**Final Project: Breast Cancer Data Set**

CAP 4601 - Fall 2020

*Wendy Slattery*

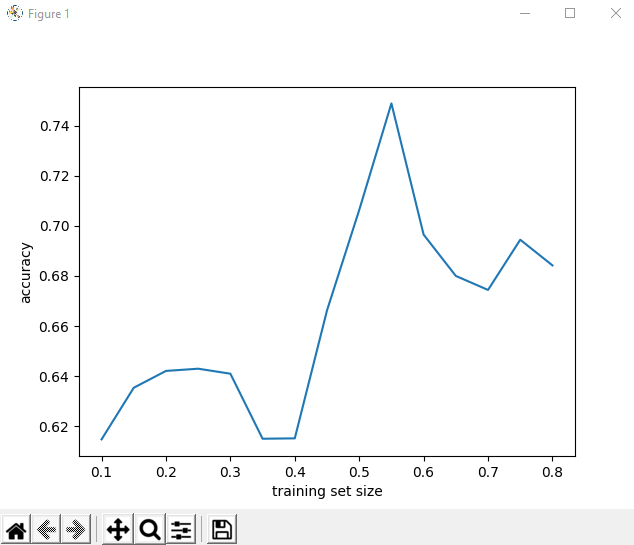
**Data Set:** Breast Cancer data source from Institute of Oncology, Ljublijana, Yugoslavia supplied and pulled from University of California Irving at <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer>

This set contains 285 data points listed by the class of Recurrence (85) vs No Recurrence events (200) and associated with nine attributes:

1. Age: discrete ranges from 10 - 99 and not all actually show up in the data
2. Menopause: lt40 (late 40s), ge40 (generally 40s), and premeno (pre-menopausal)
3. Tumor-size: discrete ranges from 0 - 59
4. Inv-nodes: number of lymph nodes involved with tumor, discrete ranges 0 – 39
5. Node-caps: Boolean yes/no if these nodes were capped or walled in cancer
6. Deg-malig: the degree to which the cancer is malignant 1, 2, or 3 (worst)
7. Breast: left or right
8. Breast-quad: which quadrant of breast tumor is located
9. Irradiat: was tumor irradiated, yes or no

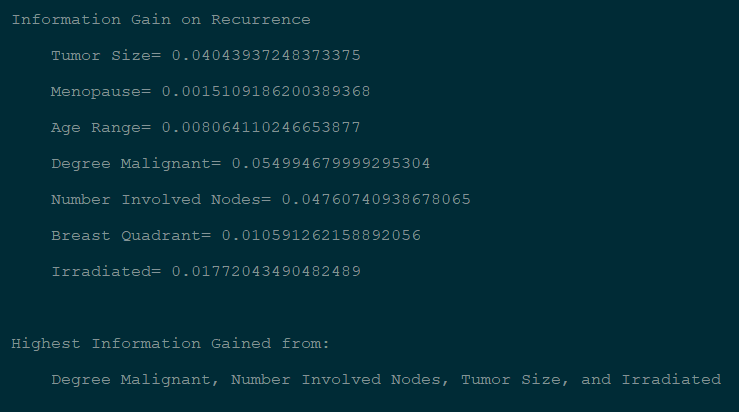
**Model 1: Decision Tree**

I started with several attributes (age range, tumor size, degree malignant, involved num nodes, menopause, breast quadrant, and irradiate) to apply to my classifier which was ‘recurrence’ vs. ‘no recurrence’ and plotted accuracy vs training set size resulting in:

