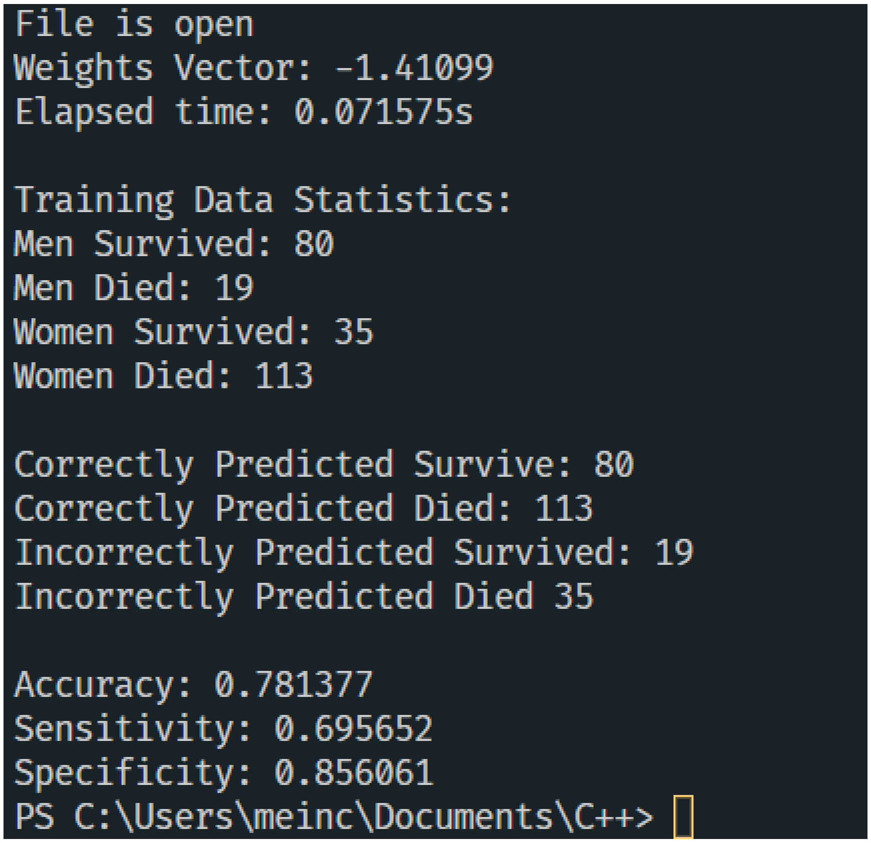
Meinhard Capucao
Khang Thai

Logistic Regression

A.



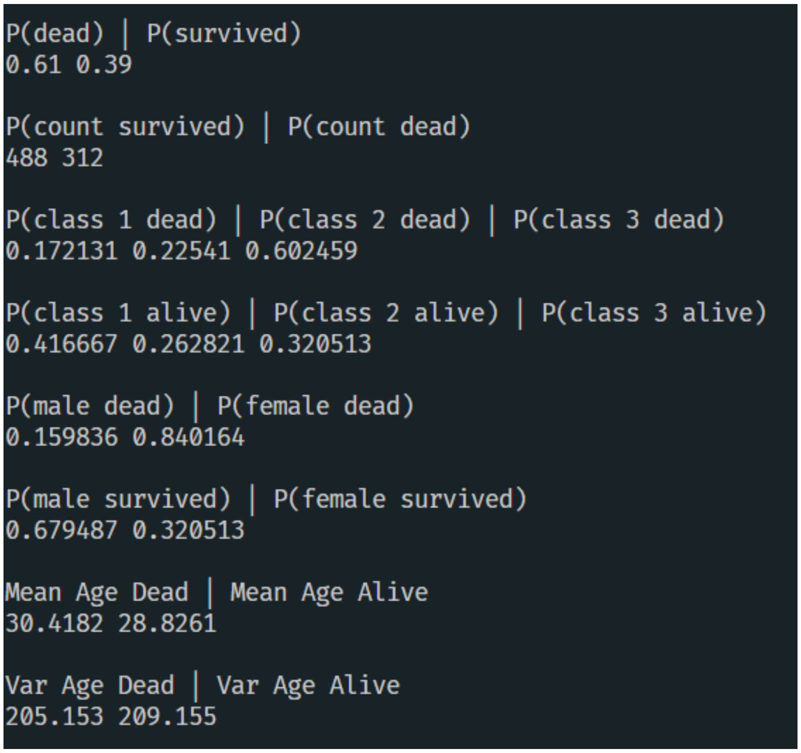
```
File is open
Weights Vector: -1.41099
Elapsed time: 0.071575s

Training Data Statistics:
Men Survived: 80
Men Died: 19
Women Survived: 35
Women Died: 113

Correctly Predicted Survive: 80
Correctly Predicted Died: 113
Incorrectly Predicted Survived: 19
Incorrectly Predicted Died 35

Accuracy: 0.781377
Sensitivity: 0.695652
Specificity: 0.856061
PS C:\Users\meinc\Documents\C++> 
```

