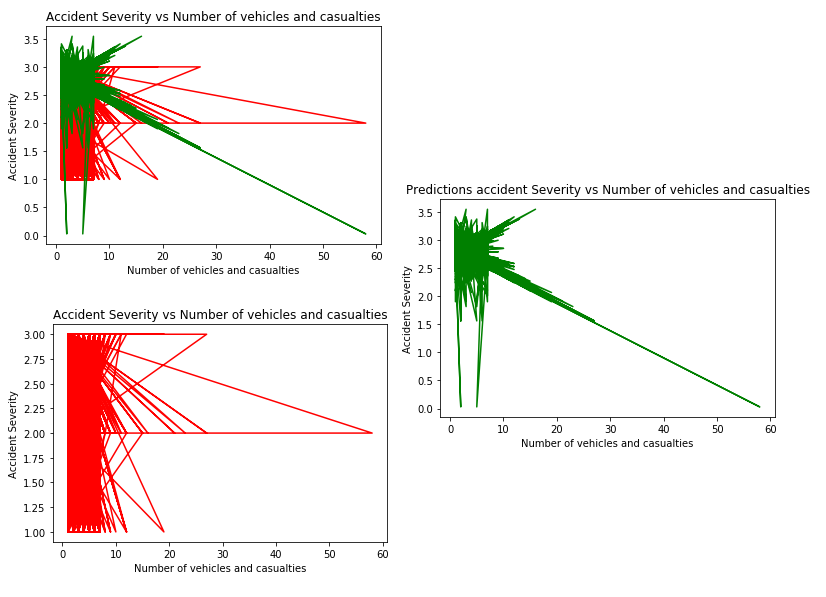


Table 9 - Do the light and weather conditions affect the severity rating in a traffic accident?

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 | Average |
| Runtime | 45.64 | 44.96 | 44 | 49.63 | 43.47 | 43.85 | 48.97 | 49.25 | 50.74 | 43.53 | 46.40 |
| Accuracy | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

From the results that algorithm gave, runtime is the lowest from the three questions and comparisons. Runtime varies between 43.47 and 50.74 with an average time of 46.40. According to the data that algorithm had to evaluate, I tend to believe that the different of 7.27 seconds from the lowest to the quickest is also a big different. From the visualisation we can see that some of the predictions (green) are in the limits (1-3) and that all of the predictions are between a small range 2.74-2.88 in comparison with current data (red) that are from 1 to 3.

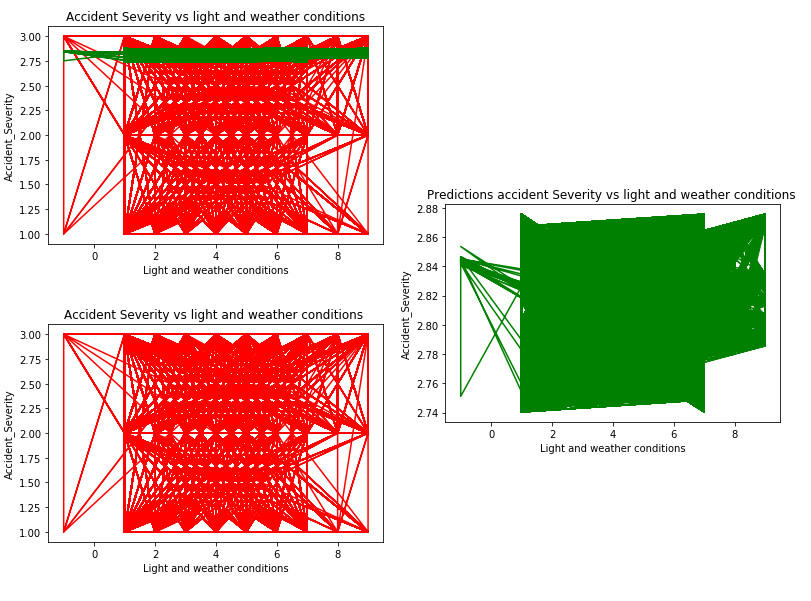
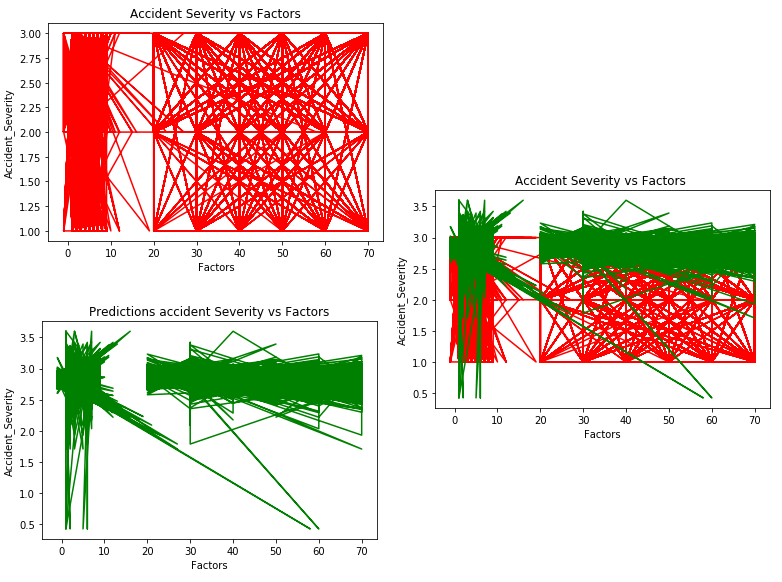


Table 10 - How numerous factors affect the traffic accident severity rating?

