1. Write a document:
   a. copy/paste runs of your code showing the output (coefficients and metrics), and run times

```
Opening file titanic_project.csv.
Reading line 1
heading:"string","pclass","survived","sex","age"
new length 1047
Closing file titanic_project.csv.
Number of records: 1047

Weights/coefficients: 0.999869 -2.41085
Accuracy: 0.788618
Sensitivity: 0.695652
Specificity: 0.870229
Run Time: 0.336992
Program ended with exit code: 0
```

```
Microsoft Visual Studio Debug Console
Opening file titanic_project.csv.
Reading line 1
heading:"","pclass","survived","sex","age"
new length 800
Closing file titanic_project.csv.
Number of records: 1046
Call::
naiveBayes.default(x=x,y=y,lapace-lapace)
A-porior probilities:
Y
0     1
0.61     0.39
Female   male
0  0.159836   0.840164
1  0.679487    0.320513
p class
y    1      2      3
1     0.416667 0.262821 0.320513
0   0.172131 0.22541 0.602459
age
y mean    variance
0  30.4182  204.732
1  28.8261  208.485
Time:0.0043407
```

2.

b. analyze the results of your algorithms on the Titanic data

For Logistic regression, the coefficient 0.999869 is really good as closer it is to 1 the better because it shows that sex is good predictor for whether people survied or not. The accuracy is also pretty high which is good and shows that the model performed well. Sensitivity measures the true positive rate so sensitivity of 0.696 means that 69.6% are true positives and the rest are false negatives which is ok. Specificity is the true negative rate so specificity of 0.87 is good because it means there is a higher value of true negatives and lower false positives.

c. write two paragraphs comparing and contrasting generative classifiers versus discriminative classifiers. Cite any sources you use.

Discriminative classifiers use a decision boundary based on features of the training data. They are used in Logistic regression and are best used for supervised learning. They work better when there are outliers in the dataset but are prone to classifying a data point incorrectly. They also need more data because it needs to learn the predictive variable so it tends to be overfitted. Are computationally less expensive than generative classifiers.

For generative classifiers, the model will try to find the distribution of both the classes in a n-dimensional plane. It will try to model the individual classes using the concept of conditional probability for prediction. If there's a new data point it will check which distribution it is closer to. Are used in naive bayes and best used for unsupervised learning. If there is missing data generative classifiers can still work while discriminative classifiers can't. It needs less data to learn since it may not generalise well with new data due to bias and can also generate new data points unlike discriminative classifiers.

https://www.analyticsvidhya.com/blog/2021/07/deep-understanding-of-discriminative-and-generative-models-in-machine-learning/

https://betterprogramming.pub/generative-vs-discriminative-models-d26def8fd64a

d. Google this phrase: reproducible research in machine learning. Using 2-3 sources, atleast one of which should be academic, write a couple of paragraphs of what this means, why it is important, and how reproducibility can be implemented. Cite your sources using any format.

Reproducibility in machine learning means that every time you run the algorithms on a data set you obtain the same or similar results. Even though the same dataset and algorithm is used the data distribution and slope and intercept will be different because when we split into train and test, a random shuffle of the dataset is performed. For this reason we get different results and reproducibility doesn't always occur..

Some reasons for why it's important are, it improves correctness and data harvesting. Regarding correctness, if we get different results every time an algorithm is run that's a problem for correctness because it can recommend the wrong things. For example, in the past Watson gave unsafe recommendations for treating cancer which obviously isn't good. Another importance is data harvesting. Obtaining a large data set is difficult and takes a long time so creating data synthetically is considered. If something bad happens to the data it's hard to reproduce as synthetic data generation is hard to reproduce. So there is no use if results can't be reproduced.

To implement reproducibility, some of the things we have to do are, control the randomness and version control. To do this we seed the randomness and manage the seed using configuration. Also version control the code so we have the same code to use everytime the algorithm is run to make sure there are no differences.

https://suneeta-mall.github.io/2019/12/21/Reproducible-ml-research-n-industry.html

https://www.determined.ai/blog/reproducibility-in-ml

Heil, B.J., Hoffman, M.M., Markowetz, F. *et al.* Reproducibility standards for machine learning in the life sciences. *Nat Methods* 18, 1132–1135 (2021). https://doi.org/10.1038/s41592-021-01256-7