Logistic Regression Output

B. Analysis

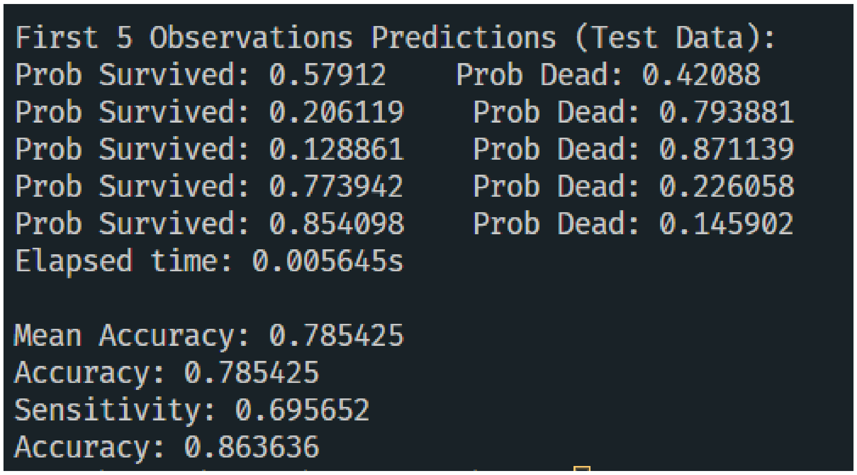
For our logistic regression function, we have sex as a sole predictor for whether a passenger survives or not. We have the weights vector that is affected by the gradient descent function run through 500 epochs. The gradient vector finds the line that best fits the logistic graph for our predictor, then gives us a coefficient of -1.4. As the gender increases by a factor of 1, the chance of survival decreases by log odds -1.4. This explains why our model predicts all men to survive. From the test data there are 109 total men, with 80/109 surviving. Because our only predictor in this C++ logistic regression function is gender, then it is more likely predicted that men survive from the test data than women. Everytime 0 for gender is fed into the logistic regression model we trained, it sees that there are higher log odds for it to survive than if the gender was 1. So, for higher accuracy our model correctly predicted over 78% of survival based on gender. Accuracy was calculated by the percentage of our model correctly predicting if a passenger survived or died. Sensitivity calculates the true positive rate, and specificity calculates the true negative rate. Our model predicted a passengers survival rate from our test data about 78% of the time based on seeing the gender, and had a higher true negative rate.

Naive Bayes
A.



```
P(dead) | P(survived)
0.61 0.39

P(count survived) | P(count dead)
488 312

P(class 1 dead) | P(class 2 dead) | P(class 3 dead)
0.172131 0.22541 0.602459

P(class 1 alive) | P(class 2 alive) | P(class 3 alive)
0.416667 0.262821 0.320513

P(male dead) | P(female dead)
0.159836 0.840164

P(male survived) | P(female survived)
0.679487 0.320513

Mean Age Dead | Mean Age Alive
30.4182 28.8261

Var Age Dead | Var Age Alive
205.153 209.155
```

