CT475
Machine Learning & Data Mining

Assignment 1

Name: Taidgh Murray

Student I.D: 15315901

Discipline: College of Science

Course: 4BS2 Undenominated Science (Computing)

The first required by this assignment was to select an appropriate machine learning package. The main requirement for choosing the right package was that it be open source. Trying to find a suitable open source package was a challenge. I wanted to find a package in Python. I feel that Python is the most appropriate language for machine learning, it has an excellent choice of libraries for any number of tasks, and is quite lightweight as a language. Furthermore, it's my preferred language, I'm most comfortable using it.

The package I decided to go with was SciPy. There is a large variety of Machine Learning packages in Python, so choosing the right one to go with was a challenge. I initially looked at TensorFlow[1], it's used by Google, and is touted as "An open source machine learning framework for everyone". For the purpose of this assignment, however, I felt it was too advanced & specific. SciPy was my next choice. SciPy is more robust & 'all purpose'. I have some familiarity with NumPy, which SciPy has some connections[2] with. All knowledge on the workings of the package was taken from the documentation[3], and this[4] website outlining instructions for using said framework

With the choice of Machine Learning package sorted out, I began to prepare the data that had been given to us, the 'autoimmune.txt' file on Blackboard. SciPy works with CSV files, so converting the .txt file to a .csv file was my first step. During this step, I transposed the rows and columns in excel, it made logical sense to me to have the data displayed in this way. I saved the file as a .csv, which Python can recognise as a list.

[1] https://www.tensorflow.org/
[2] https://www.scipy.org/about.html#the-scipy-ecosystem
[3] https://www.scipy.org/docs.html
[4] https://machinelearningmastery.com/machine-learning-in-python-step-by-step/

Having sorted out the formatting of the data, I could then focus on the Machine Learning section of the assignment. The first task was to choose two Machine Learning Algorithm. One algorithm was ID3/C4.5 or k-Nearest-Neighbours, both of which we had covered in the lectures. The other is one of our own choosing.

The first algorithm I chose was the k Nearest Neighbours algorithm. The simplest definition I could find was directly from the lecture slides[5]. Each data point can be considered a point in a sample space, usually a two-dimensional graph. If two samples are close to each other in space, then they should be close to each other in target values as well. If you then wanted to add a new query to the data, you compare it to the nearby values in the data set. For k Nearest Neighbours, you base your prediction on the 'k' nearest neighbours, k usually being any number between 3 and 20.

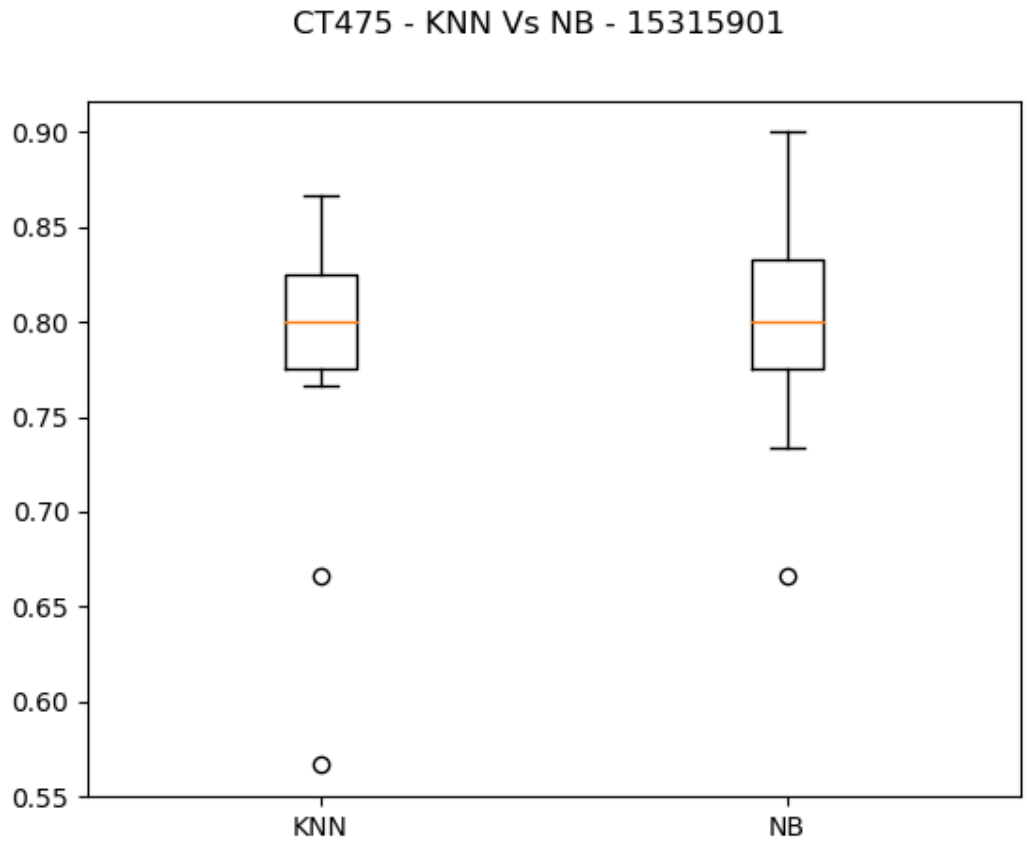
The second algorithm I chose was the Naïve Bayes algorithm. Taken from this[6] website, the Naïve Bayes theorem is based off Bayes' Theorem from Statistics, which describes the probability of an event, 'P' based off the prior knowledge of conditions, 'A' & 'B' that might relate to said event. The Naïve Bayes algorithm assumes that each conditions, 'A','B','C', etc. independently contribute to the probability, 'P', which is why it's 'Naïve'.

