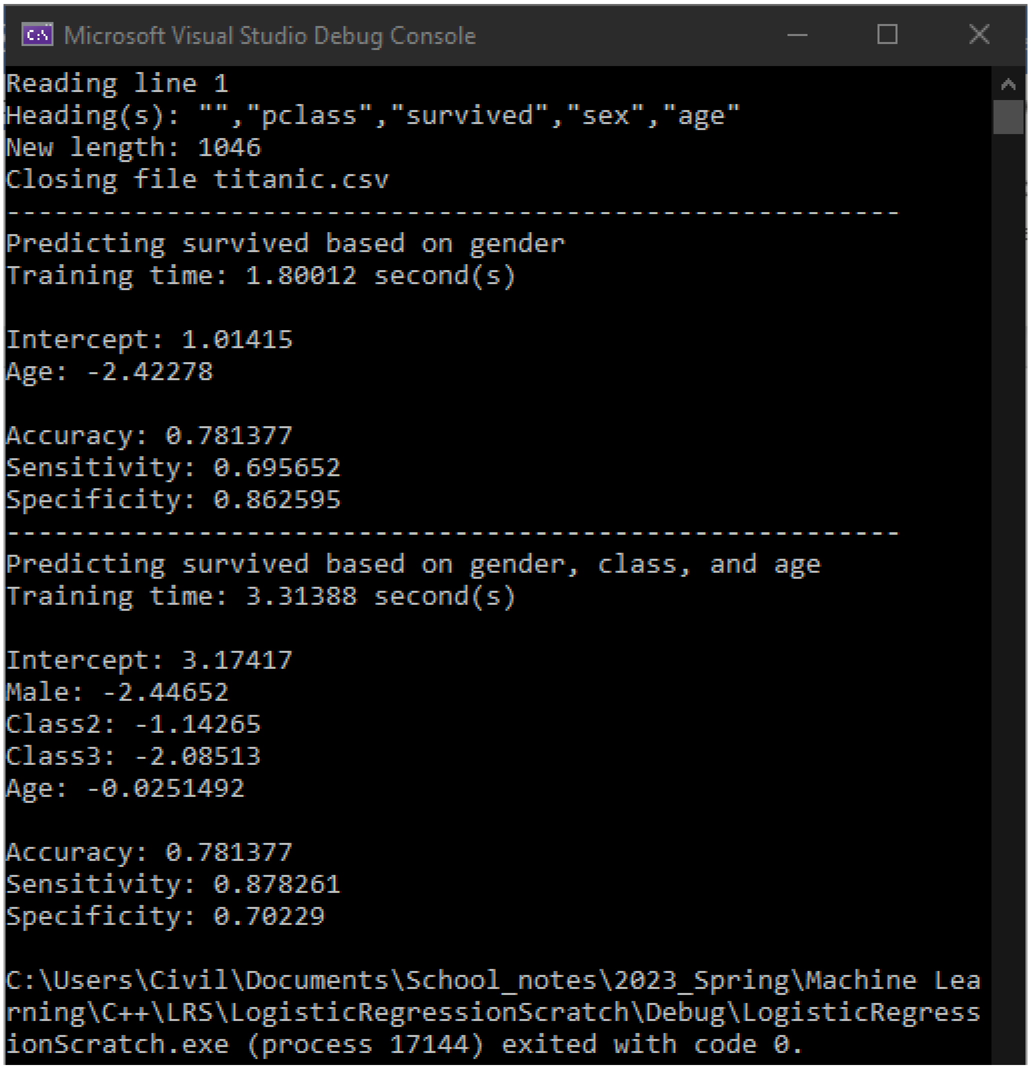
