a. Output of logreg.cpp:

```
Opening file titanic_project.csv
Reading line 1
heading: "","pclass","survived","sex","age"

Training took 0.0312993s

             Coefficient
(intercept)      0.999588
"sex"            -2.41045

accuracy         0.784553
sensitivity      0.695652
specificity      0.862595
```

Output of naivebayes.cpp:

```
Opening file titanic_project.csv
Reading line 1
heading: "","pclass","survived","sex","age"

Training took 0.0001792s

A-priori Probabilities:
       0        1
    0.61     0.39

Conditional Probabiliites:
        "pclass"
                0        1        2
        0 0.172131  0.22541 0.602459
        1 0.416667 0.262821 0.320513

        "sex"
                0        1
        0 0.159836 0.840164
        1 0.679487 0.320513

Means and Variances
        "age"
              Mean Variance
        0  30.3914   204.73
        1  28.8077  209.316

accuracy         0.784553
sensitivity      0.695652
specificity      0.862595
```

b. Both algorithms achieve similar results to one another, with both models producing the same values for accuracy, sensitivity, and specificity. Logistic regression takes many times longer to produce a model compared to naive Bayes. This makes sense, as logistic regression is an iterative algorithm while naïve Bayes requires only a deterministic number of passes over the

data to produce a result. Even if we were to add more predictors to this dataset, naïve bayes would continue to be quite fast while logistic regression may start slowing down.

c. Generative classifiers capture the joint probability distribution of the predictors with the target. Generative models estimate a prior probability P(Y) and the likelihood P(X|Y), and to make predictions, P(Y|X) is calculated via Bayes Theorem. Generative models can use this information to generate new data points [1]. Naïve Bayes is an example of a generative model.

   Discriminative classifiers do not model the conditional probabilities of the data and instead model the boundaries between classes. Unlike generative classifiers, discriminative classifiers directly estimate the value of P(Y|X). As such, discriminative classifiers cannot be used to generate new data points [1]. Logistic regression is an example of a discriminative classifier.

d. Reproducibility is the ability of an independent researcher to duplicate the results of a scientific study using the same materials and methods as the original researchers [2]. In the case of computational experiments, a study is considered reproducible if another researcher can repeat the experiment with the same data, methods, and computing power and achieve the same results [3]. Reproducibility is important because it demonstrates that a study is credible and that its results are correct. If a study is not reproducible, it indicates that the original researchers have made some sort of mistake. Reproducibility is especially important in machine learning because it is often the case that researchers cannot understand precisely why or how a model works. Furthermore, anything from randomization of starting weights to changes in software versions may make reproducing a study more difficult [2].

   Reproducibility can be achieved by documenting every element that can contribute to the creation and performance of a machine learning model. The precise algorithm must be clearly described. The exact source code and data that were used by the original researchers must be openly available. The process of data collection and partitioning must be recorded. The values of hyperparameters must be given. Finally, a complete description of the computing infrastructure used to create the model must also be provided [2]. Because a change in any of these factors can contribute to a difference in the resulting model, they must all be precisely recorded to give other researchers the best possible chance of reproducing results.

References

[1] C. Goyal, “Deep understanding of discriminative and generative models,” Analytics Vidhya, 19-Jul-2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/07/deep-understanding-of-discriminative-and-generative-models-in-machine-learning/. [Accessed: 01-Oct-2022].

[2] Z. Ding, A. Reddy, and A. Joshi, “5 - reproducibility,” Machine Learning Blog | ML@CMU | Carnegie Mellon University, 24-Aug-2020. [Online]. Available: https://blog.ml.cmu.edu/2020/08/31/5-reproducibility/. [Accessed: 01-Oct-2022].

[3] J. Shenouda and W. U. Bajwa, “A guide to computational reproducibility in signal processing and machine learning,” arXiv.org, 15-Feb-2022. [Online]. Available: https://arxiv.org/abs/2108.12383. [Accessed: 01-Oct-2022].