COMP30027: Machine Learning Assignment 1

Question 1

Discretising numeric variables does improve classification performance. A k-means clustering method was used to find the natural breakpoints in the data to discretise it, leading to higher accuracy.

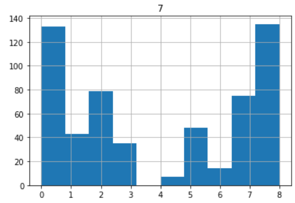
The following datasets were tested, and all hold similar results: 'WDBC.Data', 'Wine.Data', 'Adult.Data', 'Bank.Data' (see python script)

For a numeric dataset WDBC.Data, the accuracy increased by 2% when data was discretised:

```
Accuracy with no discretisation 0.9367311072056239
Max Accuracy with discretisation: 0.9560632688927944
N K-splits to receive Max Accuracy: 7
```

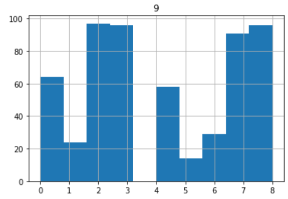
This is consistent with the attributes displaying non-Gaussian behaviour (2 examples shown below):

(i)



a. Most datapoints are located on the 'tails' of the distribution, not normal

(ii)



a. Repetitive pattern within the data, distribution of datapoints not normal

For all attributes within WDBC.Data, none exhibit Normal Gaussian behaviour (see python script), meaning that the probabilities predicted by the Naïve Bayes Gaussian Function will always be inaccurate. To address this, the data is discretised into 7 centroids to achieve a maximum accuracy of 95.6%.

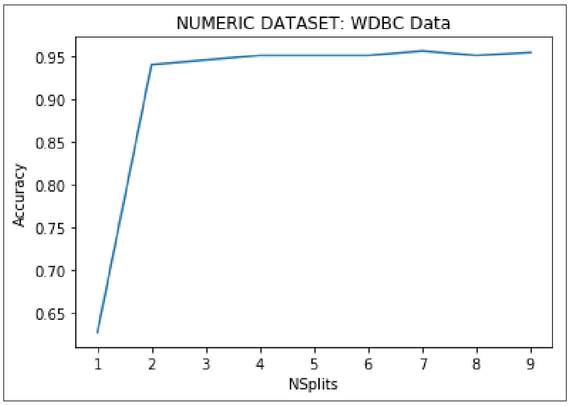


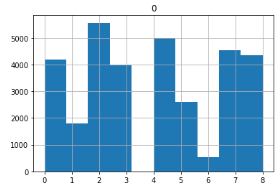
Fig.1: Accuracy of Naïve Bayes model given varying levels of discretisation

Similarly, for mixed dataset Adult.Data, accuracy decreased by over 1% when discretising:

```
Accuracy with no discretisation 0.8333282147354196
Max Accuracy with discretisation: 0.8273087435889561
N K-splits to receive Max Accuracy: 9
```

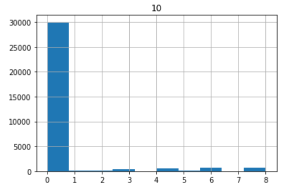
For all attributes within Adult.Data, none exhibit Normal Gaussian behaviour (see python script):

(i)



a. Multiple peaks and troughs within dataset, not normal

(ii)



a. Most of Adult.Data looks like the following distribution: large left peak with long right tail, not normal
b. This means that the standard deviation and mean are highly centralised around the left peak and most times it will correctly predict the distribution as occurring on the left - for this reason, the accuracy with no discretisation performs better

For this dataset, no matter the level of discretisation the accuracy will always be lower than using the Gaussian probabilities:

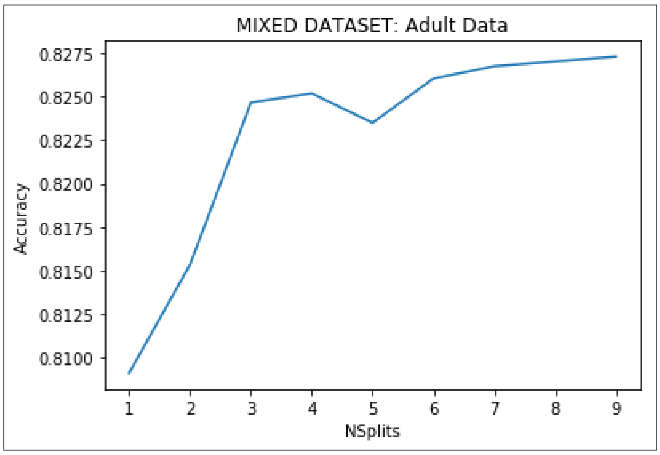


Fig.2: Accuracy of Naïve Bayes model given varying levels of discretisation

Question 2

Since the training and testing sets are equivalent, the baseline performance varies due to the predominance of the most common class within a respective dataset.