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 | Average |
| Runtime | 50.17 | 50.19 | 48.50 | 50.17 | 48.36 | 49.22 | 48.74 | 48.63 | 50.68 | 54.09 | 49.87 |
| Accuracy | 0.03 | 0.02 | 0.03 | 0.03 | 0.03 | 0.03 | 0.02 | 0.03 | 0.02 | 0.02 | 0.026 |

According to the results that algorithm gave, runtime is the lowest from the three questions and comparisons. Runtime is between 48.3 and 54 with average time 49.87. Depend on the data that algorithm had to evaluate, I believe that the different of 5.7 seconds from the lowest to the quickest is also a normal different. From the visualisation we can see that some of the predictions (green) are out of the limits (1-3) and that all of the predictions are between a big range of 0.5-3.6 in comparison with current data (red) that are from 1 to 3.



## 7.5 - Conclusion and recommendations

According to the results, accuracy of the prediction of the algorithm is low and reasons are number of vehicles and casualties that are out of limits. From a general point of view, predictions are highest in severity (1.8-3.5) most of them, in compare with the entries and data that we have.

# 8.0 - Results, comparison and discussion

Three tables were created to compare the results of each algorithm for each question. The best result for each question was showcased as bold to help differentiate it from the other results.

Table 11 - Do the number of vehicles and casualties in a traffic accident affect accident severity rating?

|  |  |  |  |
| --- | --- | --- | --- |
|  | KNN | Naïve-Bayes | Regression |
| Runtime | 46.1 | **1.70** | 16.84 |
| Accuracy | **83.81** | 81 | 1 |

Table 12 - Do the light and weather conditions affect the severity rating in a traffic accident?

|  |  |  |  |
| --- | --- | --- | --- |
|  | KNN | Naïve-Bayes | Regression |
| Runtime | 46.25 | **1.17** | 46.40 |
| Accuracy | 76.38 | **83** | 0 |

Table 13 - How numerous factors affect the traffic accident severity rating

|  |  |  |  |
| --- | --- | --- | --- |
|  | KNN | Naïve-Bayes | Regression |
| Runtime | 47.97 | **1.46** | 49.87 |
| Accuracy | 80.47 | **81** | 2.6 |

K-Nearest Neighbours and Naïve-Bayes remained close in accuracy for all three questions, whereas the regression algorithm had an extremely low accuracy rating in comparison.

For the first question (table 11), K-Nearest Neighbours had the highest accuracy of 83.81 by 2.81%.

For the second (table 12) and third question (table 13) Naïve-Bayes had the highest accuracy, leading second place in the second question by a significant 6.62% and leading the third question by measly 0.53%.

Naïve-Bayes had the fastest runtime out of all three algorithms by a large margin, taking no less than 2 seconds to run for every question, in comparison to K-Nearest Neighbors maximum of 47.97 seconds and Regressions maximum of 49.87.

# 9.0 - Conclusion and recommendations

To conclude, Naïve-Bayes seems the best algorithm for this dataset, being the most accurate two out of three times and having extremely quick runtimes in comparison to the other algorithms. However, it should be noted that only three algorithms were used for this dataset and there are many other machine learning algorithms which could be tested to give a more complete picture of which is best for the dataset. Another reason for Naïve-Bayes being the superior option is that the algorithm that was its closest performer, K-Nearest-Neighbors, doesn’t scale as well with data, having a lengthy runtime on larger datasets.

The results found that all the factors described in the three questions did have an impact on the accident severity rating, with the strongest link being between the number of accidents and casualties and the accident severity rating.

More research could be carried out with a class other than accident severity rating as the focus to determine other relationships between the features in this dataset.

# 10.0 - References

Ahmad Ashari, Iman Paryudi, A Min Tjoa (2013), Performance Comparison between Naïve Bayes,

Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No. 11, 2013, pages:33-39

Dataaspirant. (2018). Gaussian Naive Bayes Classifier implementation in Python. [online] Available at: http://dataaspirant.com/2017/02/20/gaussian-naive-bayes-classifier-implementation-python/ [Accessed 22 Jan. 2018].

