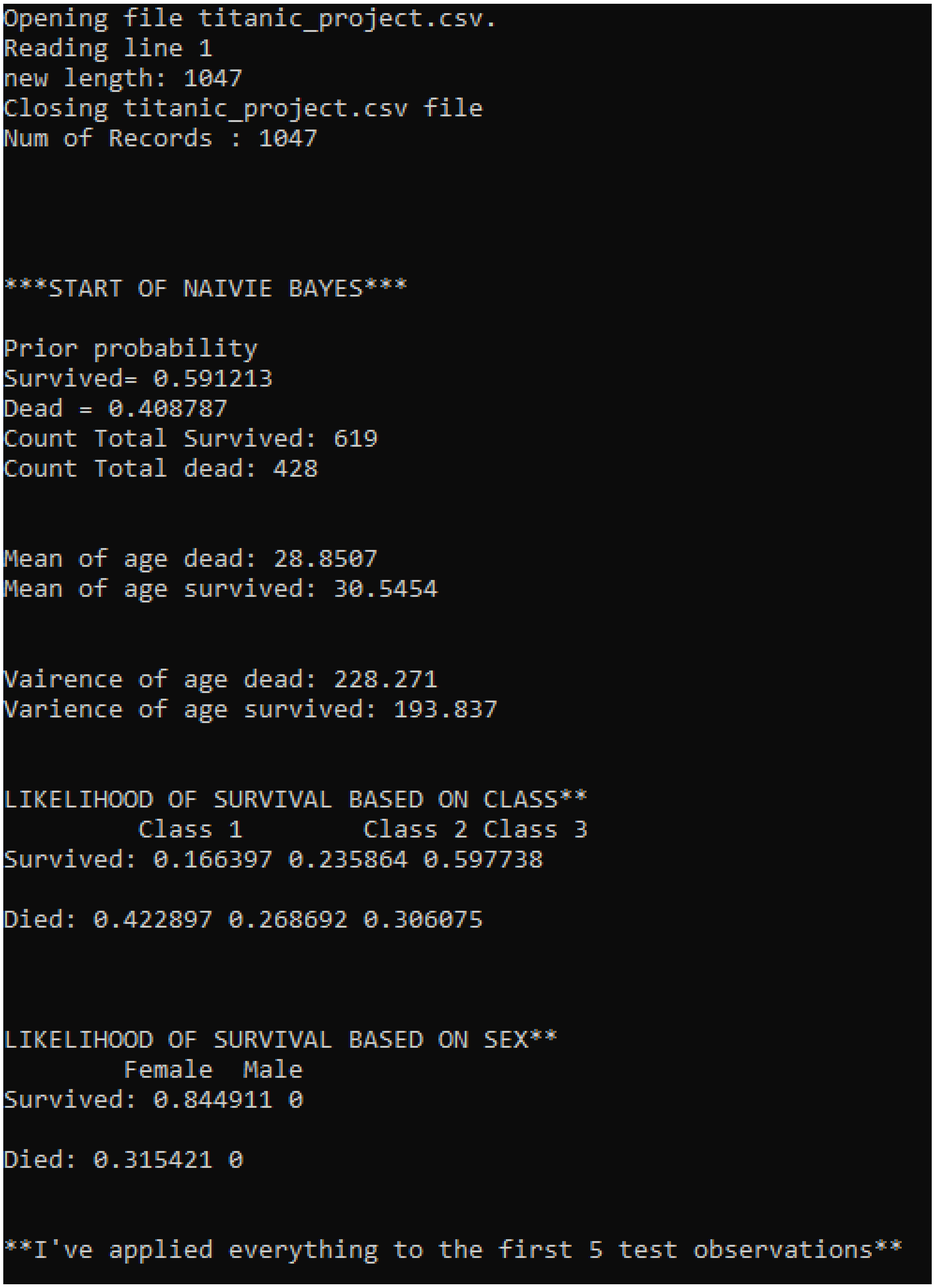
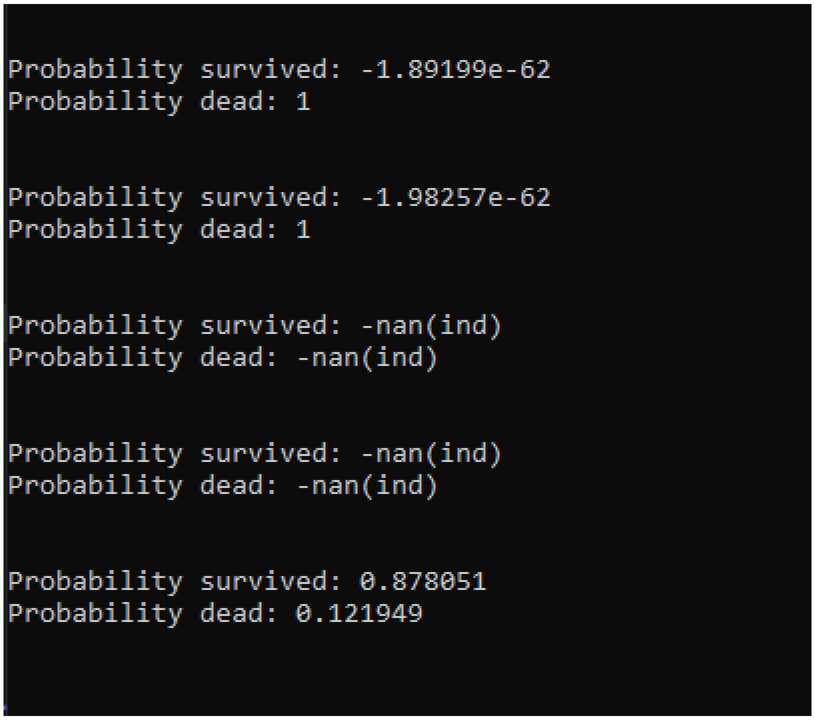
Marcela Pantoja CS 4375.004 3/4/2023