The Zero R classifier tends to perform best in cases with fewer classes and a strong bias towards one class in both the training and testing set – this is seen in Bank Data. The Zero R baseline is outperformed by the Naïve Bayes model in all other datasets (below):

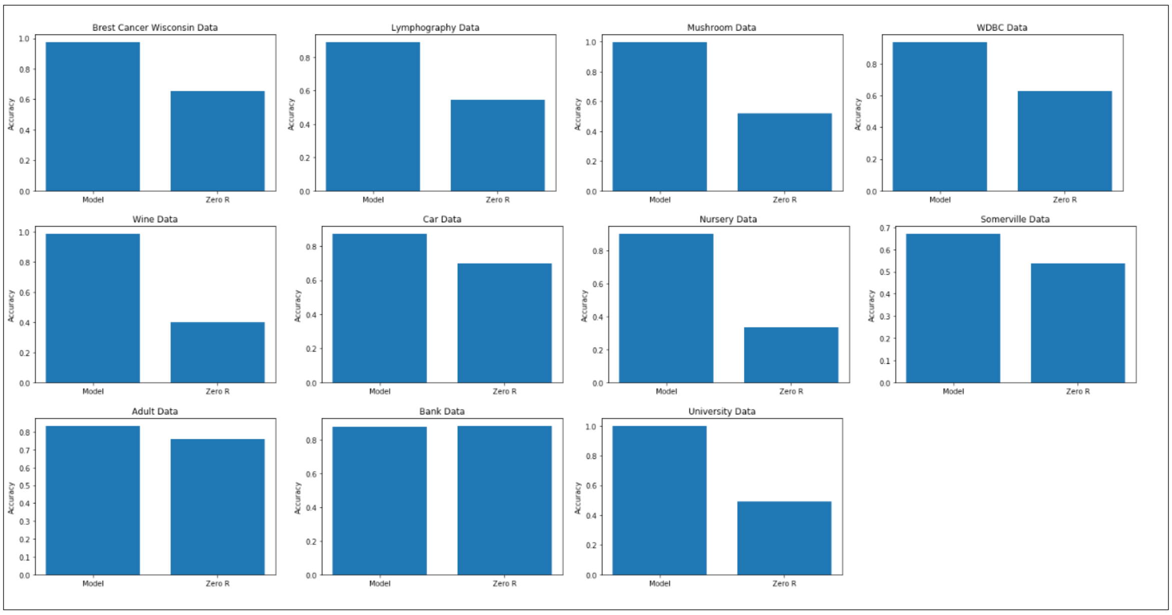
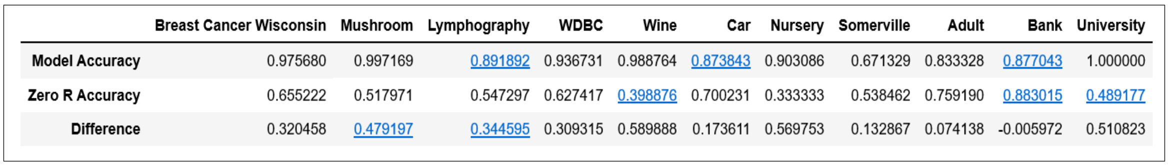


Fig.3: Accuracy of Naïve Bayes model against accuracy of Zero R baseline

The Zero R can only ever predict one class meaning that there is an upper limit to the accuracy it will face. The Naïve Bayes can predict all classes, meaning that it does not face the same upper limit, however, is limited by the learned probabilities from the training dataset. Overall, across all datasets except for Bank.Data, the accuracy is higher in the Naïve Bayes model.

Below is the extent to which the Naive Bayes classifier improves on the baseline Zero R performance:



	Breast Cancer Wisconsin	Mushroom	Lymphography	WDBC	Wine	Car	Nursery	Somerville	Adult	Bank	University
Model Accuracy	0.975680	0.997169	0.891892	0.936731	0.988764	0.873843	0.903086	0.671329	0.833328	0.877043	1.000000
Zero R Accuracy	0.655222	0.517971	0.547297	0.627417	0.398876	0.700231	0.333333	0.538462	0.759190	0.883015	0.489177
Difference	0.320458	0.479197	0.344595	0.309315	0.589888	0.173611	0.569753	0.132867	0.074138	-0.005972	0.510823

Fig.4: Accuracy of Naïve Bayes model and Zero R baseline