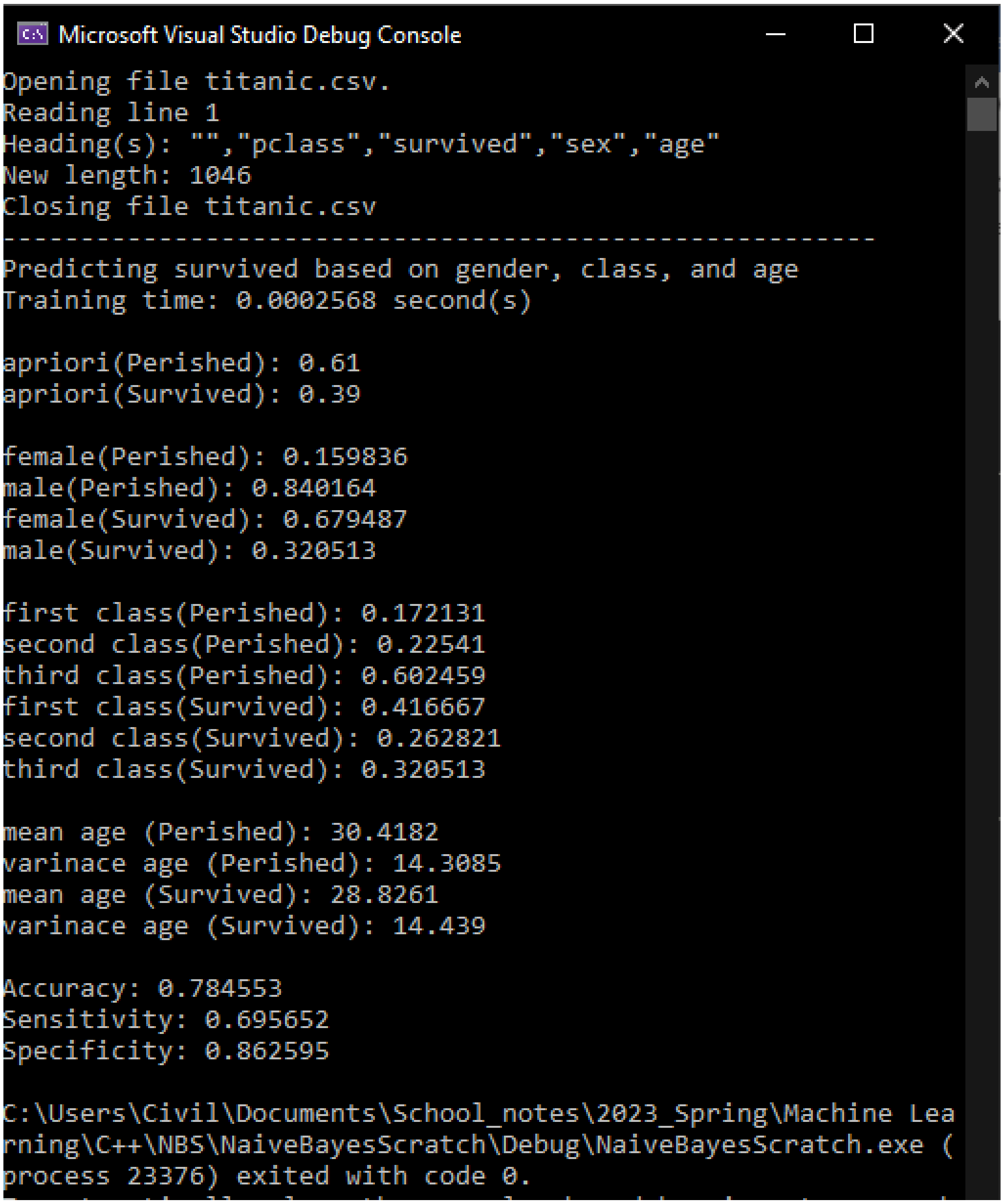
A

