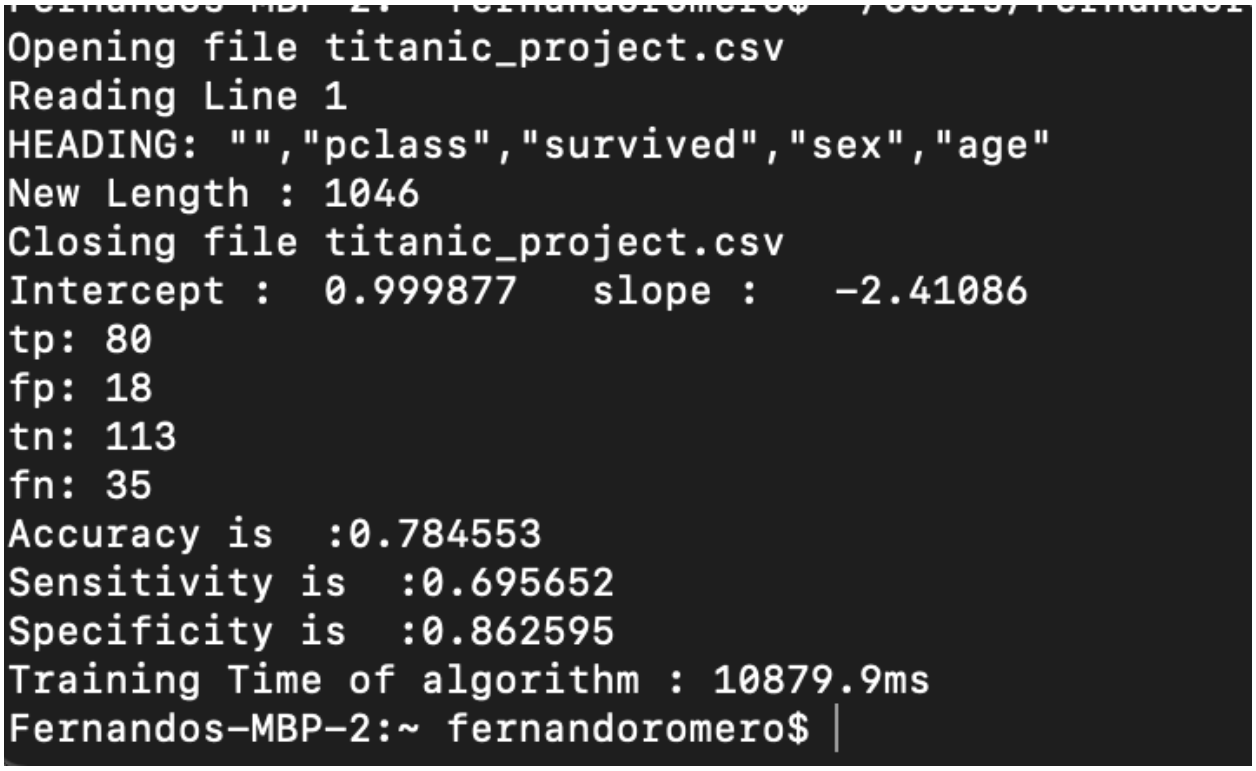
Manuel Romero

Project 1 Logistic Regression



For logistic regression we have an intercept of .999877 and a slope of -2.41086. This gives us an accuracy of 78.4553%, a sensitivity of 69.5652% and specificity of 86.2595%. This is a decent algorithm to find if a passenger survived based on their sex. An accuracy of 78% is decent. The run time being 10871.9 ms.

```
The prior probability of survived = no(0) : 0.61
The prior probability of survived = yes(1) : 0.39
Likelihoods for survival given class:
class   survived        died
c1 :    0.416667        0.172131
c2 :    0.262821        0.22541
c3 :    0.320513        0.602459
Likelihoods for survival given class:
Sex             survived        died
female :        0.679487        0.159836
male :          0.320513        0.840164

survived        age mean        age varr
T               28.8261         209.155
F               30.4182         205.153
Training Time of algorithm Naive Bayes: 0.356514ms
tp: 76
fp: 15
tn: 99
fn: 34
Accuracy is  :0.78125
Sensitivity is  :0.690909
Specificity is  :0.868421
```

For Naive Bayes our accuracy is 78.125%, our sensitivity is 69.0909% and specificity is 86.8421%. This is about the same as logistic regression but for NB our accuracy and sensitivity is slightly lower and specificity is slightly higher. However the training time for Naive Bayes is significantly smaller. It is so much faster not sure if it was an error on my part or just how the code was done.

Generative models understand how the data is spread in space, while discriminative models divide the data and focus mostly on creating a line to separate the boundaries. Discriminative model will separate the data using a line. For example logistic regression creates a line with a slope and intercept that will separate the binomial data.
On the other hand, generative is used to generate more data points and goes further in depth to understand the data. It is prone to outliers.
In other words for classification discriminative is trying to find the boundaries that separate the data in order to classify while generative tries to understand how the data target occurs with the variables in order to be able to get the variables and predict what the target would be given those variables.

Sources
https://www.turing.com/kb/generative-models-vs-discriminative-models-for-deep-learning

https://towardsdatascience.com/generative-vs-discriminative-classifiers-in-machine-learning-9ee265be859e

Reproducible Research in Machine Learning

Reproducibility in machine learning is exactly what it sounds like. Are we able to run the program/algorithm and repeatedly get the same or similar results in each run. If someone else runs the algorithm do they get the same or similar results regardless of what machine they have or software environment they have. While in real life computers are different ideally we would like to be able to get the same results or very very close results regardless of these differences in the machines. But this is often not the case.
Reproducibility should be something to keep in mind when developing algorithms in order to be able to have concise results when developing other projects around the algorithm so that regardless of how it's used we know that the algorithm will perform how it's meant to and reproduce similar results each time.

One way to make sure reproducibility is implemented is by documenting the runs of the algorithm and how it was successfully run, it should include all the important details so that if those details are handed to someone else they are able to reproduce the results. Similar to any scientific experiment, in order for a result to be valid it must be able to be reproduced. For example if we use a dataset and it's the same data values but the first run of the code used a random sample and calculated mean, variance etc, then they made a logistic regression algorithm and had an accuracy of 80%. Then another person did a run of the algorithm but instead they used the first n rows of the data set, their mean, variance, etc can and most likely will be different from the first run, furthermore their logistic regression might have an accuracy of 76%. Another run could use the last n rows of the data set, their results will also vary, one way to fix this would be by passing the subset of data that was used in the first run of the algorithm and then we would be able to reproduce the results. Or specify that n random data points were used so that the other users know that this can account for differences in the result, another way would be to just use the same number n data points starting from the first row this would also ensure that the results can be reproduced by others.

Sources

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.

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