Part A
copy/paste runs of your code showing the output (coefficients and metrics), and run times

Run of logistic regression from scratch in C++

```
sujayvadlakonda@Sujays-MBP 03-ml-algorithms-from-scratch % g++ logistic-regression.cpp -o logistic-regression.out
sujayvadlakonda@Sujays-MBP 03-ml-algorithms-from-scratch % ./logistic-regression.out
Opened file titanic_project.csv
Model Training Time: 95418 milliseconds
Slope: -2.41086
Intercept: 0.999877
Accuracy: 0.784553
Sensitivity: 0.816327
Specificity: 0.763514
sujayvadlakonda@Sujays-MBP 03-ml-algorithms-from-scratch %
```

Run of naive bayes from scratch in C++

```
sujayvadlakonda@Sujays-MBP 03-ml-algorithms-from-scratch % g++ naive-bayes.cpp -o naive-bayes.out
sujayvadlakonda@Sujays-MBP 03-ml-algorithms-from-scratch % ./naive-bayes.out
Opened file titanic_project.csv
Survived Prior: 0.39
Not Survived Prior: 0.61
Passenger Class 1 Given Survived: 0.416667
Passenger Class 2 Given Survived: 0.262821
Passenger Class 3 Given Survived: 0.320513
Passenger Class 1 Given Not Survived: 0.172131
Passenger Class 2 Given Not Survived: 0.22541
Passenger Class 3 Given Not Survived: 0.602459
Male Given Survived: 0.320513
Female Given Survived: 0.679487
Male Given Not Survived: 0.840164
Female Given Not Survived: 0.159836
Age Mean Given Survived: 28.8261
Age Variance Given Survived: 209.155
Age Mean Given Not Survived: 30.4182
Age Variance Given Not Survived: 205.153
Training Time: 128 microseconds
Accuracy: 0.784553
Sensitivity: 0.816327
Specificity: 0.763514
sujayvadlakonda@Sujays-MBP 03-ml-algorithms-from-scratch %
```

Part B
analyze the results of your algorithms on the Titanic data

Both the logistic regression and the naive bayes algorithms produce the same accuracy, sensitivity, and specificity. This seems to suggest that sex is the only relevant predictor of survival, because sex was the only predictor in the logistic regression. The naive bayes age priors seem to indicate that there is not much of a difference between the ages of the people who survived and the people that did not. The naive bayes passenger class priors show that the passenger class has a significant effect on survival but not enough to change the classification based on sex. The slope of the logistic regression indicates that being male decreases the chances of survival.

Part C

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

Generative and discriminative classifiers are two different ways of doing machine learning classification. Generative classifiers examine the underlying distributions between the predictors and the target in the training data [1]. For example, naive bayes finds the conditional distributions between qualitative predictors and the target and uses gaussian distributions between quantitative predictors and the target. Discriminative classifiers try to determine the most effective way to separate data points into groups [1]. For example, logistic regression uses a line that separates data points and uses the improvement of log odds to find the most effective line.

Generative classifiers are worse at dealing with outliers because their distributions become greatly affected by them [1]. Discriminative classifiers are not as affected by outliers because they simply put outliers on the wrong side of the boundary anyways because that minimizes error [1]. Generative classifiers require more data and computation to be effective when compared to discriminative classifiers [1]. This can be seen in my results as the logistic regression took significantly more time to train than the naive bayes model.

Part D

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. Cite your sources using any format.

Reproducible research in machine learning is creating research that can be replicated by other machine learning researchers to verify the results of the first machine learning researcher [2]. Reproducibility is important because it ensures that the machine learning model is reliable, before the machine learning model is given high-risk tasks that cannot afford to see failure [3].

Reproducibility can be implemented by having strict standards for documentation from the beginning of a machine learning project [4]. Other key ways reproducibility can be achieved in machine learning is to keep track of the specific training and testing data sets used [2]. Code that is used to generate features can be put under version control to gain an understanding of how exactly the features were generated [2]. The code environments and the specific hardware that was used can also be recorded to ensure reproducibility [2].

Citations

[1] Yildirim, 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] McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: Still a ways to go. Sci Transl Med. 2021 Mar 24;13(586):eabb1655. doi: 10.1126/scitranslmed.abb1655. PMID: 33762434.

[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/.