Department for Transport (2017). *Road Safety Data.* Available at: <https://data.gov.uk/dataset/road-accidents-safety-data/resource/91789e37-03e5-48cf-9720-2d13639c32b9> [Accessed 11 Jan. 2018]

Keller, J.M., Gray, M.R. and Givens, J.A. (1985). ‘A fuzzy K-nearest neighbor algorithm’. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol.15, No.4, pp.580-585

Kotsiantis, S.B. (2007). ‘Supervised Machine Learning: A Review of Classification Techniques’. *Emerging Artificial Intelligence Applications in Computer Engineering*, pp.3-24

R), 6., R), 6. and Ray, S. (2018). 6 Easy Steps to Learn Naive Bayes Algorithm (with code in Python). [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/ [Accessed 22 Jan. 2018].

Ray, S. 7 (2015) *Types of Regression Techniques you should know!* , Available at: <https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/> [Accessed 21 Jan. 2018]

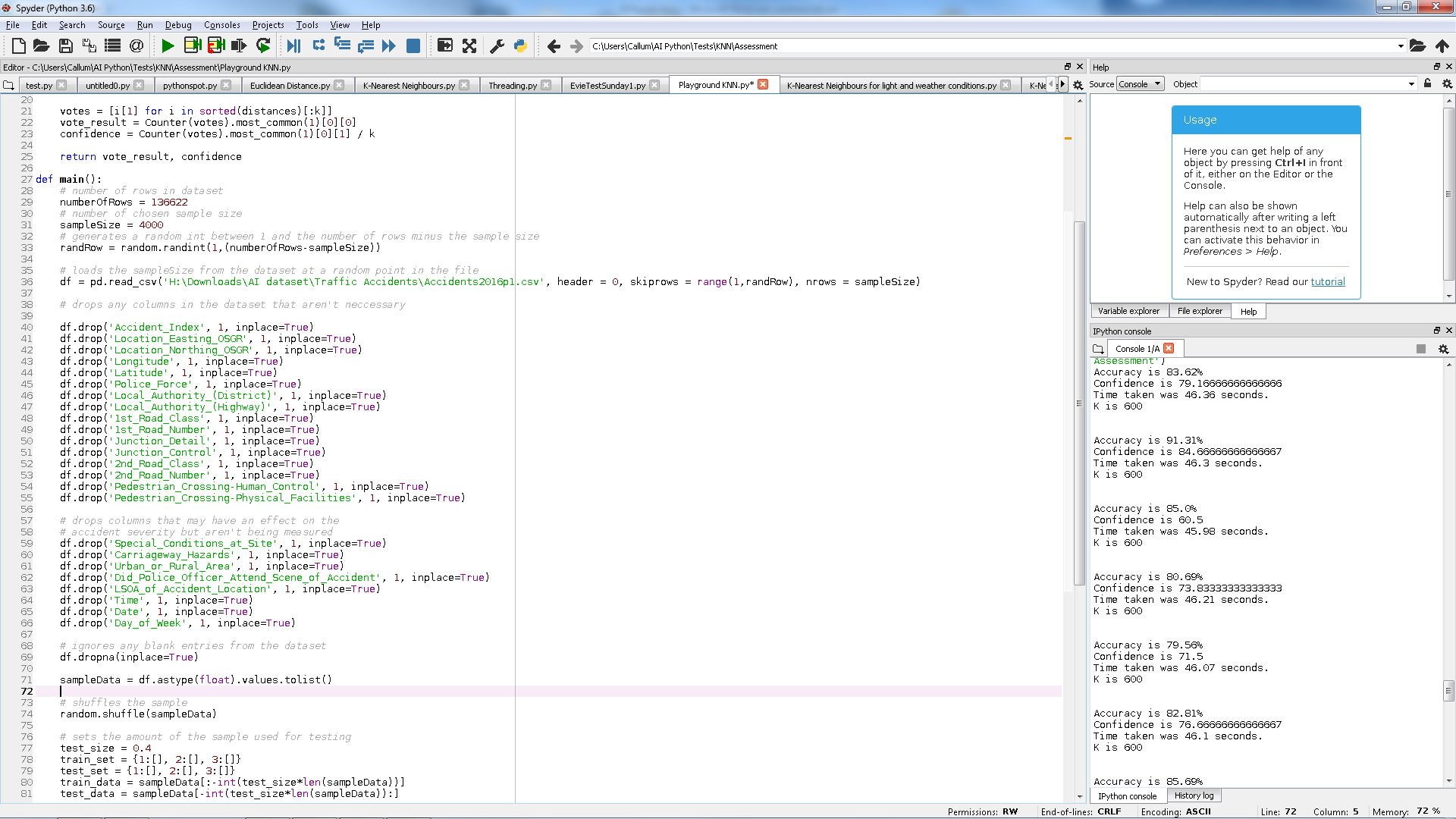
VanderPlas, J. (2018). In Depth: Naive Bayes Classification | Python Data Science Handbook. [online] Jakevdp.github.io. Available at: https://jakevdp.github.io/PythonDataScienceHandbook/05.05-naive-bayes.html#Gaussian-Naive-Bayes [Accessed 22 Jan. 2018].

