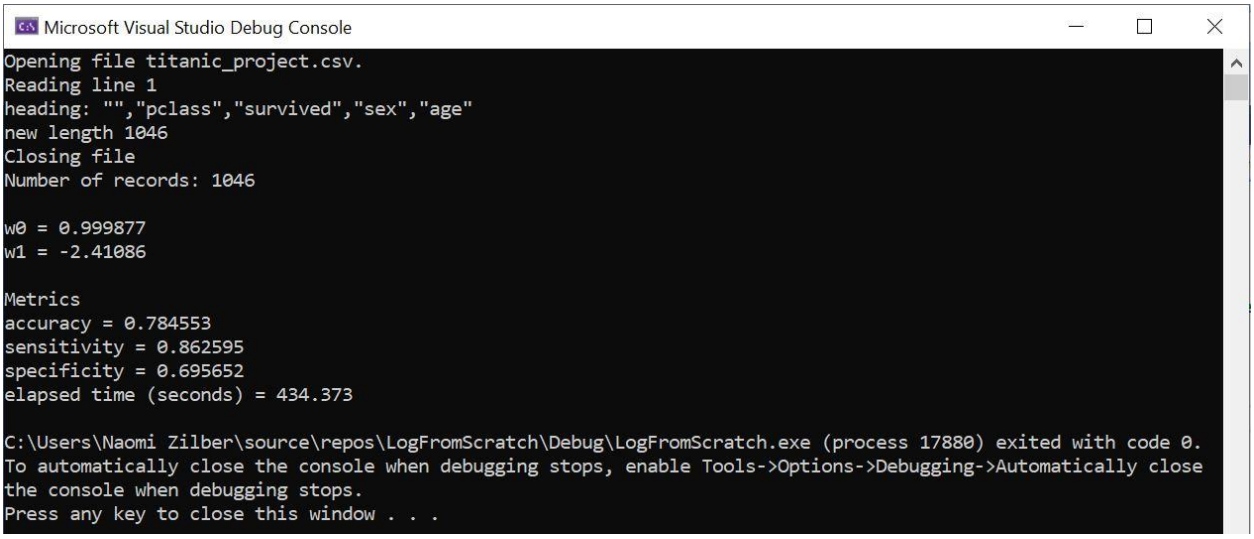
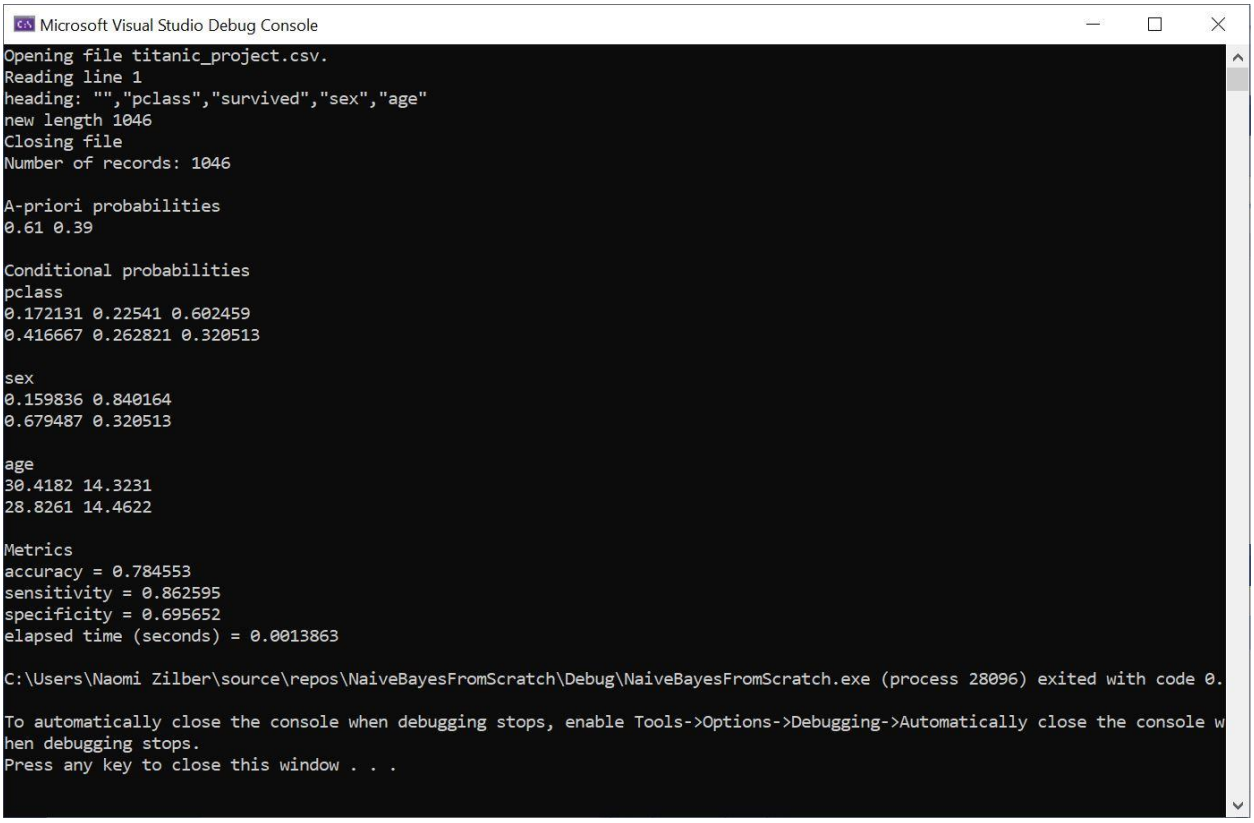
ML Algorithms from Scratch

