Report on Assignment 3

Part A:

```
Training time: 0.418436

Coefficients [intercepts, sex]: [0.999869, -2.41085]

Accuracy: 0.784553
Sensitivity: 0.695652
Specificity: 0.862595
```

Logistic Regression Output

```
A-priori probabilities:
0         1
0.61   0.39

Conditional probabilities:
Pclass
  1        2        3
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

Age
  Mean         Variance
0 30.4182      205.153
1 28.8261      209.155

Training time: 0.000188887
Accuracy: 0.764228
Sensitivity: 0.521739
Specificity: 0.977099
```

Naive Bayes Output

Part B:

For this assignment, I implemented Logistic Regression and Naive Bayes in C++. These algorithms analyzed the Titanic data and predicted passenger's survivability. Regarding predictors, Logistic Regression selected gender, while Naive Bayes selected passenger class, gender, and age. In the context of metrics, these algorithms offered similar performance. Logistic Regression and Naive Bayes have respective accuracies of ~0.78 and ~0.76. In this case, Naive Bayes possessed some disadvantages. Both algorithms analyzed a medium-sized dataset of 800 observations. Naive Bayes tend to underperform on larger datasets. It only offered similar accuracy due to the usage of multiple predictors. In the context of training time, Naive Bayes quickly analyzed the data. It assumed conditional independence and calculated conditional probabilities. In contrast, Logistic Regression performed gradient descent. Through multiple iterations, it modified the weights in an incremental manner.

Part C:

Discriminative classifiers are used for supervised machine learning. They estimate parameters of P(Y|X). With a given training set, discriminative classifiers determine some boundaries between classes. They model the data and determine some function that maps inputs to class outputs. Discriminative classifiers are robust when handling outliers. However, they can suffer from misclassification.

Generative classifiers are used for unsupervised machine learning. They estimate parameters for P(Y) and P(X|Y). They learn the data distribution and use Bayes theorem to calculate joint probabilities. They use likelihood to separate class labels. Generative classifiers do not handle outliers well.

Source:
https://www.analyticsvidhya.com/blog/2021/07/deep-understanding-of-discriminative-and-generative-models-in-machine-learning/

Part D:

Reproducible ML research refers to the replication of studies by different individuals. Specifically, individuals can reproduce ML work and generate the same findings. Reproducibility helps enforce transparency and reliability. With regards to ML, researchers can compare baseline models and determine whether novel work yielded improvement. By also verifying proof of correctness, researchers can check for consistent results.

A lack of reproducibility poses various repercussions. This can be demonstrated by some interdisciplinary studies. These studies leverage existing ML techniques and apply them to different fields. However, some of these studies don't follow proper protocols. This leads to the generation of spurious results. By analyzing their methodology, researchers can identify different flaws or errors. As a result, some ML models might yield dubious findings.

Reproducible ML research can be implemented with various means. Researchers should provide a thorough and clear overview of the algorithm and complexity analysis. They should also provide links, so that other individuals can access the study's source code and dataset. Researchers can also provide a description of their computing tools and software. With these methods, researchers can help enforce ML reproducibility.

Sources:
https://blog.ml.cmu.edu/2020/08/31/5-reproducibility/
https://www.wired.com/story/machine-learning-reproducibility-crisis/