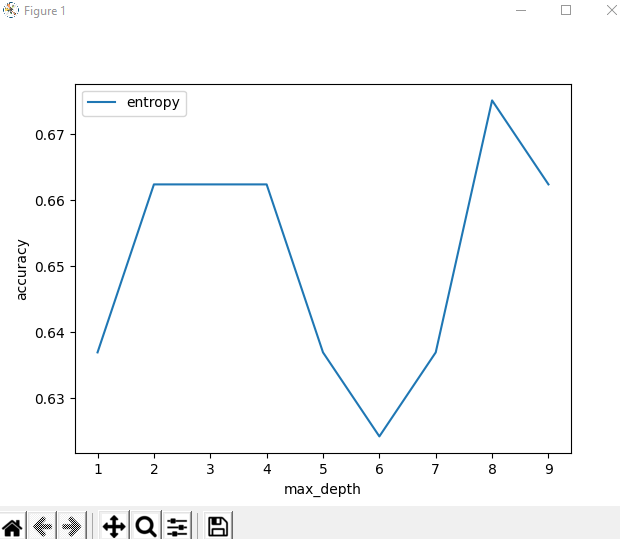


We can see from this graph that our attributes produce an accuracy that maxes out at approximately 75%. That max is found with a training set size close to .55 and there is another high accuracy found at about 64% for training sets roughly .20 -.30 in size. So, from this information, I would deduce that the best training set size would be an average of these in the middle. So, I will then use a training set size of .45 to determine the best tree level depth of accuracy I might achieve with this attribute set.

I then ran an information gain upon these attributes to discover which attributes were contributing the most correlation to the accuracy:

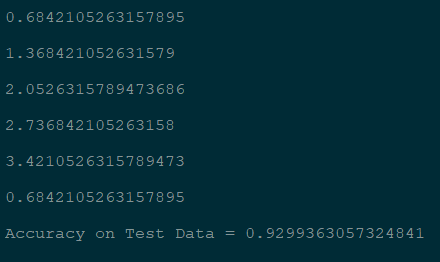


As can be seen by these results and in order of gain, ‘degree malignant’, ‘num involved nodes’, ‘tumor size’, and ‘irradiated’ offered the most value. We will use this information to recreate our decision tree and apply the accuracy vs training set size of .45 to our data to discover a best max tree depth value.

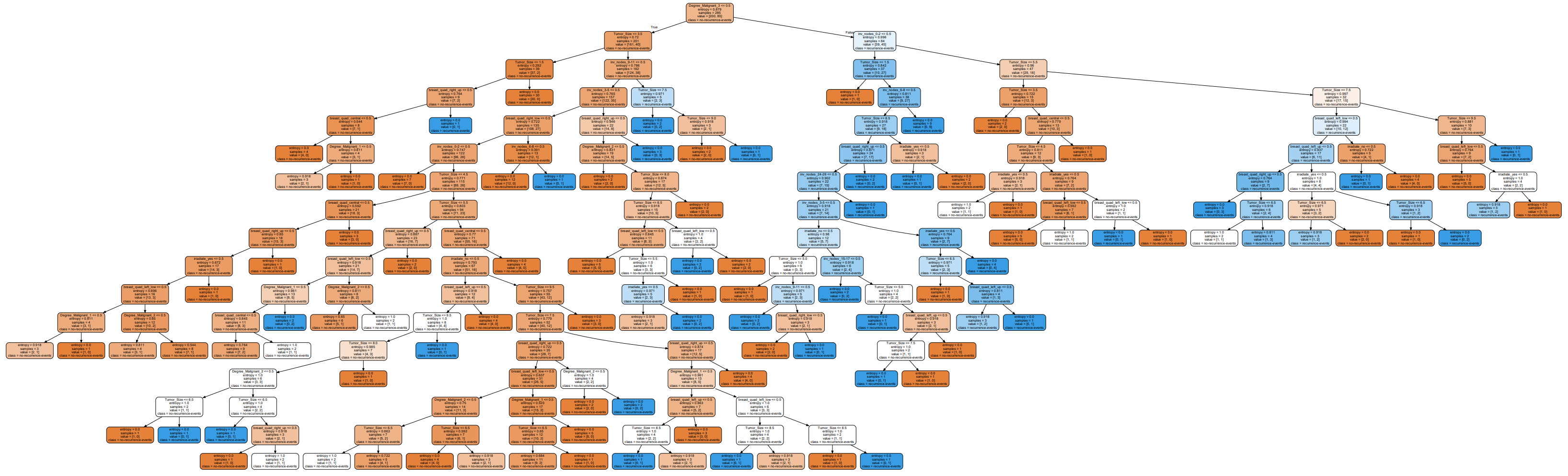


The accuracy vs max depth graph shows us a couple of things. First, we plateau our accuracy at 4 levels. After that, we don’t gain anything by adding more levels to our tree. We do observe a spike at level 8, however, the degree of gain from that is so inconsequential that it doesn’t justify adding all that extra data to the tree just to jump .01 in accuracy, so we will set our tree max depth to 4.

