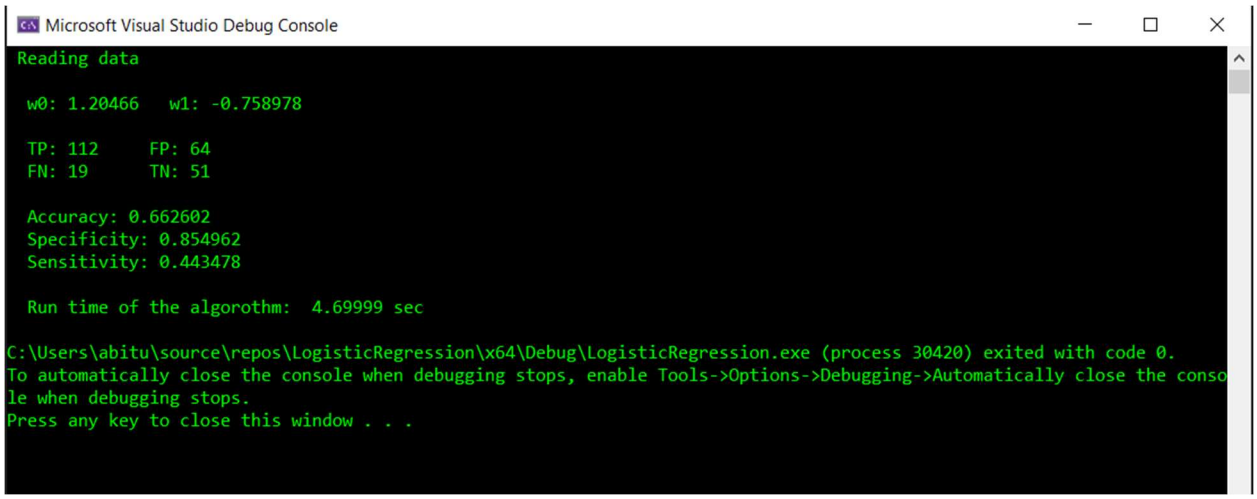
## Report on Portfolio: ML Algorithms from scratch

## Logistic regression run time result




## Naïve Bayes run time result

**Analyzing the results of algorithms:**

The logistic regression model, a true positive value is 112, meaning the model predicts that 112 have survived and they really survived. The true negative value is 51, that is 51 people died, and they actually died. The model fails to predict correctly that 64 people survived and 19 died, which are the values of the false positive and false negative values respectively. Sensitivity, the proportion of survived, who were correctly identified as survivors, is 44 % , Similarly specificity, those who died and were predicted so is 85 %. Whereas Accuracy, that measure the proportion of survived and died, and were predicted so is 66%. Run time of the algorithm is 4.7 sec.

The Naïve Bayes model, true positive value is 113, and the true negative value is 80, false positive value and false negative values are 35 and 18 respectively. The classification test results, Sensitivity, specificity, and accuracy are 70% , 86%, and 78% respectively. It looks much better that logistic regression, the run time for this model is way faster, it's 0.0005 sec.

**Generative vs Discriminative classifier**

Classification is used to create a model based on evidence, E and hypothesis, H that can predict the class H given a set of new, unseen attributes E. Generative and discriminative classifiers work the classification differently. Generative classifiers learn a model of the joint probability, p( E, H), of the inputs x and the label y, and make their predictions by using Bayes rules to calculate P(E|H ), and then picking the most likely label H, it first tries to learn the model that generates the data behind the scenes by estimating the distributions of the model and then use it to predict unseen data, whereas, discriminative classifier tries to model by just depending on the observed data. It makes fewer assumptions on the distributions but depends heavily on the quality of the data, estimate P(H|E) directly. A decision boundary is created that creates a dividing line between instances of one class and instances of another class. Unseen instances are then classified based on which side of the line they fall. In this way, a direct mapping is generated from attributes E to class labels H, example is the Logistic Regression.

It is good to know when to use generative or discriminative classifier, when we have a big data, discriminative model performs better than generative model. As Andrew Ng concluded, probabilistic models can be fit either to optimize the joint likelihood of the evidence and the hypothesis or fit to optimize the conditional likelihood. The generative model does indeed have a higher asymptotic error (as the number of training examples becomes large) than the discriminative model, but the generative model may also approach its asymptotic error much faster than the discriminative model- possibly with several training examples that is only logarithmic, rather than linear, in the number of parameters.

**Reproducible research in machine learning**

Reproducibility is the ability to be recreated or copied. In machine learning context, reproducibility is being able to recreate a machine learning workflow to reach the same conclusions as the original work, , it relates to getting the same output on the same algorithm, parameters, and data on every run. Machine learning (ML) aims at solving problems that cannot be solved by applying already known logic, rather it is much of its *intelligence* derived from data upon application of complex mathematics, so a tiny change can bring a change in the outcome, and the whole work of getting a solution fails. We also fail to have a same result at every run when we don't set the seed for the random shuffle. An algorithm from new research without the reproducibility aspects can be difficult to investigate and implement. Reproducibility is also an important step to promote open and accessible research

A reproducible version of a model has much importance, to mention some, Reproducibility helps with understanding, explaining, and debugging, and is also a crucial means to reverse engineering, since ML is inherently difficult to understand and debug, obtaining different results at every run could have worsen it. Reproducibility helps to achieve correctness through understanding and debugging, and ultimately it helps the ML to be credible. Reproducibility in preceding layers is needed to build out and extend. , it's more important that reproducibility to be ensured, as ML applications are being used to generate training data now, the world's most valuable resource is no longer oil, but data! Implementing Reproducible Research covers many of the elements necessary for conducting and distributing reproducible research. It explains how to accurately reproduce a scientific result, Stodden in his '*Implementing Reproducible Research*' book in computational science is divided into three parts, the tools, practices, and dissemination platforms. Computational tools, such as Sweave, knitr, VisTrails, Sumatra, CDE, and the Declaratron system open-source practices, good programming practices, trends in open science, and the role of cloud computing in reproducible research Software and methodological platforms, including open-source software packages, RunMyCode platform, and open access journals Each part presents contributions from leaders who have developed software and other products that have advanced the field.

Works Cited

Y. Ng, A., and Jordan, M. *On Discriminative vs. Generative Classifiers: A Comparison of ...* - *Neurips*. 2001, https://proceedings.neurips.cc/paper/2001/file/7b7a53e239400a13bd6be6c91c4f6c4e-Paper.pdf.

"Reproducibility in Machine Learning - Research and Industry." *Suneeta Mall*, 21 Dec. 2019, https://suneeta-mall.github.io/2019/12/21/Reproducible-ml-research-n-industry.html.

Stodden, Victoria, Friedrich Leisch, and Roger D. Peng. Implementing Reproducible Research. Ed. Victoria Stodden, Friedrich Leisch, and Roger D. Peng. First edition. Boca Raton, FL: Chapman and Hall/CRC, 2018. Print.

Lucic, Ana et al. "Towards Reproducible Machine Learning Research in Natural Language Processing." ACL (2022).