Calculating Likelihoods for Qualitative and Quantitative Data, and Mean/Var of Age



```
First 5 Observations Predictions (Test Data):
Prob Survived: 0.57912    Prob Dead: 0.42088
Prob Survived: 0.206119    Prob Dead: 0.793881
Prob Survived: 0.128861    Prob Dead: 0.871139
Prob Survived: 0.773942    Prob Dead: 0.226058
Prob Survived: 0.854098    Prob Dead: 0.145902
Elapsed time: 0.005645s

Mean Accuracy: 0.785425
Accuracy: 0.785425
Sensitivity: 0.695652
Accuracy: 0.863636
```

Naive Bayes Output

B.

Our naive bayes model uses the predictors sex, class, and age to determine wheter or not a passenger survives. Naive Bayes works by calculating raw probabilities from our predictors by splitting likelihoods for quantitative data and the mean/variance for qualitative data based on the input given. After calculating all raw probabilities and implementing a function for the probability for quantitative data, we create a function that gives probabilities for survival based on sex, class, and age. If the probability for survival is greater than or equal to than 0.5, then we have the model predict that the passenger survived. With this, our accuracy is about 78.5%, just slightly better than our logistic regression model.

Our run time for both models is relatively fast. For logistic regression, we had it run through 500 epochs for the gradient descent function. If we did more, it would take a lot longer. Naive bayes is quicker than logistic regression since it bases the prediction on a calculated probability function instead of training a lot of times with the gradient descent.

C.

Generative classifiers give a probability for a dataset while a Discriminative classifier will make a prediction on unseen datasets. The differences between the two are based on how it uses the data it is given. Discriminative classifier will receive the data and then make a prediction based on the training data while the generative models find the “prior probability and likelihood probability to calculate a probability” (Goyal).

Generative Classifiers and Discriminative Classifiers are different based on the type of dataset. In terms of performance, Generative Models are better because they need far less data then Discriminative models to make a strong assumption. However, Discriminative Classifiers are able to make predictions even if there is missing data while Generative Classifiers need to take out the missing data or have the missing data filled.

D.

The phrase “reproducible research in machine learning” means that running the same algorithm on certain datasets can obtain the same or similar results. This means that if anyone was given the same data, they would be able to come to similar predictions and probability. Being able to replicate the project allows the project to be more accurate and shows that the dataset is a reliable source to use for predictions.

Being able to reproduce the research in machine learning means that if everyone can achieve the same statistics, the dataset becomes more reliable. If everyone that tries to recreate the project ends up with different predictions and accuracy then the information is questionable if it can even be used to correctly predict something in the future. According to Zihao’s blog, 70% of researchers are not able to reproduce another scientist’s experiment (Ding). Having such a high chance of failure of replication indicates that many projects are too diverse or complex that getting the same results is difficult. If this continues, then eventually the predictions of machine learning become useless as nobody else would be able to replicate the data.

Reproducibility can be implemented to show that there is an improvement made or that using different predictors was able to have higher accuracy or predictions. To implement reproducibility better, according to Decisivedge, having documentation from the beginning and clear reasons as to why the project chose certain things to another helps replicate the data (DecisivEdge). By having clear instructions, people are able to replicate the data and should have the same results, however, according to Zihao's blog, about 6% of all projects share their code and a third share the data (Ding). While replication of projects on machine learning is important, many don't see it that way and would rather keep the data to themselves.

Work Cited

Ding, Zihao. “5 - Reproducibility.” *Machine Learning Blog | ML@CMU | Carnegie Mellon University*, 24 Aug. 2020, https://blog.ml.cmu.edu/2020/08/31/5-reproducibility/.

Goyal, Chirag. “Deep Understanding of Discriminative and Generative Models.” *Analytics Vidhya*, 19 July 2021, https://www.analyticsvidhya.com/blog/2021/07/deep-understanding-of-discriminative-and-generative-models-in-machine-learning/.

“The Importance of Reproducibility in Machine Learning Applications.” *DecisivEdge*, 14 Oct. 2020, 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.