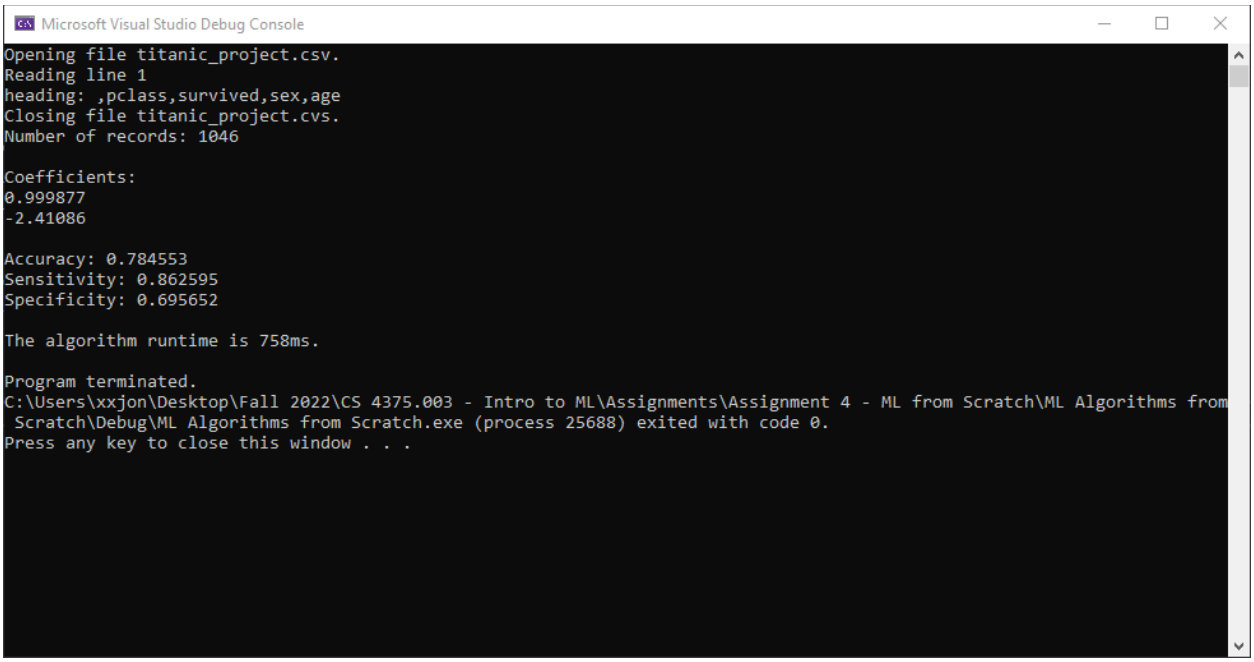
Jonathan Ho

CS 4375

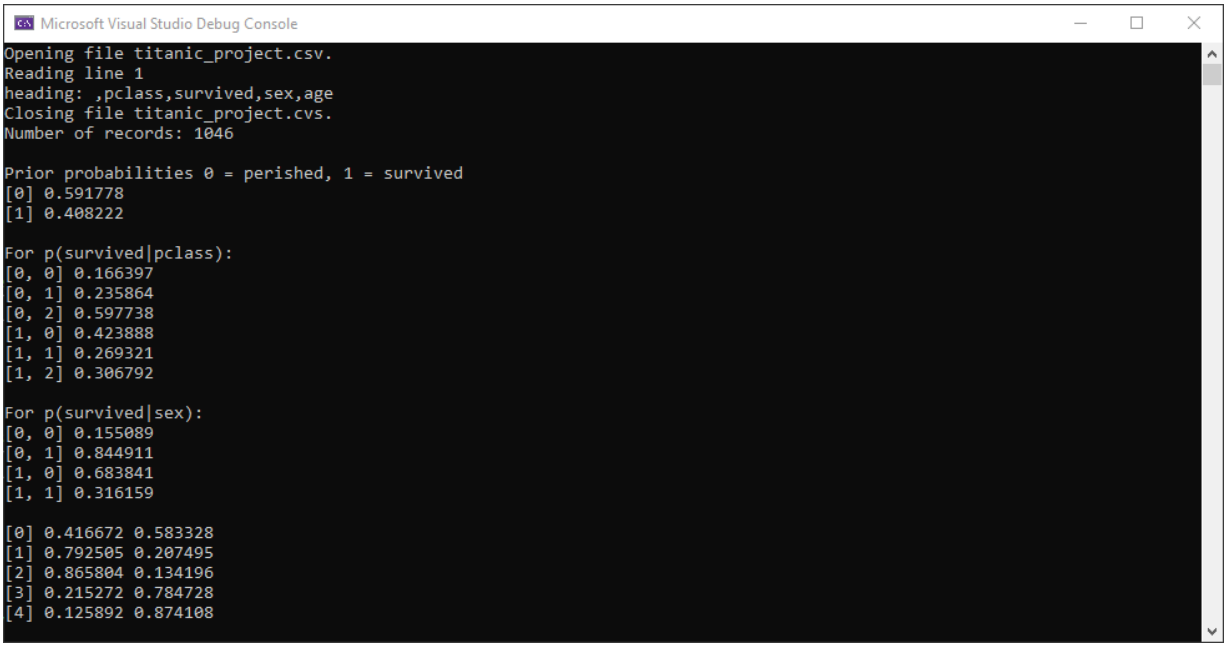