Weedmark, D. (2017) *The Advantages & Disadvantages of a Multiple Regression Model*, Available at: <https://sciencing.com/advantages-disadvantages-multiple-regression-model-12070171.html> [Accessed 21 Jan. 2018]

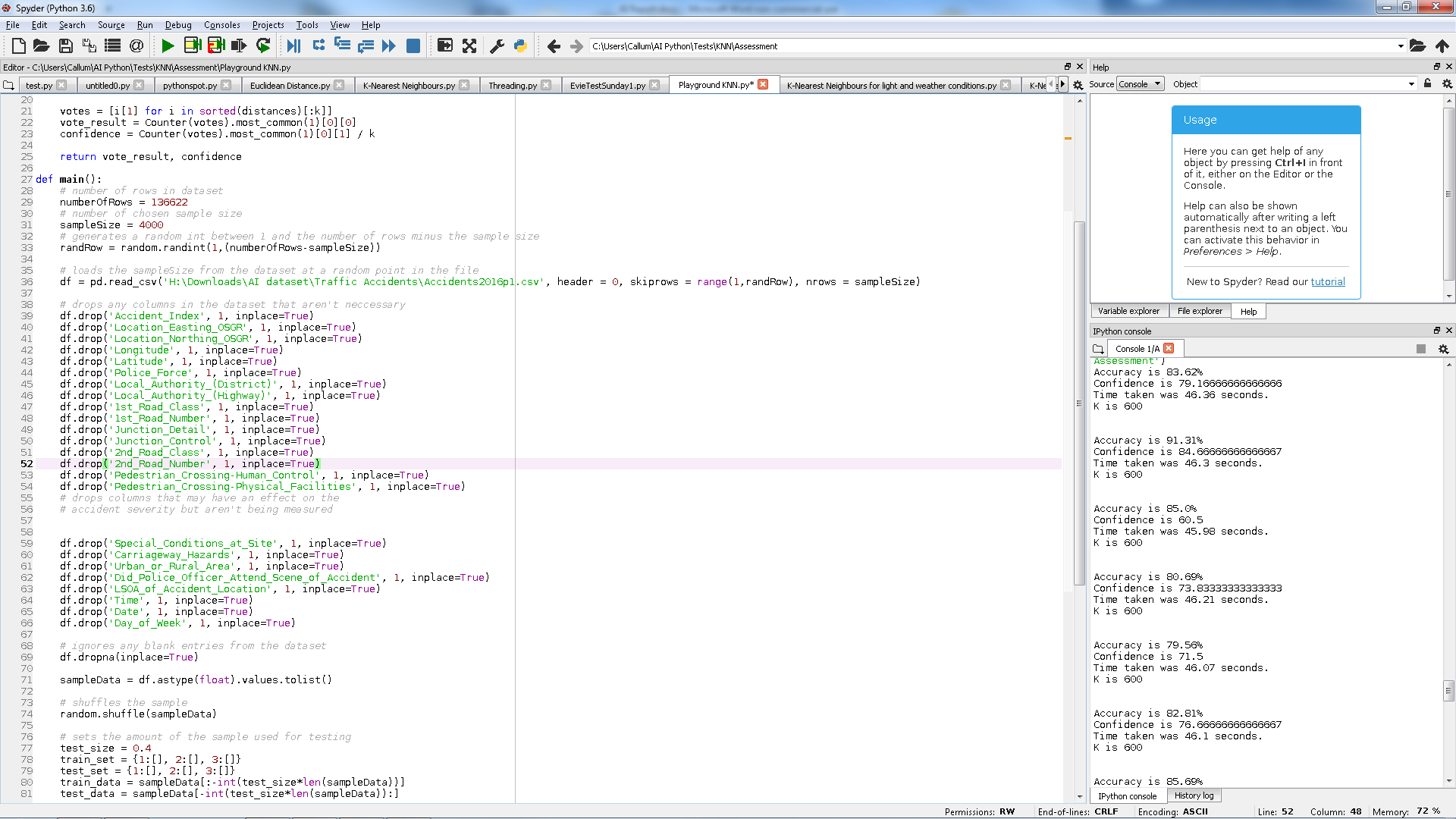
Weinberger, K.Q. and Saul, L.K. (2009). ‘Distance Metric Learning for Large Margin Nearest Neighbor Classification’. *The Journal of Machine Learning Research*, Vol.10, pp.207-244