C++ ML Algorithms from Scratch

NAÏVE BAYES CODE

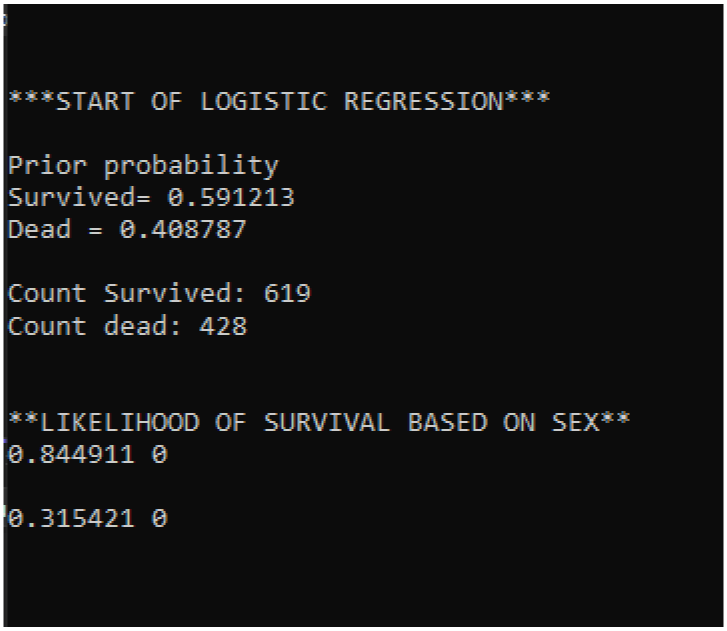


```
Opening file titanic_project.csv.
Reading line 1
new length: 1047
Closing titanic_project.csv file
Num of Records : 1047




***START OF NAIVIE BAYES***

Prior probability
Survived= 0.591213
Dead = 0.408787
Count Total Survived: 619
Count Total dead: 428


Mean of age dead: 28.8507
Mean of age survived: 30.5454


Vairence of age dead: 228.271
Varience of age survived: 193.837


LIKELIHOOD OF SURVIVAL BASED ON CLASS**
         Class 1        Class 2 Class 3
Survived: 0.166397 0.235864 0.597738

Died: 0.422897 0.268692 0.306075



LIKELIHOOD OF SURVIVAL BASED ON SEX**
        Female  Male
Survived: 0.844911 0

Died: 0.315421 0


**I've applied everything to the first 5 test observations**
```



```
Probability survived: -1.89199e-62
Probability dead: 1


Probability survived: -1.98257e-62
Probability dead: 1


Probability survived: -nan(ind)
Probability dead: -nan(ind)


Probability survived: -nan(ind)
Probability dead: -nan(ind)


Probability survived: 0.878051
Probability dead: 0.121949
```