I had been asked to perform 10-fold cross-validation to estimate likely future performance on both of the classification models. I split up the data into 10 parts, 9 'train' parts, and one 'test' part. From the lecture notes, All the data is used 9 times for training, and all data is used once for testing. The following lines are the accuracy scores of the two algorithms (K Nearest Neighbours & Naïve Bayes).

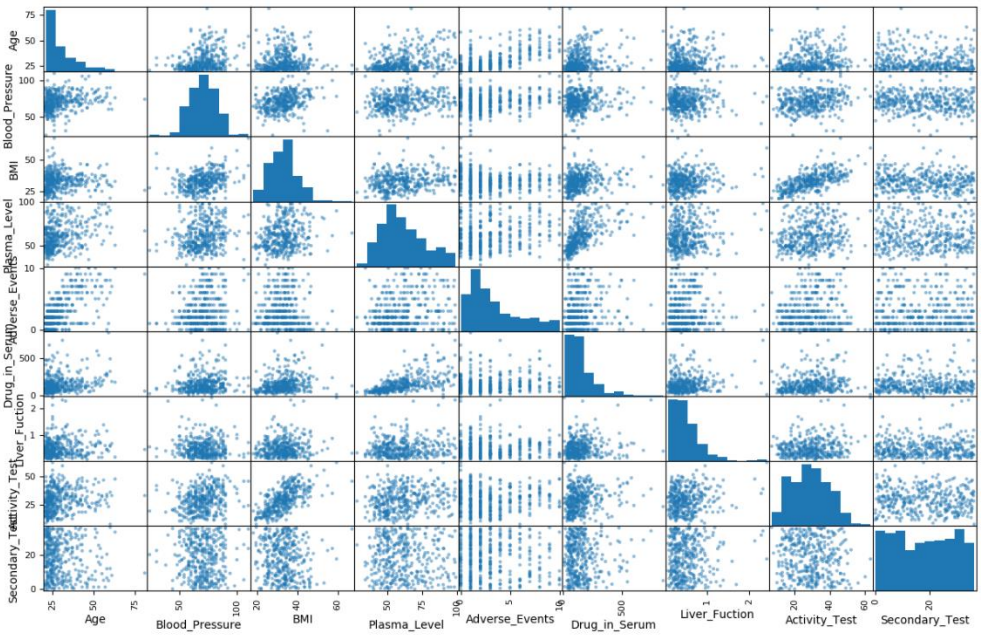
[5] https://nuigalway.blackboard.com/bbcswebdav/pid-1559028-dt-content-rid-11877344_1/courses/1819-CT475/CT475_03_IBL.pdf
[6] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

KNN: 0.7733333333333333 (0.08537498983243799)
NB: 0.7966666666666666 (0.06046119049072354)

A boxplot comparing the algorithms can be found here:



A multivariate plot of the data can be found here:



Comparing the two algorithms, we can see that the Naïve Bayes algorithm has the larger estimated accuracy score, but has a larger spread. KNN's estimated accuracy is slightly lower, but has a much tighter spread. Their average is close enough that there should be no issue choosing one algorithm over another. They're both nonlinear algorithms, and unlike algorithms such as Logistic Regression, are designed to work with non-binary data. The NB algorithm assuming every data point in independent of the result is likely what results in the larger spread, and its higher estimated accuracy score.

Below is the accuracy score of the two algorithms, as well as the classification report.

```
k Nearest Neighbours Prediction:

Accuracy Score: 0.7894736842105263

Confusion Matrix:
 [[45  8]
 [ 8 15]]

Classification Report:
              precision    recall  f1-score   support

    negative       0.85      0.85      0.85        53
    positive       0.65      0.65      0.65        23

   micro avg       0.79      0.79      0.79        76
   macro avg       0.75      0.75      0.75        76
weighted avg       0.79      0.79      0.79        76


Naive Bayes Prediction:

Accuracy Score: 0.75

Confusion Matrix:
 [[41 12]
 [ 7 16]]

Classification Report:
              precision    recall  f1-score   support

    negative       0.85      0.77      0.81        53
    positive       0.57      0.70      0.63        23

   micro avg       0.75      0.75      0.75        76
   macro avg       0.71      0.73      0.72        76
weighted avg       0.77      0.75      0.76        76
```

Looking at the classification report, we can see that the Nearest Neighbour actually has the higher accuracy, at 78.9%, with the Naïve Bayes slightly lower, at 75%, unlike the predicted scores above. The confusion matrix & classification report shows that the kNN algorithm correctly predicted more negative results, while the NB algorithm correctly more positive results.