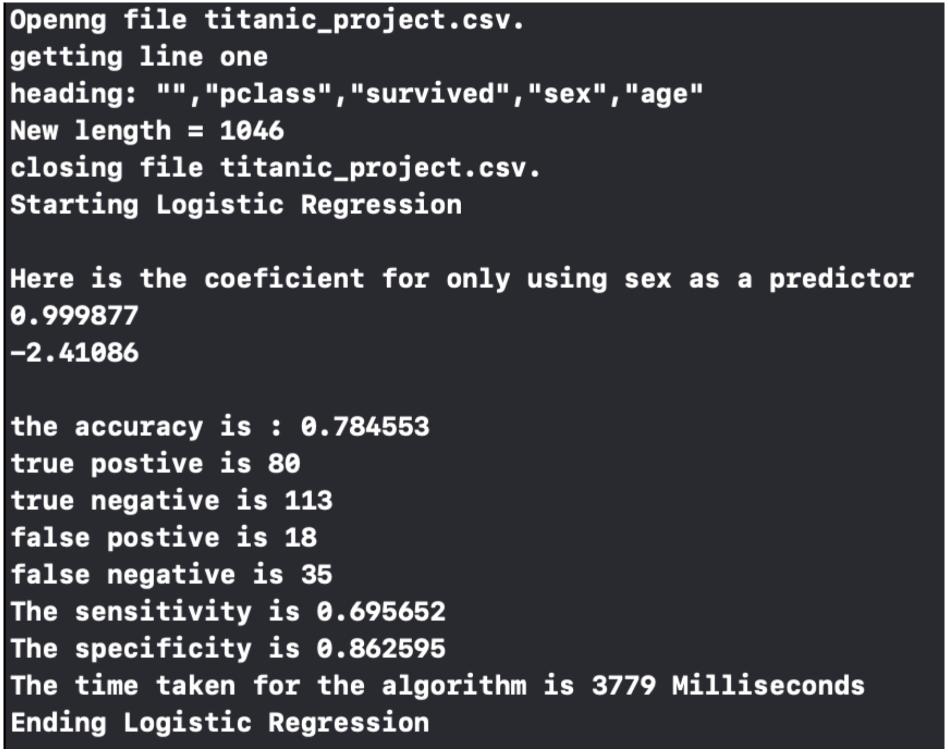
This is for logistic regression

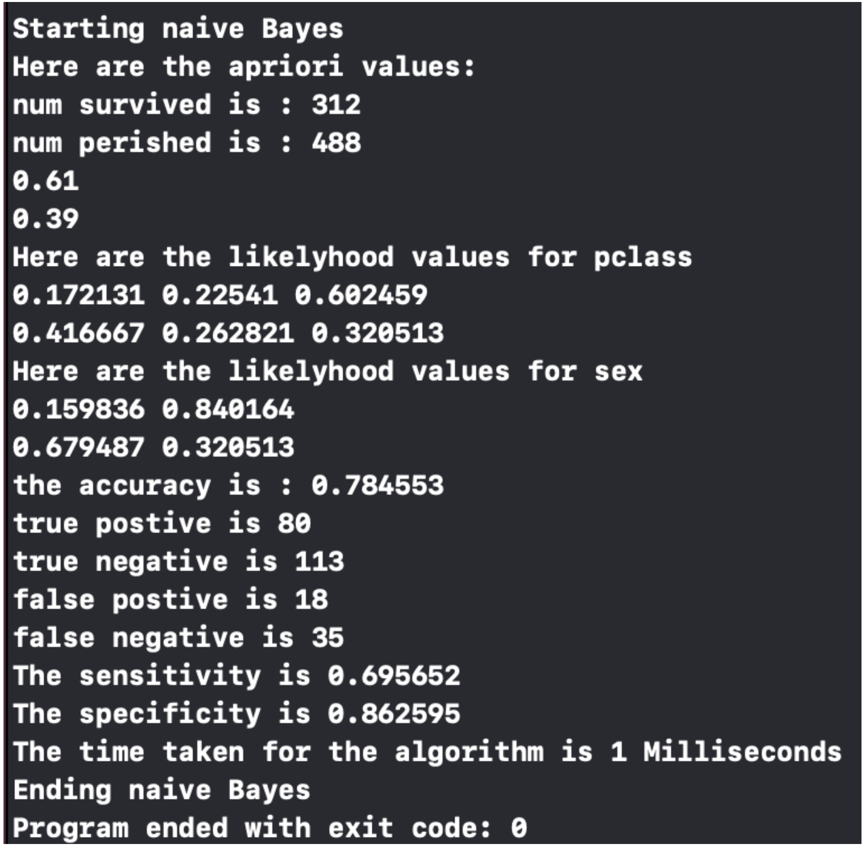


```
Openng file titanic_project.csv.
getting line one
heading: "","pclass","survived","sex","age"
New length = 1046
closing file titanic_project.csv.
Starting Logistic Regression

Here is the coeficient for only using sex as a predictor
0.999877
-2.41086

the accuracy is : 0.784553
true postive is 80
true negative is 113
false postive is 18
false negative is 35
The sensitivity is 0.695652
The specificity is 0.862595
The time taken for the algorithm is 3779 Milliseconds
Ending Logistic Regression
```

