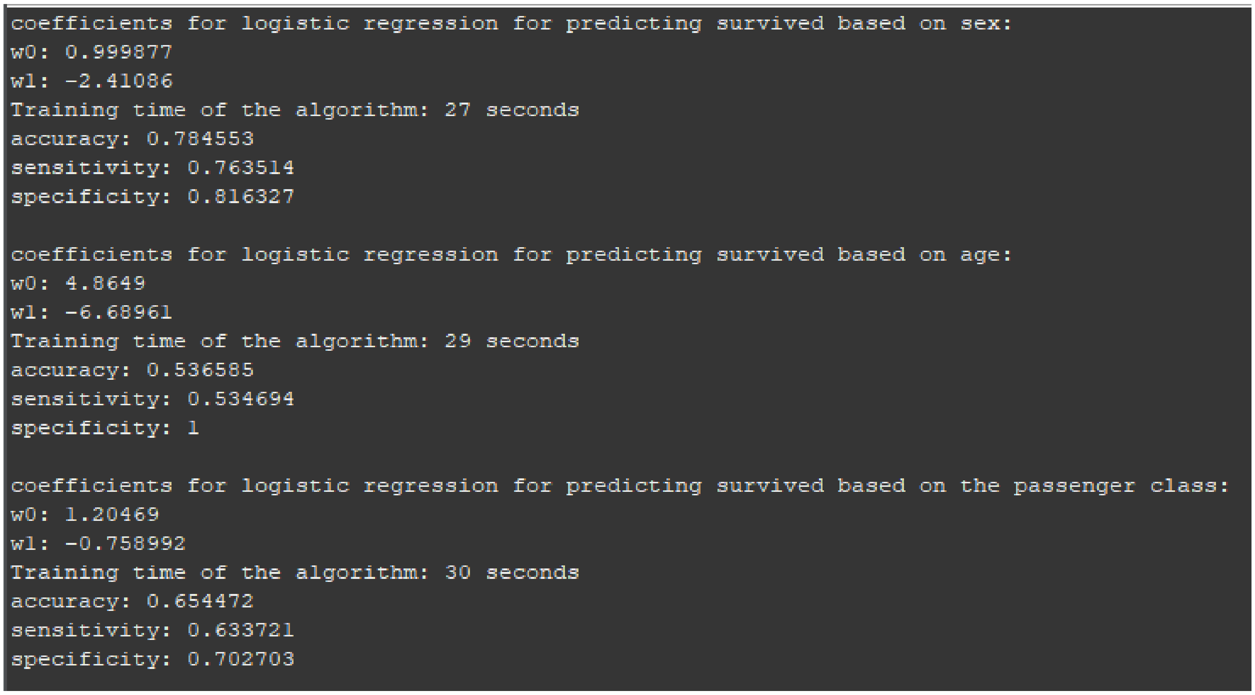
By: Dinesh Angadipeta(DXA190032),
Alejo Vinluan (ABV210001)

- Logistic Regression Code Output:



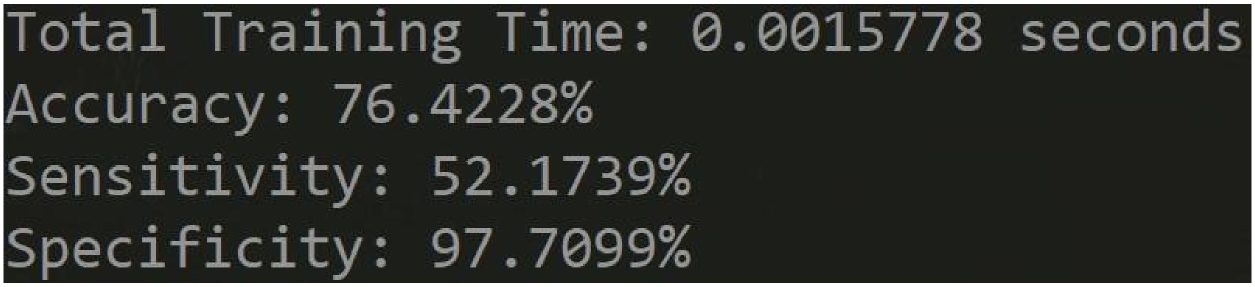
```
coefficients for logistic regression for predicting survived based on sex:
w0: 0.999877
w1: -2.41086
Training time of the algorithm: 27 seconds
accuracy: 0.784553
sensitivity: 0.763514
specificity: 0.816327

coefficients for logistic regression for predicting survived based on age:
w0: 4.8649
w1: -6.68961
Training time of the algorithm: 29 seconds
accuracy: 0.536585
sensitivity: 0.534694
specificity: 1

coefficients for logistic regression for predicting survived based on the passenger class:
w0: 1.20469
w1: -0.758992
Training time of the algorithm: 30 seconds
accuracy: 0.654472
sensitivity: 0.633721
specificity: 0.702703
```

