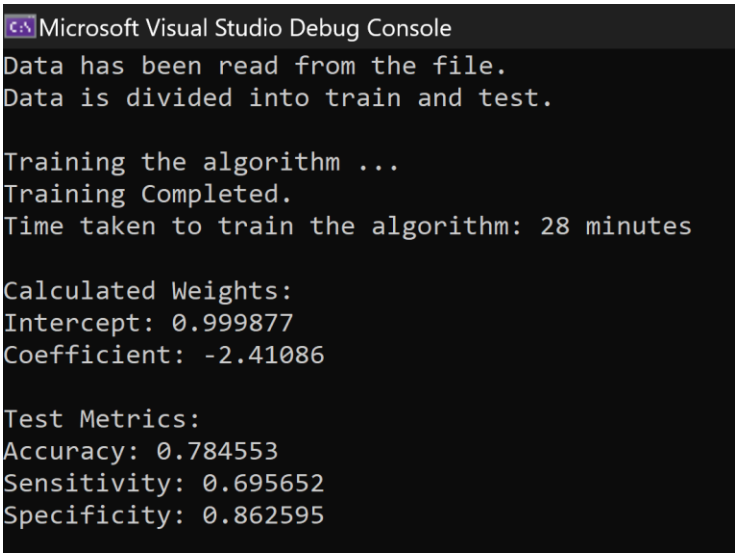
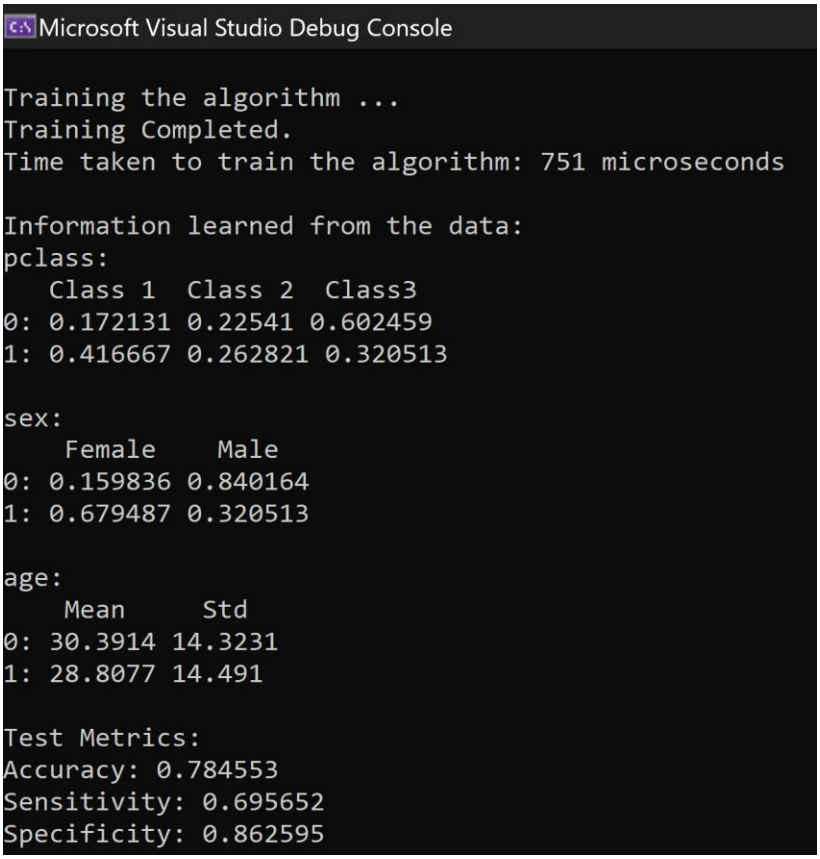
ML Algorithms from Scratch

A- Output of Program 1:



A- Output of Program 2:



B- Analyze the results:

By performing logistic regression on the data, we calculated the coefficient for the variable 'sex' and the intercept. The coefficient tells us the log odds of change in our target variable. For our data, this means that, for one unit of change in x, the probability of survival changes by exp(-2.41)/(1+exp(-2.41)). The intercept is used to fit the data.

By performing Naïve Bayes on the same data, we calculated the likelihood of survival based on each attribute. The probabilities of survival for the three classes of passengers are 41%, 26%, and 32% respectively. The probability of survival for female passengers was 67.9%, and the probability of survival for male passengers was 32%. The mean age of people who survived was 28, and the mean age for people who died was 30.

Now, if we look at the test metrics for both tests, we can see that both of the models had the same performance. They had an overall accuracy of 78%, sensitivity of 69.5%, and a specificity of 86%. This means that both models were 16.5% better at identifying true negatives, than they were at identifying true positive.

Lastly, if we look at the time it took to run both of the models, we can see that the Logistic Regression model took 28 minutes to achieve the same thing that the Naïve Bayes model could achieve in only 751 microseconds. This means that if we are given a choice of implementing Logistic Regression or Naïve Bayes on a data set, we should try to use Naïve Bayes because it will save us both time and money.

C- Generative vs Discriminative Classifiers:

Let's consider a case where we need to predict the value of the target variable y, based on the input value x. Generative classifiers will first learn the joint probability P(x, y), in order to explain how the data is generated. Then, after the model captures the data generation process, it will use Bayes rule to convert the joint probability P(x, y) to conditional probability P(y | x). Lastly, the model uses those conditional probabilities to make predictions on the new data points.

On the other hand, Discriminative classifiers try to find a distinct separation between two classes, where all the values of one class is on one side, and all the values of the other class are on the other side. Also, discriminative classifiers directly learn the conditional probability P(y | x), instead of figuring out how the data is generated.

Choosing the right classification method depends on the specific task because both of these methods have their pros and cons. For example: Generative classifiers require more data to accurately represent the joint probability distribution; Discriminative classifiers are more robust against outliers; Generative models require more computational power, making it more expensive than the discriminative models (Yıldırım).

D- Reproducible Research in Machine Learning:

In Machine Learning, reproducibility means running an algorithm on a certain data set multiple times and getting the same results each time. It also means that if someone else runs your algorithm on the same data set, they should be able to get the same result as your original work. This concept is very important when you are making a Machine Learning model because it represents the correctness of the model, it shows that the model is reliable, and that it can be used as a baseline to improve upon. Most importantly, it shows that the results produced from the model are actually through learning from the data and not by randomness (Reproducibility…).

When it comes to implementing reproducibility in Machine Learning, there are different methods suggested by different organizations and different authors. According to the paper I analyzed, in order to generate high reproducibility, the person who created the model should not only share the code and the data, but they should also share the environment used to execute the program. Environment includes all the libraries and dependencies required to run the code on a new machine. This method solves the “it runs on my machine” problem. Three main techniques used for sharing a computational environment is by using a hosting service, providing a container, or by providing a virtual machine image (Tatman).

Resources

"Reproducibility to Improve Machine Learning.", https://www.section.io/engineering-education/reproducibility-machine-learning/.

Tatman, Rachel. A Practical Taxonomy of Reproducibility for Machine Learning Research. 2018, https://openreview.net/pdf?id=B1eYYK5QgX.

Yıldırım, Soner. "Generative vs Discriminative Classifiers in Machine Learning." Medium, Towards Data Science, 14 Nov. 2020, https://towardsdatascience.com/generative-vs-discriminative-classifiers-in-machine-learning-9ee265be859e.