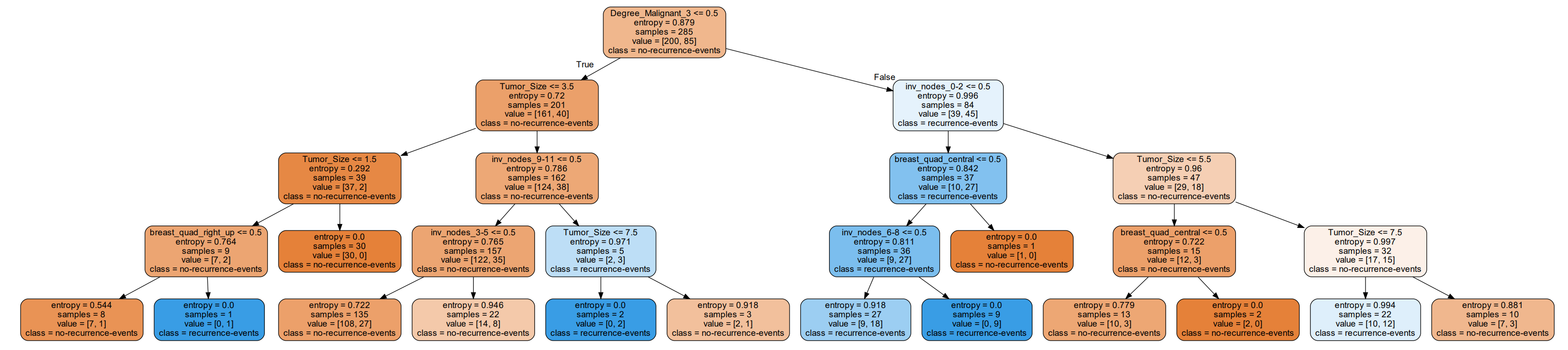
The accuracy achieved by applying our information gain, utilizing those attributes, and applying the best training set size that was calculated was an **accuracy of 93%** overall for this set. Nice!



There are several PDF files linked within the project to view the decision trees created with all the data. This version was pared down from an even larger tree:



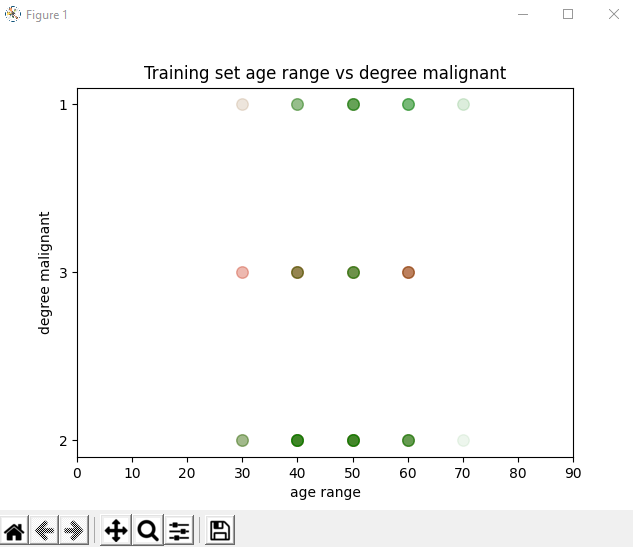
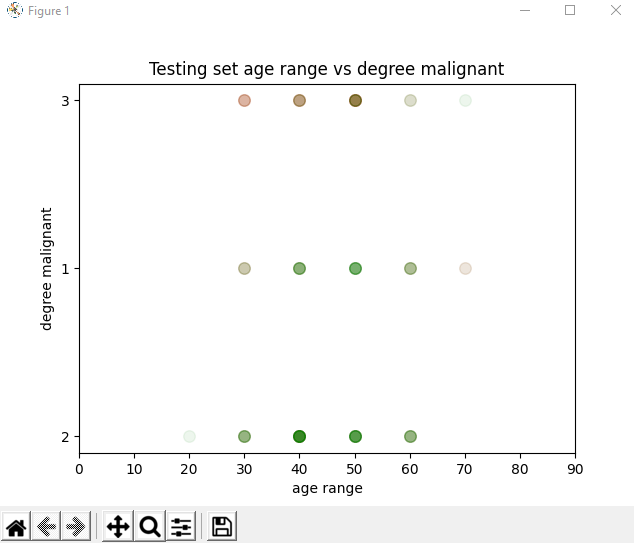
down to the max depth level 4 for which we predicted offered us the most in terms of accurate data:



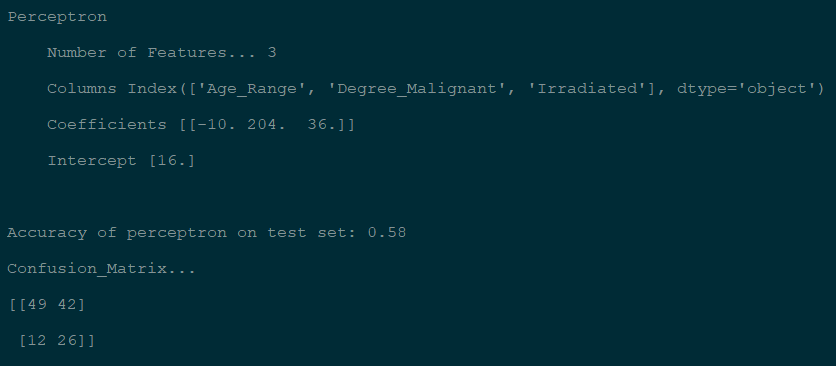
That is much more manageable and if we were working with an even larger data set, would make memory management better as well.

**Model 2: Perceptron and Logistical Regression**

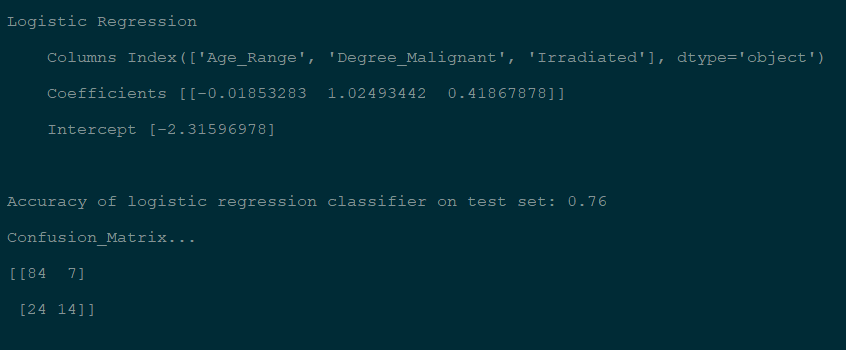
In creating a Perceptron, we try to find attributes that can be binarily separable or classified. I did not evaluate independence on my data set attributes to help me determine this. Looking in hindsight, that would have been the path to go to determine what was possible with this data set. Instead, I looked for clues within the information gain from the decision tree model and did notice a strong correlation between the tumor size, degree malignancy, and involved number of nodes…which makes sense. The higher degree of malignancy would imply a larger or more involved tumor cluster. I did notice that age range and irradiation had less information to offer and so started, by plotting ‘age range’ vs. ‘degree malignant’ in the context of ‘recurrence’ and eventually added ‘irradiate’ into the mix as well as it enhanced the accuracy of the data. Set up with a training set size of .45 as this seemed a good fit for the decision tree data, these were the results:

Both, the training, and the testing sets were plotted with alpha layering so that we could observe the overlaps and density between recurrence (red) and cancer free (green) results for each attribute. Clearly, those women within ages of 20 and 70 were the bulk of data, though I am unsure as to why the y axis, degree malignant, values fluctuate on the graphs unless this relates to values of data in order of hierarchy showing up? Yet, we can see that the data is not ideally separated linearly, however, there is some separation. I believe therefore the perceptron and linear regressions comes out with less accuracy than did the decision tree model:



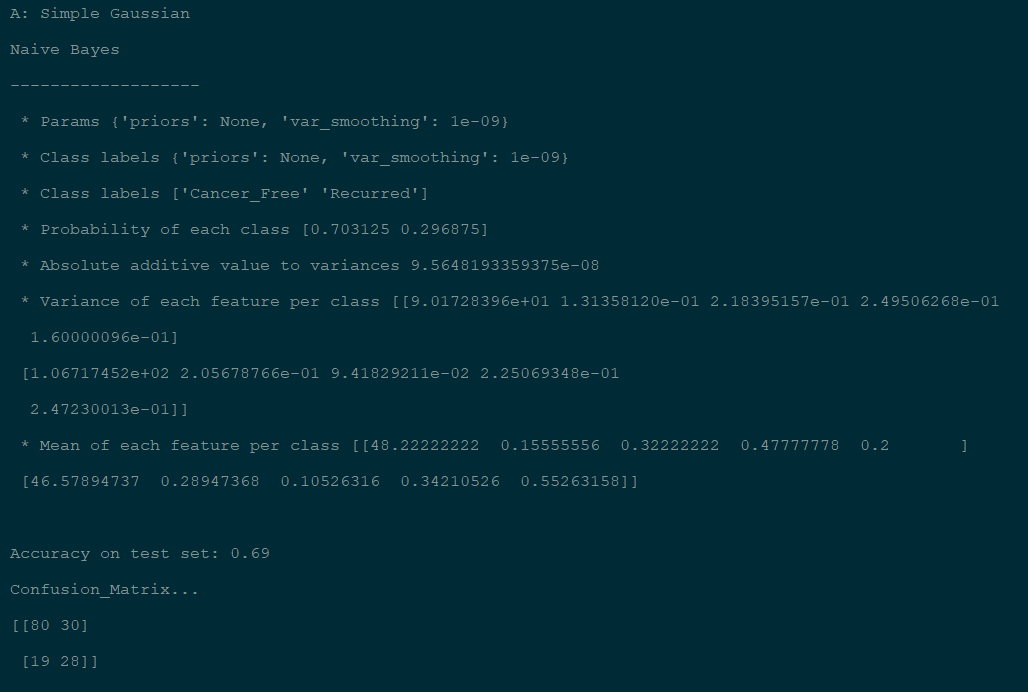
The Perceptron, with 3 features only provides us with an overall accuracy of 58%. However, this model offers more numerous true positive and true negative predictions than false ones as seen by the confusion matrix.

The logistic regression results in a better overall accuracy of 76%, though not the more ideal 93% we found in the decision tree model, but in this confusion matrix, we see a higher incidence of false negatives which might be concerning when trying to predict recurrence in cancer.

**Model 3: Naïve Bayes**

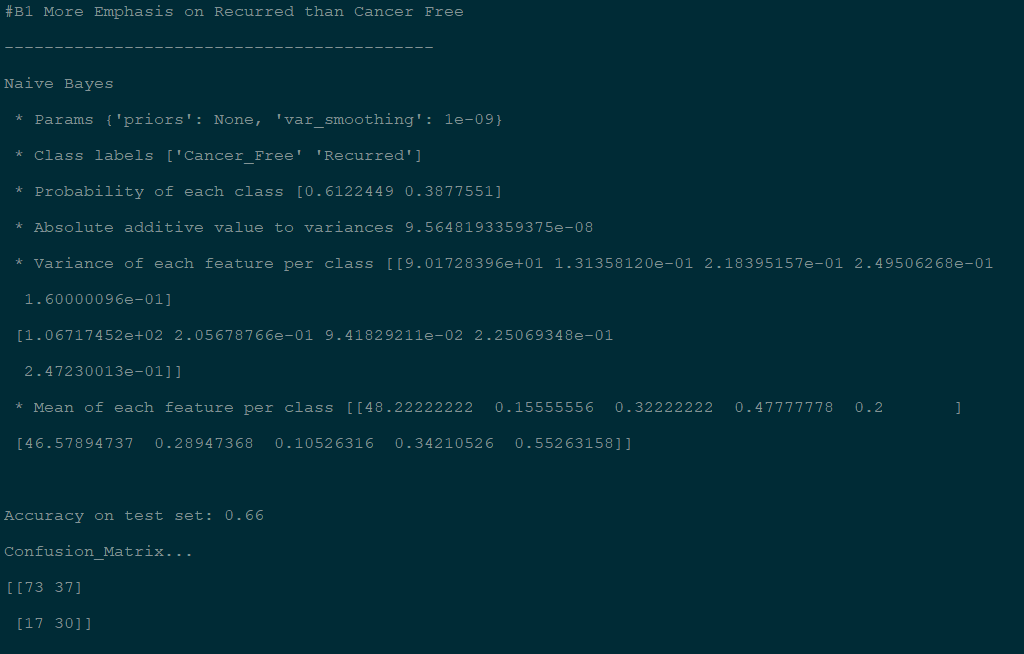
In this model, we play around with gaussian distributions and weighted attributes to see what kind of predictions we can make with our data. Again, we compare ‘age range’ with ‘degree malignant’ and ‘irradiate’ and first apply a gaussian distribution. Keeping in line with previous models using a training set size of .45, I’ve applied a testing size of .55 here:

**Simple Gaussian**



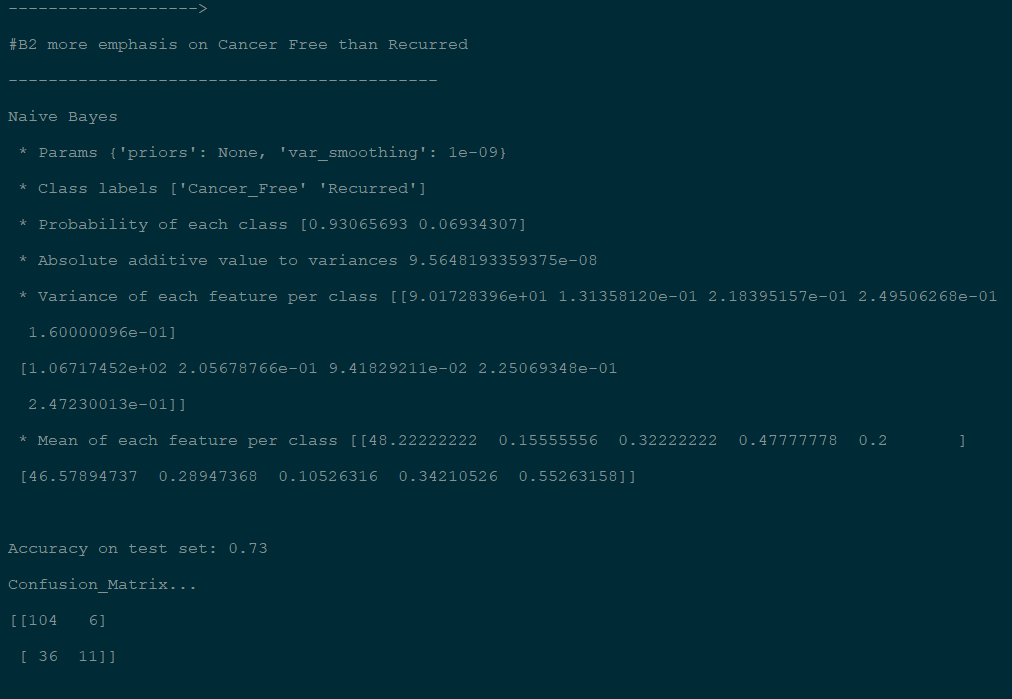
Simple gaussian only offers a 69% overall accuracy, though the confusion matrix does reveal that we have more true predictions than false ones for both positive and negative events.