Logistic Regression(LR): The two runs are separated by lines. The training time it listed first in seconds. Then the coefficients are listed. Last, the metrics for the test predictions are shown.



```
Reading line 1
Heading(s): "","pclass","survived","sex","age"
New length: 1046
Closing file titanic.csv
--------------------------------------------------------
Predicting survived based on gender
Training time: 1.80012 second(s)

Intercept: 1.01415
Age: -2.42278

Accuracy: 0.781377
Sensitivity: 0.695652
Specificity: 0.862595
--------------------------------------------------------
Predicting survived based on gender, class, and age
Training time: 3.31388 second(s)

Intercept: 3.17417
Male: -2.44652
Class2: -1.14265
Class3: -2.08513
Age: -0.0251492

Accuracy: 0.781377
Sensitivity: 0.878261
Specificity: 0.70229

C:\Users\Civil\Documents\School_notes\2023_Spring\Machine Lea
rning\C++\LRS\LogisticRegressionScratch\Debug\LogisticRegress
ionScratch.exe (process 17144) exited with code 0.
```

Naïve Bayes(NB): The training times is listed first. Then the coefficients for both the perished and survived are listed. Lastly, the metrics for the test are shown.



```
Opening file titanic.csv.
Reading line 1
Heading(s): "","pclass","survived","sex","age"
New length: 1046
Closing file titanic.csv
--------------------------------------------------------
Predicting survived based on gender, class, and age
Training time: 0.0002568 second(s)

apriori(Perished): 0.61
apriori(Survived): 0.39

female(Perished): 0.159836
male(Perished): 0.840164
female(Survived): 0.679487
male(Survived): 0.320513

first class(Perished): 0.172131
second class(Perished): 0.22541
third class(Perished): 0.602459
first class(Survived): 0.416667
second class(Survived): 0.262821
third class(Survived): 0.320513

mean age (Perished): 30.4182
varinace age (Perished): 14.3085
mean age (Survived): 28.8261
varinace age (Survived): 14.439

Accuracy: 0.784553
Sensitivity: 0.695652
Specificity: 0.862595

C:\Users\Civil\Documents\School_notes\2023_Spring\Machine Lea
rning\C++\NBS\NaiveBayesScratch\Debug\NaiveBayesScratch.exe (
process 23376) exited with code 0.
```

B

The training time for LR was longer as compared to NB.

The NB model reveals that by default, one's chances of survival started off at only 39%.

Both models determined being male severely reduced chances of survival. The LR had a -2.4 weight associated with males while NB showed a 84% of the victims were male and only 32% of the survivors were male.

Being of a lower class also reduced chances of survival. LR has a -1.14 weight if the passenger was second class and a -2.08 weight for third class. NB determined that 60% of the victims were third class, 22% were second class, and only 17% were from first class.

Neither model was able to glean much information about the effect of age on the model. There appears to be a very slight trend that younger people survived more. LR has only a tiny -.02 with age. NB shows the average age and variance of age were pretty much the same. It's likely that replacing NAs with the average age has muddled the effectiveness of age as a predictor.

Both models attained a roughly 78% accuracy rate, which is pretty good. Oddly enough, their sensitivity and specificity mirrored each other. LR had a better sensitivity rate of 87% and a lower specificity rate of 70%. NB has a better specificity rate of 86% and a lower sensitivity rate of 69%. This means that LR had better survival predictions and NB had a better perished prediction.

C

The main difference between the two are exactly what the two estimate. In discriminative classification, it estimates P(Y|X). This can be seen in LR, as each predictor is given a weight and its influence on the result is directly measured. Generative classification instead determines the P(Y) and P(X|Y). NB shows this as first the prior of the goal is determined, and then separate observations are made for either value of Y.

The second difference is how well they work with data sizes. NB will typically do better with smaller data sizes while LR will work better with larger data sizes. This is because NB has a higher bias but lower variance than LR. As such, NB is less likely to be influenced by noise in smaller data sets. However, if there is enough data to reduce the effect of noises and the data is actually complex, then LR should do better.

D

Reproducible research means that one researcher should be able to get the same results when performing the same test as another researcher. As stated by Sunita Mall, reproducible research in the context of machine learning means "getting the same output on the same algorithm, (hyper)parameters, and data on every run"[3]. Unfortunately, this is hard to achieve in machine learning given the complexities of it.

This topic is important since it is essentially the only real way to verify machine learning results. As stated by Carnegie Mellon, "Reproducibility is important not just because it ensures that the results are correct, but also because it ensures transparency and gives us confidence in understanding exactly what was done”[1]. This verification is needed to justify any actions taken based on the results of machine learning.

Reproducibility can be implemented with the help of documentation. As stated by DecisivEdge, “The documentation process should explain why certain choices were made, as well as a range of important details needed to successfully execute the project – what Philip Stark refers to as “reproducibility”[2]. If the entire thought process and methods used are properly shared, then this will serve to increase reproducibility as repeat experiments won’t have to guess on what the previous team did.

Works Cited

[1]University, Machine Learning Department, Carnegie Mellon. “5 - Reproducibility.” *Machine Learning Blog | ML@CMU | Carnegie Mellon University*, 31 Aug. 2020, blog.ml.cmu.edu/2020/08/31/5-reproducibility/.

[2]“The Importance of Reproducibility in Machine Learning Applications.” *DecisivEdge*, www.decisivedge.com/blog/the-importance-of-reproducibility-in-machine-learning-applications/. Accessed 2 Mar. 2023.

[3] “Reproducibility in Machine Learning - Research and Industry.” *Suneeta Mall*, suneeta-mall.github.io/2019/12/21/Reproducible-ml-research-n-industry.html. Accessed 2 Mar. 2023.