# 11.0 – Appendices

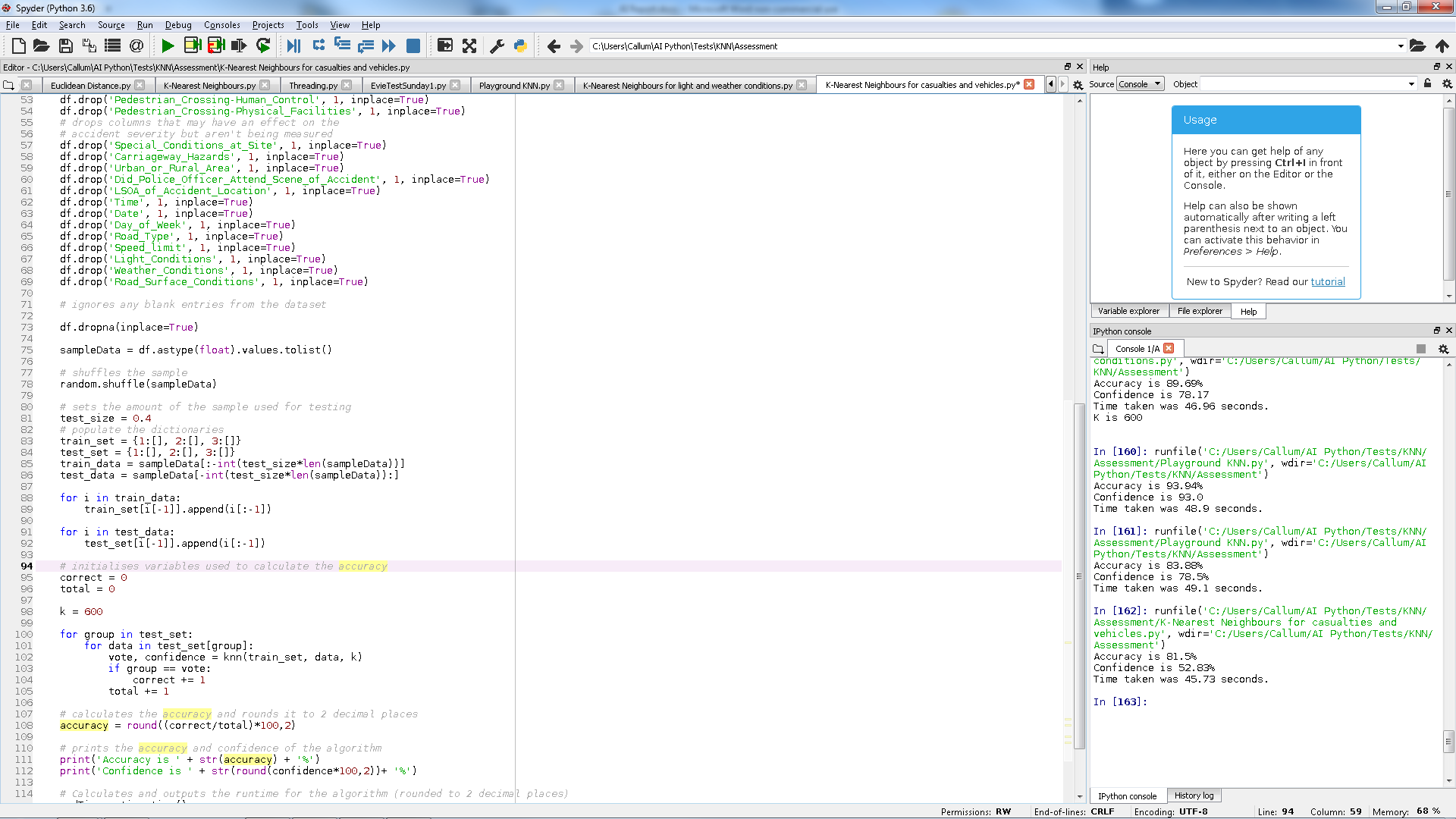
## Appendix A



