Comparison of Naïve Bayes and Logistic Regression on predicting individuals with heart attacks

Arwa Guudle [redacted]

Description and motivation

The goal of this study is to compare the two methods. We will be comparing the performance of the models and to evaluate which model is better in terms of predicting patients with heart disease.

Initial Analysis of the data set

- Dataset: Heart Disease data, from UCI Machine Learning Repository.
- The dataset contains of the patients from Cleveland, Hungary, Switzerland, and the VA Long Beach information such as age, sex, cholesterol etc. as well as whether they have a heart disease.
- We will be focusing solely on the patients in Cleveland
- Despite this data being "processed data", there is some missing data. To which, I had them removed.
- The data set contains 303 rows, each reflecting the patient's information, 14 columns including one target feature
- The target feature contains a 0 if there's no induction that the patient has heart disease, and contains 1,2,3,4 to distinguish levels of heart disease (Eg: 1 for mild, 2 for moderate, 3 for severe, and 4 for very severe). However, with some manipulation, the target in this data will show 1 for any case of heart disease.
- The first table on the right shows the statistical information (mean, standard deviation, minimum and maximum values) of the selected features
- The correlation heatmap shows the relationship between each feature, although there isn't much of a correlation between most of the features.
- Lastly, we see multiple histograms of the features for all patients but also with the patients that have heart disease.

Features	Mean	Std	Min	Max
Age	54.542	9.0497	29	77
Cholesterol	247.35	51.998	126	564
Resting Blood Pressure	131.69	17.763	94	200
Max Heart Rate	149.6	22.942	71	202

Figure 1: Statistical Value of Predictors

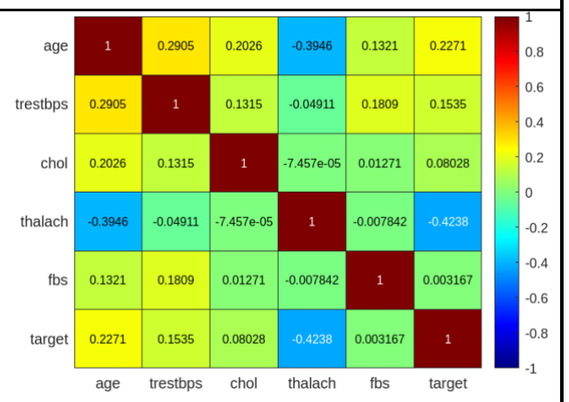
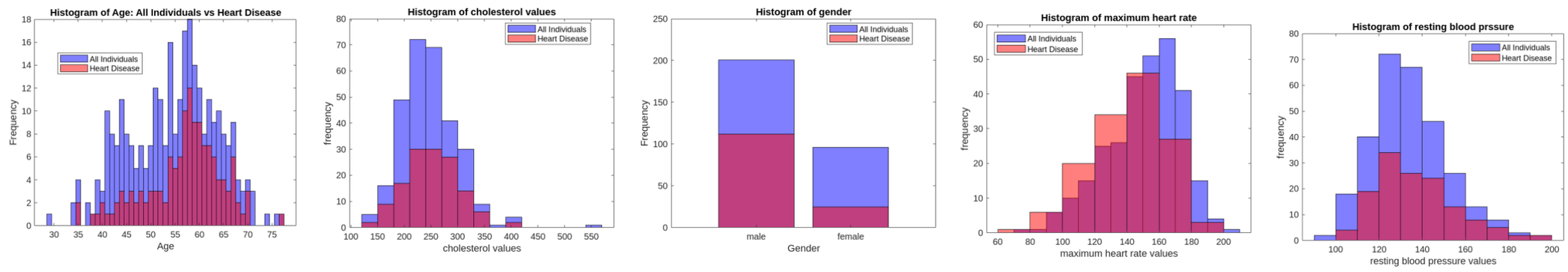


Figure 2: Correlation Heatmap

Figure 3: Histograms



The Models

Logistic Regression:

• Logistic regression is a statistical method used in machine learning for building models where the dependent variable is binary, represented by 0 and 1 (Simplilearn.com, n.d.).
• It is employed to describe data and the relationship between a dependent variable and one or more independent variables (Geeks for Geeks, 2024).

Pros:

• Simple (Simplilearn.com, n.d.)
• Performs well when the relationship between independent variables and the log-odds of the dependent variable is approximately linear (Geeks for Geeks, 2024).
• Less prone to overfitting (Kumar Bhowmik, 2015).

Cons:

• Not as effective for multi-class classification, as it is primarily designed for binary classification (Simplilearn.com, n.d.).
• Sensitive to outliers, which can skew results (Geeks for Geeks, 2024).

Naïve Bayes:

• Naïve Bayes is a classification algorithm based on Bayes' Theorem, which calculates the probability of a feature vector being associated with a specific label (Learning Classifiers based on Bayes Rule, 2015).
• It assumes conditional independence for each feature, which means that the algorithm expects the features to be independent—a condition that may not always be met in real-world data (Geeks for Geeks, 2024).

Pros:

• Effective for high-dimensional data (Geeks for Geeks, 2024).
• Works well with categorical data (Kumar Bhowmik, 2015).
• Handles missing data effectively (Learning Classifiers based on Bayes Rule, 2015).