Overview Document

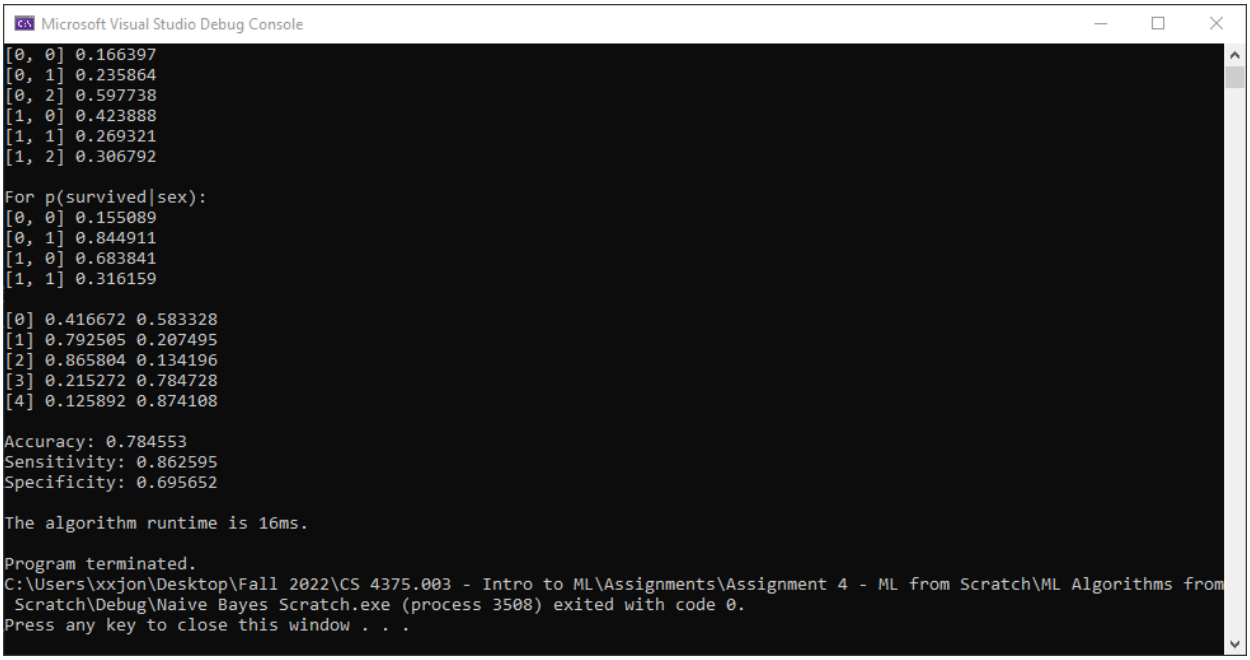
Output:

Logistic Regression:



Naïve Bayes:





Analysis:

When it came to getting the metrics of both Logistical Regression and Naïve Bayes, I was able to reproduce the same exact numbers. This is most likely due to the data set being used was made specifically for this assignment. Comparing these metrics with R given the train and test conditions, they are the same as well. The only real difference is the algorithm runtime between the two. Logistical Regression took significantly longer than Naïve Bayes. This is most likely due to the portion of Logistical Regression that requires running the Gradient descent at many iterations to make sure each point fits along the curve. Each iteration of the Gradient descent loop makes the total time complexity approximately $O(n^3)$ since a function must be run for each line.

Generative vs. Discriminative Classifiers:

Discriminative classifiers are those that try to find boundaries separating classes. It estimates what the parameters of P(Y|X) are. All potential threshold values for boundaries are generated and the lowest error one is accepted. Depending on the algorithm a discriminative classifier uses, the boundaries found could be hard or soft. Hard would imply very few to no data is misclassified. Soft means that some data could be misclassified. One of the techniques that uses this is logistical regression.

Generative classifiers are ones that try to find the joint probability distribution P(X, Y). In other words, it estimates the parameters for P(X) and P(X|Y). The main goal of generative classifiers is trying to explain how the data is created. This is done through focusing on the relationship between features and the target variable. Overall, a generative classifier is able to learn the underlying data distribution. When the joint probability distribution is found, Bayes rule is used to transform it into a conditional probability P(Y|X). One of the techniques that uses this is Naïve Bayes.

(Source: https://towardsdatascience.com/generative-vs-discriminative-classifiers-in-machine-learning-9ee265be859e and the text book).

Reproducible research in ML:

Reproducible research in ML is research that can be recreated or copied. This means that a machine learning workflow can be recreated to obtain the same exact conclusions as the original work [2]. This process encompasses design, reporting, data analysis, and interpretation. If ML research is not reproducible, then the credibility of it goes down and may not be reliable [1]. This has plagued the field of ML as there is a "reproducibility crisis" caused by lack of transparency and reporting while a data-drivel model is being built [3].

There are many reasons why research must be reproducible. One example is an algorithm that is developed from new research but cannot be reproduced would be difficult to investigate and implement. If this algorithm was implemented into a ML or AI system, then there could possibly be major consequences unforeseen [2]. Not only is trustworthiness a factor, but it helps to reduce errors and ambiguity moving projects from development to production. Reproducibility ensures data consistency, which leads to confidence in peoples' ability to confirm correct results from a model [1]. Reproducible ML research can also naturally scale with business growth [1].

There are many ways to improve or implement reproducibility. One way is through data reporting. Since the quality of a model relies significantly on the quality and characteristics of the data, the data collection process needs to be properly discussed and reported. This helps to look more into the data as there could be biases which may or may not have been known. This also makes researchers aware of any limitations while using the model. Data cleaning and data curations are also crucial, so they must also be reported in detail [3]. Another way is through model reporting. Since more complex models can lead to lower transparency, interpretability, and increased training times, models and their level of complexity should be justified. Reporting how long a step takes is also essential as it helps the reproducer know the practicality of a step given their current resources [3]. Lastly, it is important that all the pertinent code and data are available publicly. This includes the code to train, validate, and test the model as well as code for data collection, cleaning, and curation. Report of what hardware was used as well as software library versions should be available, as well as the training models themselves [3].

Sources:

[1] https://www.decisivedge.com/blog/the-importance-of-reproducibility-in-machine-learning-applications/#:~:text=Reproducibility%20with%20respect%20to%20machine,reporting%2C%20data%20analysis%20and%20interpretation

[2] https://towardsdatascience.com/reproducible-machine-learning-cf1841606805

[3] https://www.nature.com/articles/s43588-021-00152-6