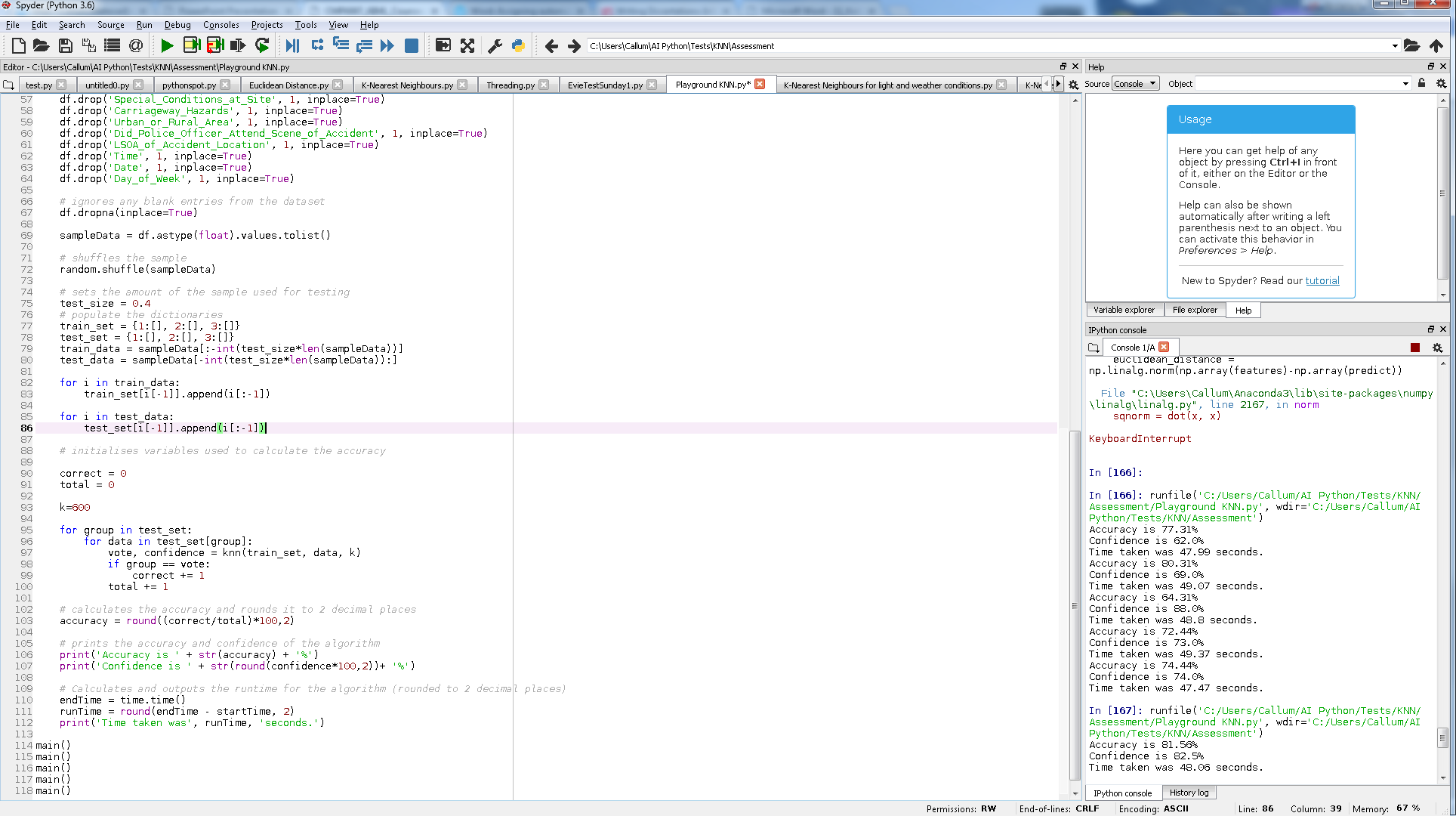
## Appendix B



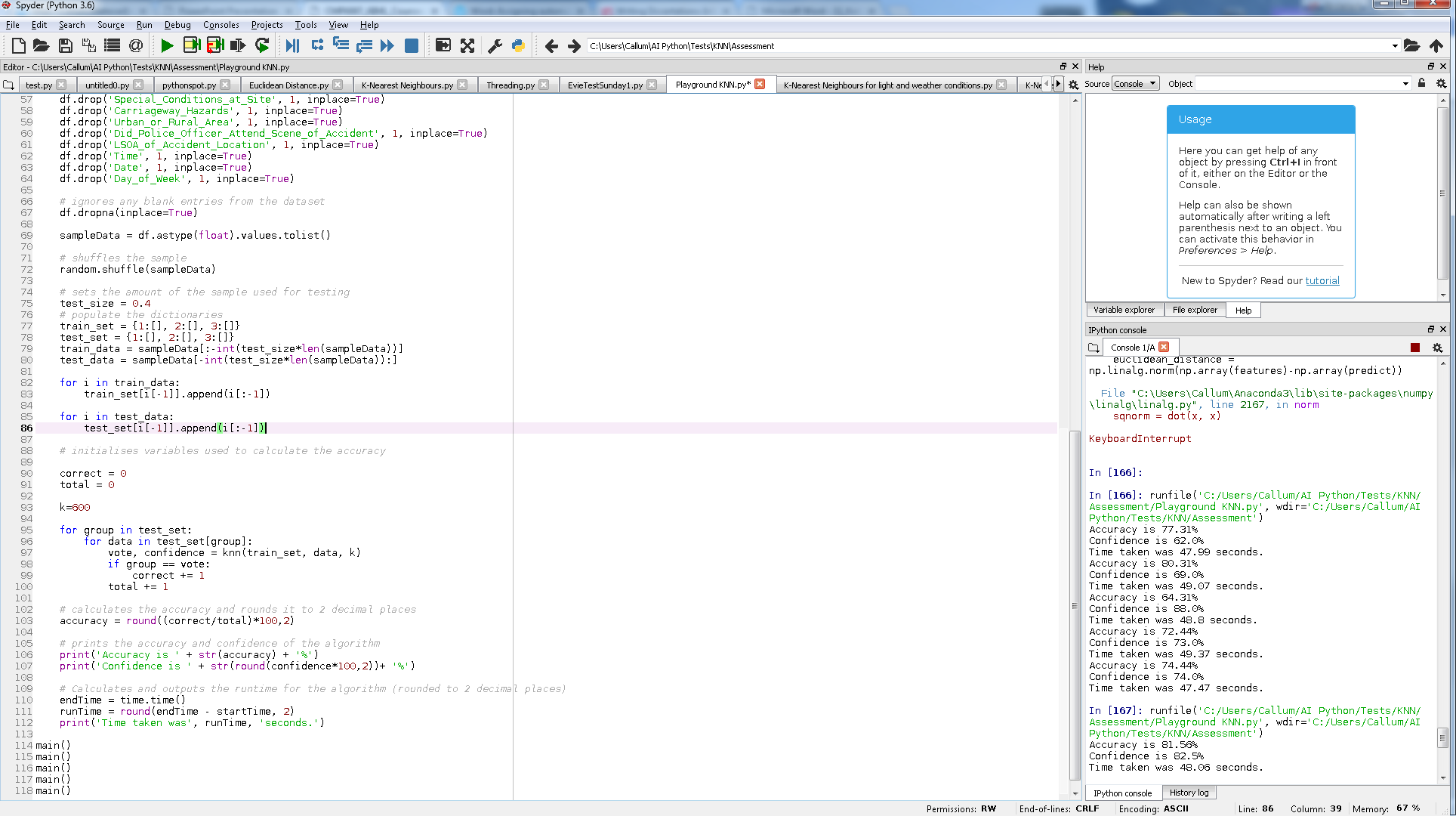
