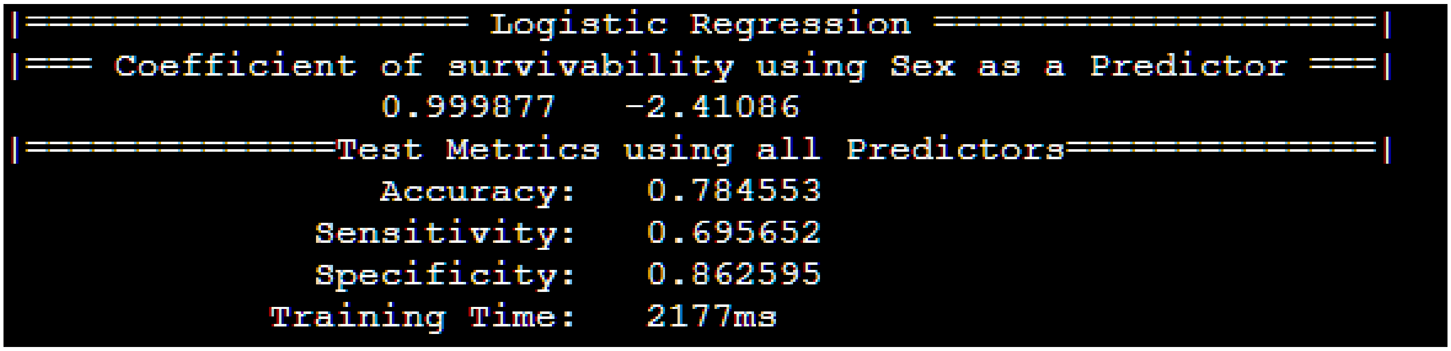
Charles Wallis
CS4375 Intro to Machine Learning
October 2, 2022

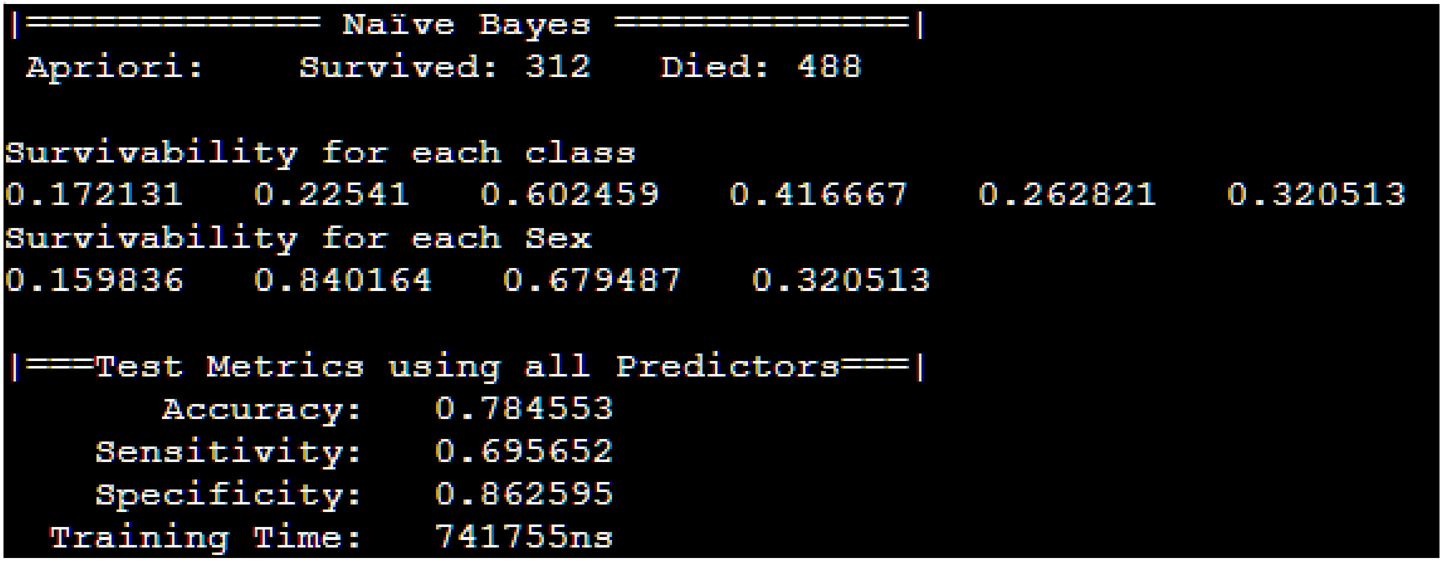
Portfolio Component 3: Machine Learning from scratch using C++