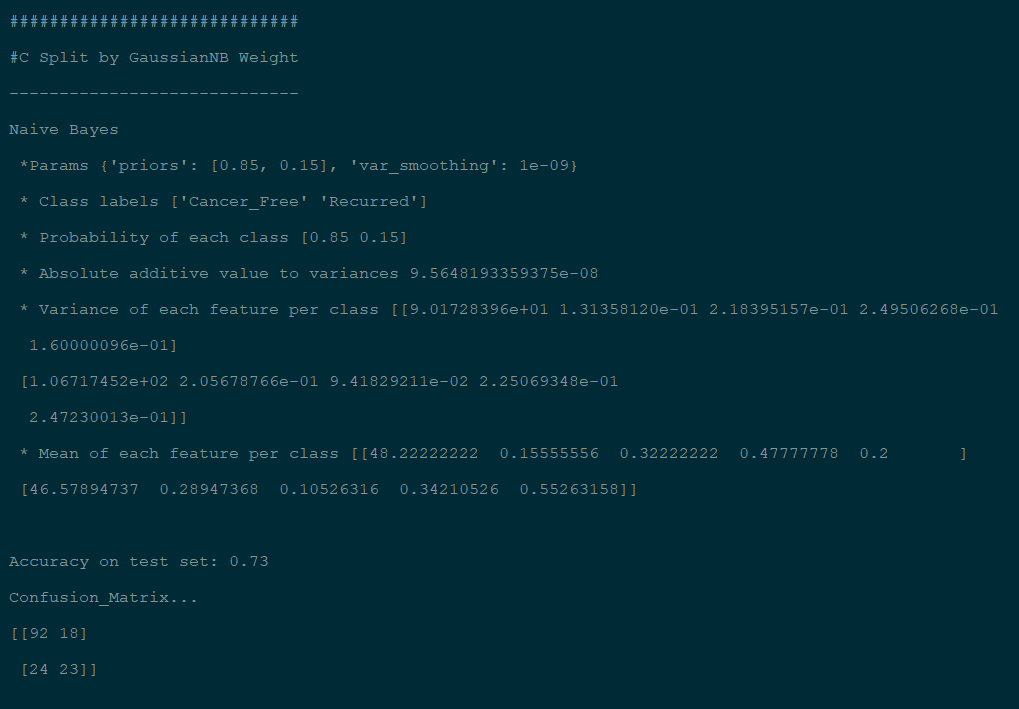
**Weighted Gaussian**

*****Weighted heavier on ‘recurred’*

Weighting the classifier with more emphasis on ‘recurred’ required only a minimal weighted advantage as the data seems to weigh more toward recurrence anyway. With weights of 60% and 40% respectively to recurred vs cancer free we get an accuracy of 66% with a well distributed confusion matrix…more true values than false ones.

*Weighted heavier on ‘cancer free’*

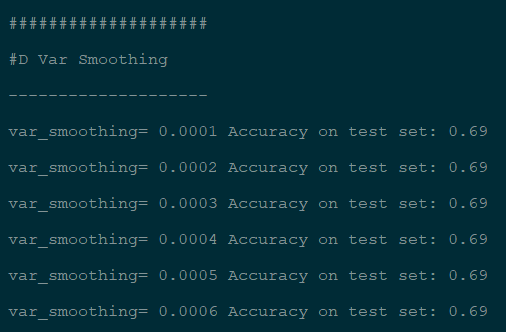
To get a similar accuracy in weighing ‘cancer free’ attribute higher, we are required to add more impact to the weighted value. Here we apply 85% and 15% to ‘cancer free’ and ‘recurred’ respectively. And, we get a 73% accuracy and an even better confusion matrix with more true values, however, the false negatives are higher.

**Split by Gaussian NB Weight (adding Priors)**

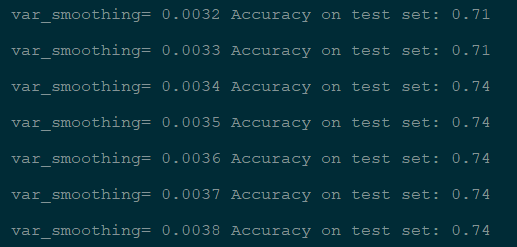
We incorporate the addition of priors or weighted values up front for each of the attributes, essentially working in the same manner as weights per attribute. We get the same accuracy though a slightly different confusion matrix.

**Variable Smoothing**

This calculation did something a bit odd. I was expecting the calculations to grow in accuracy, but they fluctuated going up and then down, but overall was still a higher ending accuracy compared to the starting value.



. . . . . . . . . . . . . . .



. . . . . . . . . . . . . . .