## Appendix C



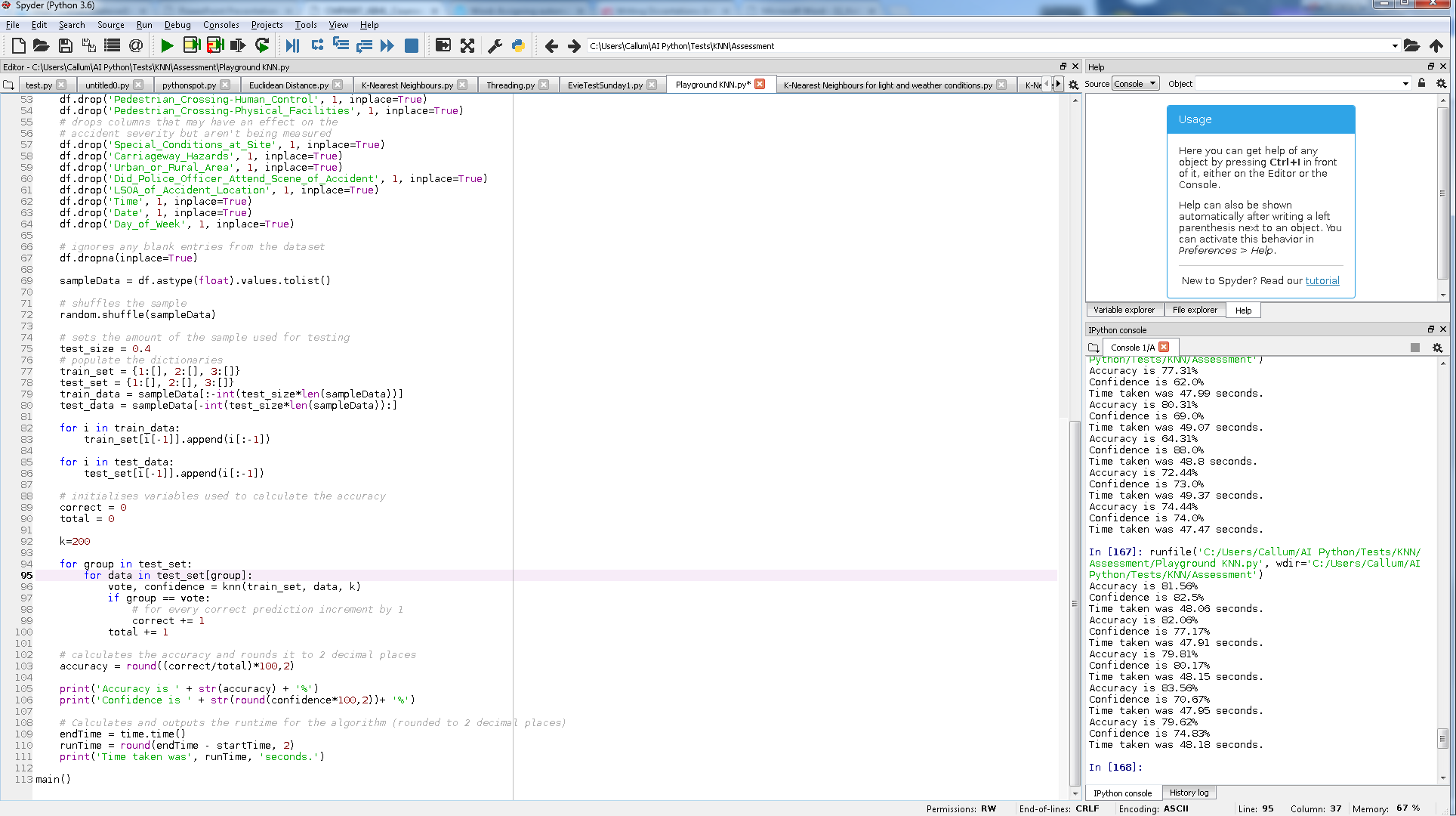
## Appendix D



