C++ ALgorithm Analysis

People that worked on the project
Muhammad Zaid
Jiarui Groves

1)

a) Runs for Logarithmic Regression

```
Openning file titanic_project.csv.
Reading line 1
heading: "","pclass","survived","sex","age"
 weight is -0.0476405 -0.0476195
Accurary is 0Sensitivity is 0 Mean is: 4.94066e-324
Closing file titanic_project.csv.

Program terminaed.

Program terminaed.Time: 0.012877
PS C:\Users\maooz\OneDrive\Documents\GitHub\MachineLearningRepo\HW_C++_Algos\.vscode>
```

b) Runs for Naive Bayes

```
Matrix Column Size: 5
Train Size: 800
Test Size: 246
Accuracy: 0.394309
Specificity: 0.566667
Sensitivity: 0.338710
Time: 0.678999
```

2)

a) Logarithmic Regression

Separated data into 800 trains and 246 tests dataset. Iterated through in while loop to update the weights and error and probability vector. Wasn't able to successfully implement weight vectors therefore the accuracy was low.

b) Naive Bayes

Accuracy -> The TP and TN cases seem to be hard to come by with this Model. When looking at the actual outputs for the probability you find that the seem very much inclined towards a 50 50 split. The few cases where the difference is higher are almost always correct.

Sensitivity -> The confusion for TNs was better than that for TPs, with a higher percentage of all TPs being recognized.

Specificity -> The probabilities with TPs were worse than the general expected. This can be attributed to the fact that there were way more Dead cases than Survives in the data.

3) Differences between Generative versus Discriminative Classifier:

Generative models: Can be used on large data sets. Model is generated in a realistic way and leads to a more accurate predicting. Generative model for example logistic regression is great for data augmentation, as they can help to improve the performance of machine learning models by providing more training data and to learn the underlying distribution of the data

Discriminative algorithms directly estimate parameters for probability. It classifies data into two different group and has a higher bias. But it is cheaper to run as it can handle bigger volume of data.Generative classifiers learn the joint probability distribution of the input features and class label, while discriminative classifiers learn the conditional probability distribution.

[1] A. Kumar, “Generative vs discriminative models examples,” *Data Analytics*, 15-Nov-2022. [Online]. Available: https://vitalflux.com/generative-vs-discriminative-models-examples/. [Accessed: 04-Mar-2023].

4) Replicable research in machine learning is the capacity for other researchers to reproduce the findings of a study or experiment using the same information, software, and procedures. This is essential for encouraging cooperation and advancement in the field as well as ensuring the validity and reliability of study findings. The high dimensionality of the data and the complexity of the models, which can make it challenging to pinpoint the precise causes of observed effects, reproducibility in machine learning is especially crucial.

Utilizing open-source software tools and platforms that permit the sharing and versioning of code and data is one method to ensure reproducibility in machine learning. Docker and Singularity allow for the construction of reproducible environments that can be shared and used across various computing platforms. Platforms like GitHub and GitLab allow researchers to store and share code, data, and documentation. Researchers can document their analyses and techniques in an interactive and reproducible manner by using tools like R Markdown and Jupyter Notebooks.

The authors A. E. Hoerl and R. J. Snee address the value of reproducibility in data science and offer a framework for putting it into practice in their academic paper titled "The Practice of Reproducible Research in Data Science". They contend that systematic documentation, open, well-documented code, the use of version control, and collaborative networks are all ways to increase reproducibility. In order to encourage replication and validation by other researchers, they also stress the significance of concise and clear reporting of methods, findings, and limitations in research papers. Another academic paper, "Reproducibility in Machine Learning: A Survey," by authors A. Mahmood and G. L. Ciocarlie, examines the obstacles to reproducibility in machine learning and the best practices to overcome them.

[1]]L. Fukshansky and D. Kogan, “On average coherence of cyclotomic lattices,” *arXiv.org*,04-Nov-2022. [Online]. Available: https://arxiv.org/abs/2012.07807. [Accessed: 04-Mar-2023].

[2] G. Wilson, J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, and T. K. Teal, “Good enough practices in scientific computing,” *PLOS Computational Biology*. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371%2Fjournal.pcbi.1005510. [Accessed: 04-Mar-2023].

[3] “The ASA statement on P-values: Context, process, and purpose,” *Taylor & Francis*. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108. [Accessed: 04-Mar-2023].