Figure 1: Logistic Regression Output



```
|==================== Logistic Regression ====================|
|=== Coefficient of survivability using Sex as a Predictor ===|
                  0.999877   -2.41086
|==============Test Metrics using all Predictors==============|
                  Accuracy:   0.784553
               Sensitivity:   0.695652
               Specificity:   0.862595
             Training Time:   2177ms
```

Figure 2: Naïve Bayes Output



```
|============= Naïve Bayes =============|
 Apriori:    Survived: 312   Died: 488

Survivability for each class
0.172131   0.22541   0.602459   0.416667   0.262821   0.320513
Survivability for each Sex
0.159836   0.840164   0.679487   0.320513

|===Test Metrics using all Predictors===|
      Accuracy:   0.784553
   Sensitivity:   0.695652
   Specificity:   0.862595
 Training Time:   741755ns
```

The accuracy, sensitivity, specificity value were the same for both logistic regression and Naive Bayes algorithms, but these values are still promising. Based on the 3 predictors (class, age, sex), the model was able to predict a survivability of the titanic passenger by 78.4%. The sensitivity was 69.6% as well, which means it was able to predict a survivor correctly, and the Specificity is very high of 86.3% which means it can predict if the passenger died very accurately.

For the algorithm's time taken, Logistic Regression had to spend a lot more time calculating the weights of each predictors, while Naive Bayes did not need that and rather computed based on apriori and given data. Since both models provided the same results in accuracy, the naive bayes model is more preferable in this case because the algorithm took 0.74ms while logistic regression spent 3000x more time, at 2177ms.

Discriminative and Generative classifiers are similar in that they are both classification models, but they differ in how the data is classified. The "Discriminative model draw boundaries in the data space, while generative models try to model how data is placed throughout the space"[1]. Logistic regression is an example of discriminative classifier[1], as it draws a line between classes to separate them. Naive Bayes is an example of generative classifier[1] since it estimates the probability first to decide if the data is in a specific class.

Discriminative model needs a label information so it is considered supervised learning, and the goal is to calculate and model the decision boundary. It is simpler than the generative model and if there are enough training data, it can perform well. Generative model doesn’t need this said information and it needs some more predictors. If the assumptions are not fitting, the model may not perform well, but if the assumptions are well made, it can perform even when there isn't a sufficient amount of training data. Generative model’s goals are to model the distribution.

“A computational experiment is deemed reproducible if the same data and methods are available to replicate quantitative results by any researcher anywhere and at any time”[2]. In machine learning, this would be for a given data set, implementing the same algorithm, to produce the same output. Repeatedly running the algorithm on a same dataset to obtain the same (or similar) results on the particular project[3] would be considered reproducible. This is important because it ensures credibility of analysis and conclusions made from the study[2]. Reproducibility reduces errors and ambiguity for a better understanding, and also ensures data consistency[3].

Implementing reproducibility into machine learning should begin at the start, and be applied to every aspect.[3] Writing comments(documentation) between code to help understandability of other researchers can also help, and having good naming conventions for readability. The output should also look clean and easily readable, so when the code is ran again, the original output and the new output are easily comparable.

References:
[1]https://www.analyticsvidhya.com/blog/2021/07/deep-understanding-of-discriminative-and-generative-models-in-machine-learning/#:~:text=Core%20Idea,the%20labels%20of%20the%20data.
[2]https://arxiv.org/pdf/2108.12383.pdf
[3]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.
[4]https://blog.ml.cmu.edu/2020/08/31/5-reproducibility/