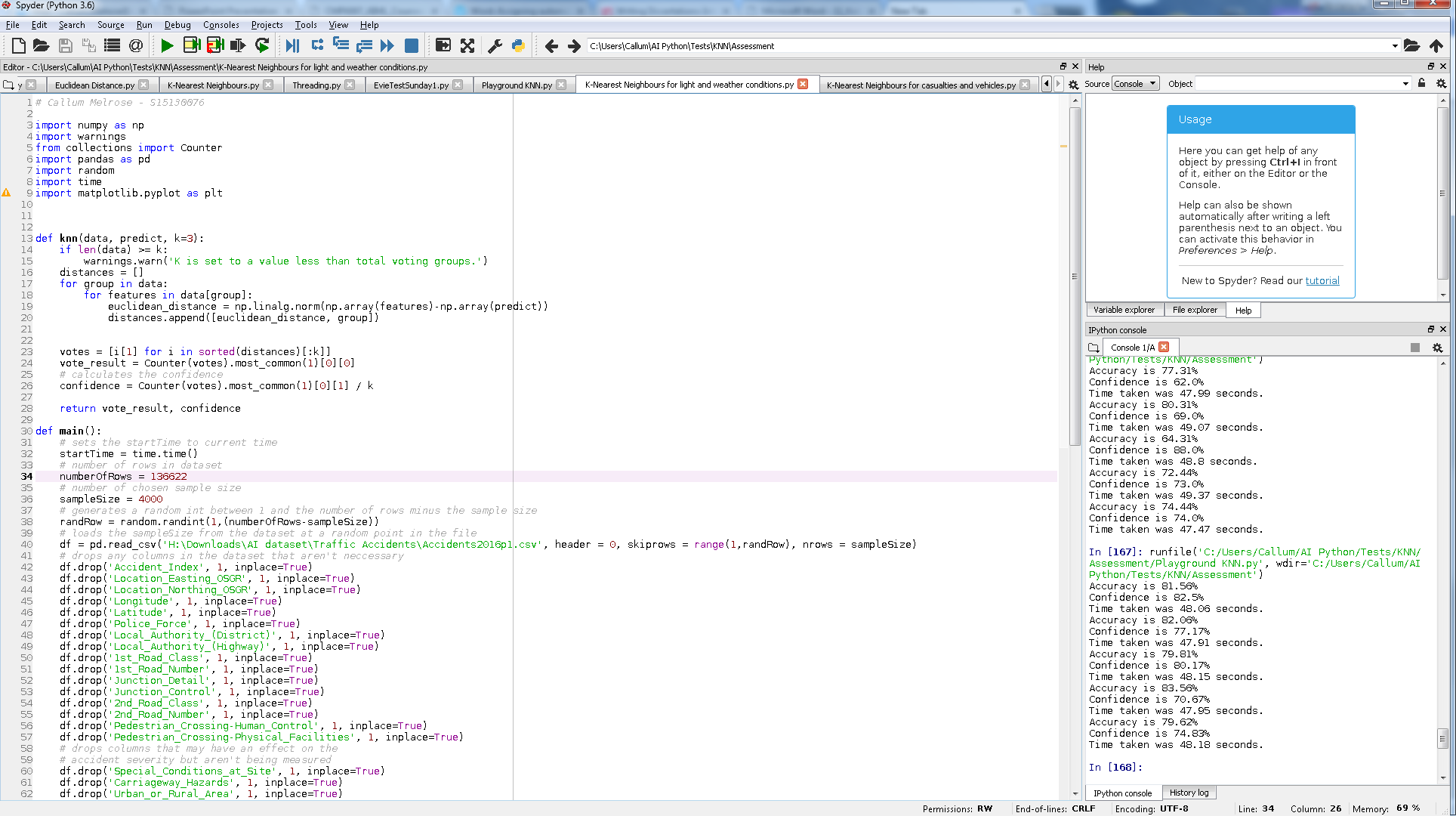
## Appendix E



## Appendix F



## Appendix G



## Appendix H