LOGISTIC REGRESSION CODE



```
***START OF LOGISTIC REGRESSION***

Prior probability
Survived= 0.591213
Dead = 0.408787

Count Survived: 619
Count dead: 428


**LIKELIHOOD OF SURVIVAL BASED ON SEX**
0.844911 0

0.315421 0
```

Analyze the results of your algorithms on the Titanic data

The Naïve Bayes Data model modeled the prior probabilities and the likelihood of survival based on Age, Class, and Sex before calculating the raw probabilities for each factor. This information lets us know that overall, people were more likely to survive than die, people of a high class were more likely to survive, and female were likelier to survive than men.

When calculating the raw probabilities for the first 5 rows of data, it found that the first 2 people were likely to have died, the 3rd and 4th gave inconclusive results, and the 5th person was likelier to survive.

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

Generative classifiers work by learning the joint probability P(x,y) and transform that by using Bayes rule into a conditional probability (P(y|x)). Discriminative classifiers on the other hand, will directly learn the conditional probability (P(x|y)) with no needs for Bayes rule. Naive Bayes in an example of a generative classifier, while Logistic Regression is an example of a discriminative classifier.

These two models have may differences, such as generative models being more sensitive to outliers and discriminative models having outliers as a misclassified example. Naive Bayes also has a tendency to perform better on small data sets, with Logistic regression outpacing and overtaking it as the data size grows. Naïve Bayes has a higher bias, but a lower variance than logistic regression, meaning that the Baye model makes more assumptions than Logistic Regressions and that the sample data is close to where the model predicted.

Reproducible Research in Machine Learning

Reproducibility in science (ML included) is the ability for something to be recreated or copied in the future in the same domain. Reproducible research in Machine Learning is ML work/research that is able to be recreated in the future.

Reproducible research is incredibly important as “an algorithm from a new research without the reproducibility aspects can be difficult to investigate and implement.[2]” To investigate a model is to be able to look into it and fully understand how it works. Machine Learning models are widely used, and it is dangerous to integrate any machine learning model that is not fully understood, and to be fully understood the model must be reproducible. And, as stated before, a model without reproducibility aspects can be difficult to implement, so the models uses become restricted. Reproducibility can also “help your team reduce errors and ambiguity…[and] ensure data consistency [4]” in a project.

Some people have theorized standards for Machine Learning models as a way to ensure reproducibility, as "For machine-learning models… to become trusted, scientists must prioritize computational reproducibility [3]". These standards can be anything from making sure that you research is publicly available (so that it can be verified), to automating the model so that it can be re-run thoroughly with ease. But, reproducibility can also be implemented through ensuring that documentation is being done from the very beginning of the project to the end. Documentation is one of the most crucial aspects of reproducibility because it allows for you to understand how the project was developed and why choices were made.

Citations

[1] Yıldırım, Soner. "Generative vs Discriminative Classifiers in Machine Learning." *Medium*, Towards Data Science, 14 Nov. 2020, https://towardsdatascience.com/generative-vs-discriminative-classifiers-in-machine-learning-9ee265be859e.

[2] Hemant, Preeti. "Reproducible Machine Learning." *Medium*, Towards Data Science, 7 Apr. 2020, https://towardsdatascience.com/reproducible-machine-learning-cf1841606805.

[3] Heil, Benjamin J., et al. "Reproducibility Standards for Machine Learning in the Life Sciences." *Nature Methods*, vol. 18, no. 10, 2021, pp. 1132–1135., https://doi.org/10.1038/s41592-021-01256-7.

[4] "The Importance of Reproducibility in Machine Learning Applications." *DecisivEdge*, 7 Dec. 2022, 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.