Analysis: We can see that when using logistic regression in c++ to predict values compared to using R, the time needed to calculate the coefficients and metrics is drastically more extensive and more expensive. You can see that the training time for each instance of the algorithm ranges in the high 20 seconds, which is a very long time for a c++ program. It is due to the fact that the program is iterating around 400 million times for each training algorithm to ensure accurate results. We can see that the log odds of predicting survived using these data points all have a negative correlation. The accuracy for using sex and passenger class is generally high while the accuracy for using age is around 53 percent, which is not the most optimal. All in all, it can be seen that using sex as a predictor may yield more accurate results.

- Naive Bayes Code Output



```
Total Training Time: 0.0015778 seconds
Accuracy: 76.4228%
Sensitivity: 52.1739%
Specificity: 97.7099%
```

Naive Bayes was utilized to predict survival on the dataset utilizing 3 features: the passenger's class, sex, and age. When utilizing the 4 features, the code was able to correctly predict the survival of a passenger with an accuracy of 76%. When taking a look at the 52% sensitivity rate, we can see that the algorithm does not perform well when predicting passengers who survived. However, specificity had a high rate of percentage at 97.7%. This suggests that the Naive Bayes Classification was able to accurately predict when passengers would perish.

- Generative Classifiers vs. Discriminative Classifiers:
  One of the main differences between generative classifiers and discriminative classifiers is that generative classifiers are used to learn the joint probability between a feature variable and a target variable while discriminative classifiers are used to learn the boundaries that separate different classes. These classifiers are used to find the conditional probability instead of the joint probability.

  Both of these classifiers have their own pros and cons. For example, generative models need a vast amount of data to be able to represent distributions accurately. These models are also much more expensive to use. Discriminative models on the other hand are affected by outliers a lot more. Despite their differences, both of these models are very useful in machine learning, in finding patterns and characteristics.

  Source:

  Yıldırım, 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.

- Reproducible Research in Machine Learning

  Reproducible Research in Machine Learning is the idea that calculations created and research that is published should be transparent and reproducible so that it could be credible (LeVeque 13). At the moment, a significant amount of research publishes only results without the sequence to reproduce the result. This prevents researchers from the same field from reproducing the results and verifying the accuracy of the results.

  The process of how a result was generated within Machine Learning is important since it allows other researchers to reproduce the results and develop upon the process. It is also important for education. A student can learn how to implement the process and generate similar results (Lucic).

  Reproducibility can be implemented by having multiple levels in reproducibility. At the root level, both the data source code be open and available for researchers to look at and evaluate. Above that level, recommendations are given for how to utilize the AI, and which data can be evaluated. Finally, at the top level, the publication should only suggest that the data provided is only theoretical so as to not mislead other researchers (Gundersen).

Sources
LeVeque, R. J., Mitchell, I. M., & Stodden, V. (2012). Reproducible research for scientific computing: Tools and strategies for changing the culture. *Computing in Science & Engineering*, *14*(04), 13-17.

Lucic, A., Bleeker, M., de Rijke, M., Sinha, K., Jullien, S., & Stojnic, R. (2022, July). Towards Reproducible Machine Learning Research in Information Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 3459-3461).

Gundersen, O. E., Gil, Y., & Aha, D. W. (2018). On reproducible AI: Towards reproducible research, open science, and digital scholarship in AI publications. *AI magazine*, *39*(3), 56-68.