Output of code:

Logistic regression output:



Naïve Bayes output:



Analysis of results:

When comparing my results from the C++ programs and the R Studio notebooks, I obtained the same results (metrics). I found that for the logistic regression program, the number of iterations that I needed to run gradient descent didn't have to be as big as 50,000 to get the right weight coefficients, but even 1000 was enough iterations. Also, it took the algorithm around 7 minutes to run the 50,000 iterations, which is extremely long, while 1,000 took around 9 seconds, which is much quicker. On the other hand, the naïve Bayes algorithm took only around 1 ms to run. This shows that in terms of computation speed, the naïve Bayes algorithm did much better than the logistic regression algorithm.

When analyzing the test metrics, I saw that the logistic regression model and the naïve Bayes model both had the same accuracy, specificity, and sensitivity, which means that both models performed equally well on the data, and given that the accuracy was relatively high, it can be concluded that the models were good fits for the data.

From the logistic regression model, it seems like being male results in a -2.41 change in the log odds of survival for a one-unit change in the predictor sex.

Generative classifiers vs Discriminative classifiers:

A discriminative classifier directly estimates the parameters of P(Y|X) while a generative classifier directly estimates the parameters for P(Y) and P(X|Y), where X is the inputs and Y is the label (Mazidi 141). This means that discriminative classifiers model the posterior P(Y|X) and therefore "learn a direct map from inputs X to the class labels," while generative classifiers estimate the joint probability P(X,Y) and make predictions using Bayes rule to calculate P(Y|X), after which the most likely Y is picked (Ng and Michael 1).

A discriminative classifier separates data points into different classes, meaning that it separates data space into different classes by learning the boundaries, while a generative classifier generates new data points, so it attempts to understand how the data is embedded into the space ("Generative Models vs Discriminative Models"). Additionally, outliers have almost no effect on a discriminative classifier, but are more prone in a generative classifier. Lastly, discriminative classifiers are more used for supervised tasks while generative classifiers are more used for unsupervised tasks.

Reproducible Research in Machine Learning:

With respect to machine learning, reproducibility refers to the ability to reproduce the same results of a particular project after repeatedly running the algorithm on a certain data set (DecisivEdge). Therefore, reproducible research in machine learning refers to research whose results can be obtained by repeatedly performing the algorithm on certain data sets, meaning that independent verification of the results can be made (Drummond).

Reproducibility is important for several reasons. First, it makes continuous integration proceed smoothly because it makes it possible to build on previous work which can speed the scientific progress (Drummond). Secondly, it helps reduce any errors and ambiguity that could occur

during a project's transition from development to production since it ensures data consistency (DecisivEdge). Lastly, a machine learning application that is reproducible is naturally built to scale with the business's growth, and it is more credible and trustworthy since the results are guaranteed.

Reproducibility can be implemented by documenting the project, including certain decisions and important details about the project. It also requires capturing the computational environment, such as the libraries used in the project, provenance, which is the data sets used and execution parameters, and the scientific narrative, which describes the particular decisions that were made and why (LeVeque et al. 15). Therefore, to implement reproducibility, researchers would need to not only publish their paper but also their computational tools and the data used to generate their paper's results (Drummond).

Works Cited

Drummond, Chris. “Reproducible Research: a Minority Opinion.” *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 30, no. 1, 2018, pp. 1–11, https://doi.org/10.1080/0952813X.2017.1413140.

“Generative Models vs Discriminative Models for Deep Learning.” *Generative Models vs Discriminative Models for Deep Learning*, Turing Enterprises Inc, 22 Apr. 2022, https://www.turing.com/kb/generative-models-vs-discriminative-models-for-deep-learning.

LeVeque, Randall J., et al. “Reproducible Research for Scientific Computing: Tools and Strategies for Changing the Culture.” *Computing in Science & Engineering*, vol. 14, no. 4, 2012, pp. 13–17., https://doi.org/10.1109/mcse.2012.38.

Ng, Andrew Y, and Michael I Jordan. “On Discriminative vs Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes.” Stanford University.

Mazidi, Karen. *Machine Learning Handbook Using R and Python*. 2nd ed., 2020.

“The Importance of Reproducibility in Machine Learning Applications.” *DecisivEdge*, 7 Dec. 2022, 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.