This is for the naïve Bayes



```
Starting naive Bayes
Here are the apriori values:
num survived is : 312
num perished is : 488
0.61
0.39
Here are the likelyhood values for pclass
0.172131 0.22541 0.602459
0.416667 0.262821 0.320513
Here are the likelyhood values for sex
0.159836 0.840164
0.679487 0.320513
the accuracy is : 0.784553
true postive is 80
true negative is 113
false postive is 18
false negative is 35
The sensitivity is 0.695652
The specificity is 0.862595
The time taken for the algorithm is 1 Milliseconds
Ending naive Bayes
Program ended with exit code: 0
```

**Analyze the results of your algorithms on the Titanic data**

- **Analyze accuracy**
  - Both the logistic regression model of the titanic data set and the naïve Bayes model produced the same accuracy for predicting the survival of a passenger. The accuracy was about 78% meaning that when predicting on the test dataset the models were right most of the time.
- **Analyze sensitivity**
  - The sensitivity was about 70% for both models and the specificity was about 86% for both models. This means that both models were slightly better at identifying if an individual did not survive the titanic than if they did survive.
- **Analyze time taken**
  - The time taken for the logistic regression model was over 3000 milliseconds while the naïve Bayes model took only a single millisecond. The naïve Bayes algorithm proved to have a significantly faster execution speed then the logistic regression model. The huge amount of time the logistic regression model spent finding the weight values involved going through a for loop thousands of times and executing complex instructions for each iteration of the for loop. The naïve Bayes model did not have such a for loop for training and thus executed in less time.
- **Compare the two models' logistic regression and naïve Bayes.**
  - Overall, even though both models provided the same values for their metrics the naïve Bayes model should be considered the superior model for the titanic dataset. The faster execution time is the only noticeable difference between each model for this dataset and since naïve Bayes had a smaller execution time it is better suited.

**Write two paragraphs comparing generative classifiers versus discriminative classifiers.**

Both generative and discriminative classifiers can be used to build a classification model for a given set of data, but they find the probability of whether a given instance belongs to a specific class in different ways. The generative model finds the probability of P(Y|X) by estimating the prior probability of Y P(Y) and likelihood probability P(X|Y). Discriminative classifiers on the other hand they start with an assumed form of P(Y|X) and through a repetitive loop continue to optimize the form of P(Y|X) they have until it becomes more accurate [3].

Discriminative models separate the classes that they have and never make assumptions about the data. This means that the discriminative model tries to draw a line separating the datapoints until there is only class A on one side and class B on the other. The discriminative classification is more effective in datasets that have outlier data points but tends to misclassify datapoints by putting them on the wrong side of its separation [3].

Generative classifiers do not separate the datapoints by their classes instead they create new data instances by looking at the data and finding the probability estimates and the likelihood of the data and determine the class based on these probabilities. Unlike the discriminative classification model generative classification tends to be affected by outlier datapoints that can change the outcome of the entire model as it assigns equal weight to all points [3].

**Google this phrase: reproducible research in machine learning. Using 2-3 sources, at least 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.**

Reproducible research in machine learning means that the results of one person's machine learning experiment can be achieved by a completely different person if they have the necessary computing power to execute the experiment [1]. When a machine learning algorithm is performed on a given dataset the same or similar result should always be given by the algorithm [2]. Reproducible research is very important to the computer science field to validate claims and hypothesis by different computer scientists across the world. If an individual preforms a computational experiment involving machine learning and gets and excellent result, but the individual's experiment was not reproducible then the results of that experiment will not be accepted by the broader scientific community. It is only when an experiment is reproducible that it can be reliably used for a variety of different applications [1]. Additionally, having your machine learning algorithm be reproducible helps with further refining that algorithm and reducing the errors in it and improving it as a whole [2].

To implement reproducibility in machine learning research, the research should start, be conducted with, and end with reproducibility in mind [2]. Carefully document the steps the actions taken during the experiment and the reasoning behind those actions, so that when others are attempting to reproduce the experiment they have a better idea of how to do it [2]. Finally designing your machine learning code in a pipeline manner where different modules focus on the input they take in and the output they give instead of the specifics of how they get it can make the code more adaptable to different environment that the experiment can take place in making it easier to reproduce [2].

Cite your sources

- https://arxiv.org/abs/2108.12383 [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. [2]
- https://www.analyticsvidhya.com/blog/2021/07/deep-understanding-of-discriminative-and-generative-models-in-machine-learning/#:~:text=Discriminative%20models%20draw%20boundaries%20in,the%20labels%20of%20the%20data. [3]