Program 1 (Logistic Regression) CLI including coefficients and metrics.

As can be seen above, the Logistic Regression algorithm completed with about 78% percent accuracy. The specificity of the model is also relatively high compared to both accuracy and sensitivity. This seems to indicate that the model tends to over predict that a given passenger will not survive.

```
PS C:\Users\sunni\Documents\GitHub\CS4375-Portfolio\component_3> g++ -o p2.exe program_2.cpp
PS C:\Users\sunni\Documents\GitHub\CS4375-Portfolio\component_3> .\p2.exe
Opening file titanic_project.csv:
Reading line 1
heading: "","pclass","survived","sex","age"
New Length: 1046
Closing file titanic_project.csv.

training time: 0.0002408s

Sex Likelihood:
0.159836        0.679487
0.840164        0.320513

Class Likelihood:
0.159836        0.679487
0.840164        0.320513
0.172131        0.416667

Age Mean:
30.4182  28.8261

Age Variance:
235.15  237.311

accuracy: 0.743902
sensitivity: 0.608696
specificity: 0.862595
PS C:\Users\sunni\Documents\GitHub\CS4375-Portfolio\component_3>
```

Program 2 (Naive Bayes) CLI including coefficients and metrics.

The Naive Bayes model performs slightly worse than the Logistic Regression model in terms of accuracy and sensitivity. However, its specificity is almost identical compared to Logistic Regression.

Generative vs Discriminative Classifiers

A generative classifier is a classifier that that is able to "generate" new examples based on previous examples that is has already seen. Naïve Bayes is able to do this because the likelihood estimate has already been calculated. These estimates can be used to create new examples. Logistic Regression is what is known as a discriminative classifier. It is called this because instead of learning the prior probability and likelihood. It learns the posterior probability directly. This means, however, that new training examples cannot be generated because it only knows nothing of the distribution of the data.

Reproducible Research

Reproducible research is the idea of making the code behind major papers and breakthroughs in research accessible, so that the results of the papers may be reproduced in other environments other than the originator. People are pushing for more recognition of reproducible data because they believe that the scientific method is beign under

utilized[1]. One of the foundations of the scientific method is that scientist can individually and independently come to the same conclusions. This is how theories are put into place . If there is enough independent research all leading to the same thing, the phenomenon can be considered a theory. In many data and machine learning fields are not implementing this because the data is simply not accessible to other people. One of the driving reasons that reproducible research is not more common is the lack of clean well documented code and data [2].

A solution to this issue is quite however might be difficult to standardize. Documentation and data cleaning must be done throughout the research process as to not create a larger problem at the end of the project. However, this does not address the issue of the lace of infrastructure for sharing the data. The data associated with these projects is rarely small enough, to be accessible to everyone. One solution may be to create some sort of decentralized method for sharing this information so that internet traffic and storage do not become an issue. However, this leads to a lot more uptime and a great deal more collaboration and cooperation between individuals.

Works Cited

LeVeque, Randall J., Ian M. Mitchell, and Victoria Stodden. "Reproducible research for scientific computing: Tools and strategies for changing the culture." Computing in Science & Engineering 14.04 (2012): 13-17.

Chris Drummond (2018) Reproducible research: a minority opinion, Journal of Experimental & Theoretical Artificial Intelligence, 30:1, 1-11, DOI: 10.1080/0952813X.2017.1413140