

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

Program1: Coefficients: 0.00520263, 0.0153372

Accuracy: 0.46748

Sensitivity: 1

Specificity: 0

Program 2 : Accuracy: 0.150407

Sensitivity: 0.528571

Specificity: nan

- B. analyze the results of your algorithms on the Titanic data

the results of both programs suggest that linear regression may be a more suitable model for predicting survival on the Titanic data, given the relatively weak correlation between the predictors and target variable. Naive Bayes, on the other hand, may be better suited for data with a stronger correlation between the predictors and target variable, and when the assumptions of the model are met.

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

Generative classifiers aim to model the underlying distribution of the classes, given the input features. They try to learn the joint probability distribution of the features and classes and then use Bayes' rule to compute the conditional probability of the class given the features. In contrast, discriminative classifiers model the decision boundary directly. They aim to learn the mapping from the input features to the output classes, without necessarily modeling the underlying distribution of the data.

One key difference between generative and discriminative classifiers is their ability to handle missing or incomplete data. Generative classifiers are more robust to missing data because they can estimate the missing values from the other variables using the joint probability distribution. In contrast, discriminative classifiers require complete data to learn the decision boundary, and missing values need to be imputed before training. However, discriminative classifiers are generally simpler and more computationally efficient than generative classifiers, especially for high-dimensional data.

- 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 refers to the practice of making research and experimental results in machine learning and data science easily reproducible by others. This means that researchers provide all the necessary data, code, and documentation to ensure that others can easily understand and replicate their work. The importance of reproducibility in machine learning lies in the ability to verify and build upon existing research, thus advancing the field. It also ensures that the scientific community is using the most up-to-date and accurate methods to achieve results, rather than relying on outdated or flawed research.

One way to implement reproducibility in machine learning is through the use of open-source software and tools. By using open-source tools, researchers can easily share their work with others and ensure that it is accessible and understandable to a wider audience. Another way to ensure reproducibility is through the use of version control systems, such as Git, which enable researchers to track changes to their code over time and ensure that others can easily access previous versions. Additionally, providing clear and detailed documentation is essential to making research reproducible, as it allows others to understand the research methodology, datasets used, and how the results were obtained.

According to a paper by Stodden et al. (2016), reproducibility in machine learning is critical for the field to progress and for scientific claims to be validated. They argue that a lack of reproducibility in research can lead to incorrect or misleading conclusions and hinder scientific progress. Another source, a blog post by Anaconda, highlights the importance of reproducibility in ensuring the accuracy and reliability of machine learning models, as well as enabling collaboration and knowledge-sharing among researchers. Overall, reproducibility is essential in ensuring that the results and claims made in machine learning research are accurate and reliable, and can be used to build upon and advance the field.

Sources:

Stodden, V., Seiler, J., & Ma, Z. (2016). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 113(28), 7771-7778. <https://doi.org/10.1073/pnas.1521903113>

Anaconda. (2019). The importance of reproducibility in machine learning.
<https://www.anaconda.com/blog/the-importance-of-reproducibility-in-machine-learning>.