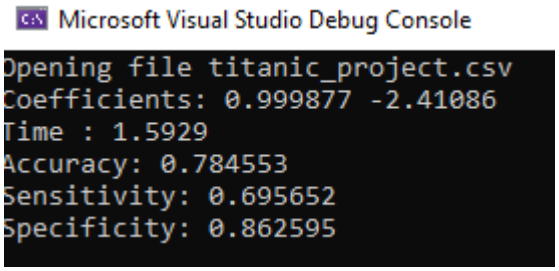
Benton Fariss
CS 4374.004

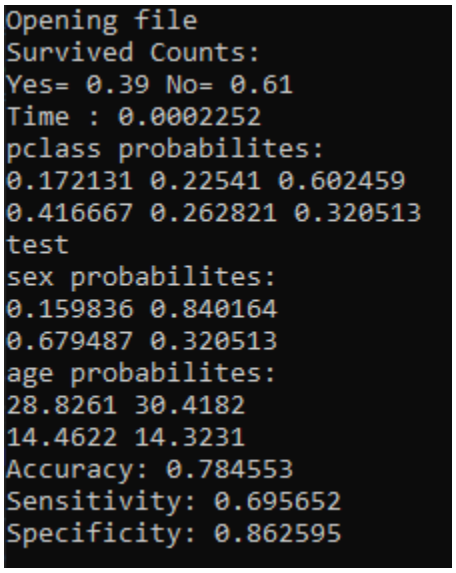
ML Algorithms from Scratch

a.

Logistic Regression:



Naive Bayes:



b.

The results of my Logistic Regression program are overall pretty good. First, we want our coefficient to be as close as possible to 1. In this case, the coefficient of .999869 is extremely close and proves that sex is a good predictor for “survived”. The accuracy is .784553=78.45% which is good enough. The sensitivity is .695652=69.53% which is the true positive rate, meaning that is the percentage of true positives, to calculate the percentage of false negatives we would do 1-.695652. The specificity is .862595=86.26% which is the true negative rate, in our case, that means that we had many more true negatives than false positives which is a very good result.

The results of my Naive Bayes program were interesting as well. The interesting part about the Naive Bayes was that it had the exact same accuracy, specificity, and sensitivity. This makes me conclude that this data was well suited for both types. Naive Bayes would be better however, because it runs at a much faster rate.

c.

Generative models are good because they need less data to train, however, they are much more biased. They are also good to use if you are working with a dataset that may have missing information, they can make assumptions upon the data and ignore the missing data points. Generative models also are more widely used, meaning, they can accomplish many more types of tasks than discriminative models. Because of all of these upsides, they are much better for unsupervised learning and they have a much greater impact on outliers in the models.

Discriminative models cannot do all of these assumptions and they cannot create any data points. However, discriminative models are good because they are more accurate and less computationally expensive. These models are only useful for classification problems. These models' main goal is to learn the decision boundaries, unlike generative which is modeling the underlying data distribution. Overall, if you are performing classification on a dataset that does not need any assumptions made on it, this is the faster and better classifier model.

Garg, Saurav. "Deep Understanding of Discriminative and Generative Models in Machine Learning." Analytics Vidhya, 16 July 2021, https://www.analyticsvidhya.com/blog/2021/07/deep-understanding-of-discriminative-and-generative-models-in-machine-learning/.

d.

Reproducible research in the context of machine learning means that whenever you run your algorithms, you will get the same results. It also means that when someone else tries to reproduce your results using your algorithms, they will also receive the same results. In machine learning this is an especially important concept because if our algorithms do not replicate relatively similar results that can cause extreme inaccuracies. In machine learning, even when we use the same datasets, we sometimes get different results because when we split the data into test/train we are getting random parts of it meaning that the data that is training the algorithm can be slightly different. The problems with this are that if we get different results then that decreases the confidence and correctness of the results. Another problem is that due to the incorrectness, the machine can give wrong or unsafe results. All of this essentially supports the fact that machine learning without reproducibility is completely useless.

To fix this "crisis" in machine learning, there are a lot of different ideas to prevent this lack of reproducibility. One of them is to use version control to track code changes. This verifies the fact that there have been no code changes in between tests. Another idea is to control the random nature of it by using seed configuration. Finally, the reports for all machine learning research needs to be extremely thorough and well-documented. If people are able to make their documentation easily-interactable and user-friendly that is also preferred.

Sipes, Jonathan. "The Importance of Reproducibility in Machine Learning Applications." Decisivedge, 6 Dec. 2018, https://www.decisivedge.com/blog/the-importance-of-reproducibility-in-machine-learning-applicat

ions/#:~:text=Reproducibility%20with%20respect%20to%20machine,reporting%2C%20data%20analysis%20and%20interpretation.

Simonite, Tom. “The Machine-Learning Reproducibility Crisis.” Wired, Conde Nast, 26 Apr. 2018, https://www.wired.com/story/machine-learning-reproducibility-crisis/.

Mahajan, Divya and Meng, Xiangrui. “5 Reproducibility.” Machine Learning @ CMU, 31 Aug. 2020, https://blog.ml.cmu.edu/2020/08/31/5-reproducibility/.