Assignment 1 Report
By: Andrew Tjen
Student Id: 939206

Q1.)

The method chosen in the data set is equal width discretization. Given that the data is 5 times randomly sub-sampled (75% train, 25% test) with add-k smoothing of k = 1.

Dataname	level 0	level 3	level 5	
wdbc.data	93.17%	94.28%	91.38%	91.35%
TIME TAKEN	0.19s	0.76s	0.93s	0.97s
wine.data	97.98%	94.34%	88.43%	82.54%
TIME TAKEN	0.09s	0.37s	0.44s	0.49s
adult.data	83.16%	81.86%	81.71%	81.57%
TIME TAKEN	1.1s	1.38s	1.42s	1.49s
bank.data	87.71%	88.89%	88.71%	88.94%
TIME TAKEN	0.68s	1.07s	1.16s	1.24s
university.data	11.27%	10.36%	12.88%	15.19%
TIME TAKEN	2.84s	4.98s	5.31s	5.35s

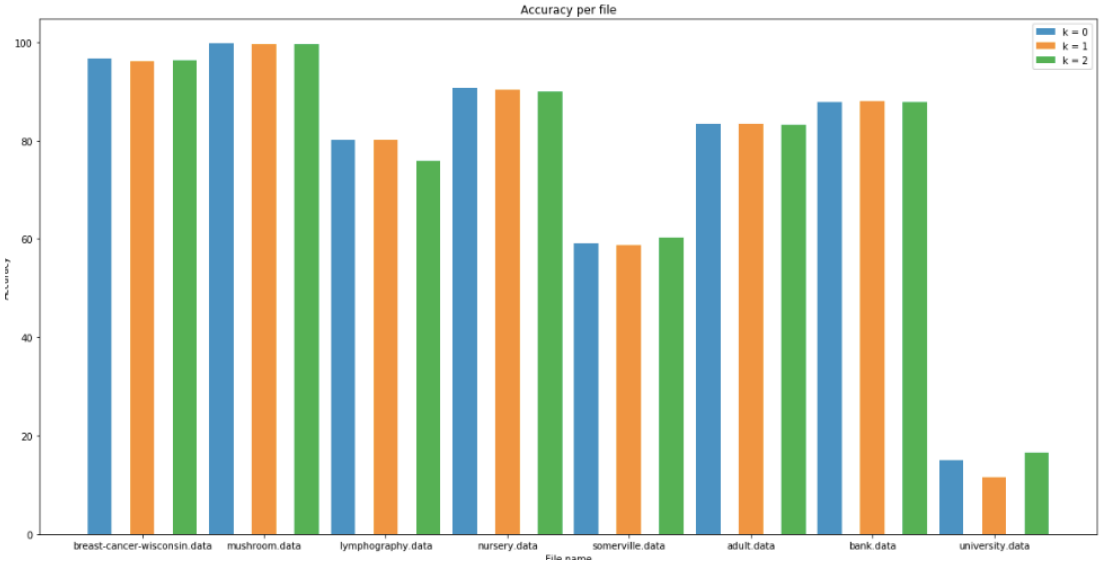
The right discretizing technique can definitely out perform Gaussian Naïve Bayes approach. However, majority of the file has a higher precision when discretizing level equals to zero. Meaning that it has a higher precision when numeric attributes are treated as Gaussian Normally distributed. A reason for this because of Gaussian Normal distribution is a good assumption that fits quite well to the train set. Another reason can be caused by the sensitivity to outlier values. Although Gaussian normal distribution is sensitive to outliers as well, equal width discretization is more sensitive. When an outlier exists, the size of the partitions can significantly differ, which will cause precision loss. And due to the random sampling nature of train and test set, the presence outliers are very possible when model is trained.

Q5.)

The data is produced by add-k smoothing ranging from 0 to 2 with random subsampling of 5 and no discretization.

Dataname	k = 0	k = 1	k = 2
breast-cancer-wisconsin.data	96.62	96.03	96.27
TIME TAKEN	0.22s	0.23s	0.23s
mushroom.data	99.77	99.54	99.51
TIME TAKEN	0.55s	0.63s	0.54s
lymphography.data	80.1	80.06	75.82
TIME TAKEN	0.32s	0.29s	0.28s
nursery.data	90.75	90.36	90.0
TIME TAKEN	0.43s	0.41s	0.36s
somerville.data	59.06	58.74	60.26
TIME TAKEN	0.08s	0.09s	0.09s
adult.data	83.43	83.39	83.16
TIME TAKEN	1.05s	1.06s	1.06s
bank.data	87.78	88.07	87.85
TIME TAKEN	0.66s	0.67s	0.65s
university.data	14.9	11.6	16.54
TIME TAKEN	2.66s	2.73s	2.85s

In theory, changing smoothing regime from originally no smoothing regime will increase the effectiveness of the Naïve Bayes classifier. This is due to the prevention of the likelihood probability to not be zero. However, we can't that conclude the data follows the theory as half of the dataset has a higher precision and another half has a lower precision. This could be caused, again, by random sub sampling of 5 samples. As the sample size increase, it will have higher chance of unique attribute values to be trained to the model. Thus, average precision might be higher even though there add-k level is zero.



We can also see that there not really much change from add-k level one to two. However, in theory, we should have lower precision. As add-k smoothing increases the probability of a data appears less than average more than their supposed MLE. Thus, could result in classification mistake. Therefore the higher

the add-k level, the less precision it may get. The reason why data with add-k level one to two does not show a difference is the same as the previous reason where it is caused by random sub sampling.