Cons:

• Assumes conditional independence among features, which can lead to inaccurate predictions when this assumption doesn't hold (Geeks for Geeks, 2024).
• Struggles when the relationship between features is complex or correlated (Kumar Bhowmik, 2015).

Analysis and Evaluation

- Our experiment shows that overall, the Logistic Regression model have outperformed the Naïve Bayes model in most of the experimental results. Especially in terms between validation result (as Logistic Regression has 76.4% in accuracy while Naïve Bayes have at least 67.6%). As our result, the entire result supports the hypothesis statement claimed earlier due to its due to its ability to model feature interactions implicitly. As mentioned by
- AUC is a measurement in which it represents the degree or measure of separability. When the AUC is initially 0.5, there isn't any capacity to distinguish between positive class and negative class however once you increase from 0.5 onwards, one can be able to see a classification .Looking at the table we can see that Logistic Regression has a significantly higher AUC (0.875) compared to Naïve Bayes (0.80324), suggesting that Logistic Regression performs better in distinguishing between classes.
- If we were to look at the Training Accuracy between LR and NB, we are to see that the NB has a higher training accuracy compared to LR, this may be because NB generally performs better when the data distribution is more probabilistic. However, this may not be as seen as a reliable result since Training Accuracy could possibly lead to overfitting.
- Logistic Regression has a lower error rate compared to Naïve Bayes, which means it is making fewer incorrect predictions on the validation set. This supports the observation that Logistic Regression is better at generalizing to new data.
- Logistic Regression has a significantly higher Precision than Naïve Bayes (LR: 76.7% NB: 64.6%) meaning that it has a better approach when dealing with predicting the presence of patients with heart disease.
- The time for Logistic Regression is faster than the time of Naïve Bayes, making it more efficient.
- The hyperparameters for the Logistic Regression were picked from a range [0.0001, 0.001, 0.01, 0.1, 1], most of the time; the hyperplane with the highest accuracy was 0.0001, making sure that the model minimizes the loss function.
- For Naïve Bayes, the smoothing value, was handpicked at 3, through basic trial and error, providing model classification results that looked good enough, and for the ROC curve to get a curve that is not too overfitting.

Lessons learned and future work:

- While using grid search seem for hyperparameter optimization, I believe that they are other alternatives that could be used like Random Search, Bayesian Optimization, which can lead to more efficient findings.
- Rather than just picking a smoothing value manually and going through trial and error to ensure that the results are well, training the Naive Bayes model with different smoothing values (using grid search) and evaluate it on the validation set.
- When dealing with finding the best value for lambda (the hyperparameter) in the Logistic Regression model, exploring a wider range of values to see which hyperparameter is the best of them, and then using it can produce a higher result.
- As I mentioned before, there were noisy data (missing entries), and so removing them would most likely have made error and mistake in terms of my findings, making them no as reliable as hoped. And so, preprocessing the data and adding more steps of data wrangling to address any possible outliers.
- Applying the SMOTE (Synthetic Minority Over-sampling Technique) to deal with imbalanced data.
- Explore other model such as Random Forests and SVM.

Hypothesis Statement:

- Logistic regression will outperform Naive Bayes due to its ability to model feature interactions implicitly through its linear model. (Geeks for Geeks, 2024).
- However, Naïve Bayes may perform well in probabilistic scenarios, as it is grounded in Bayes' Theorem, which provides a solid foundation for handling conditional dependencies (Learning Classifiers based on Bayes Rule, 2015).

Methodology:

- As previously mentioned at the beginning the target dataset shows a 0 if there's an absence in heart disease and numbers 1-4 for specific cases of heart disease
  - From this we shall manipulate the dataset so that the target will only show 1 (the presence of heart disease in general) and 0 for no heart disease seen)

From this point we will:

- Split the dataset into training (80%) and testing (20%) subsets
- Apply the 10-fold classification validation on the training set
- From this we can find patterns such as accuracy, error, etc.
- we will also be looking out for any misclassifications

Parameter choices and the experimental results,

Logistic Regression

- using grid search to find the optimal hyperparameters for a logistic regression model with Lasso regularization as a feature selection.
- The choice of parameter is when lambda (the regularization strength) is 0.0001, giving the best of its accuracy in comparison to other values of lambda.

Naïve Bayes

- Also using grid search to find the optimal hyperparameters foe the naïve bayes using Kernel smoothing.
- Smoothing value picked as 3 through trial and error.

Choice of Parameters:

- **Logistic Regression**: Default parameters with L2 regularization.
- **Naive Bayes**: Gaussian distribution assumption for continuous features.
-

Experimental Results:
Performance metrics include AUC, accuracy, precision, recall, and F1-score, averaged across all folds for consistency.

Metrics	AUC	Validation Accuracy	Training Accuracy	Avg. Time	Precision	Recall	F1 Score	Error
Logistic Regression	0.875	0.76431	0.78477	0.0052916	0.76743	0.71818	0.73369	0.23569
Naïve Bayes	0.80324	0.67627	0.80626	0.0088004	0.646695	0.65455	0.73369	0.32373

Figure 4: Classification Results

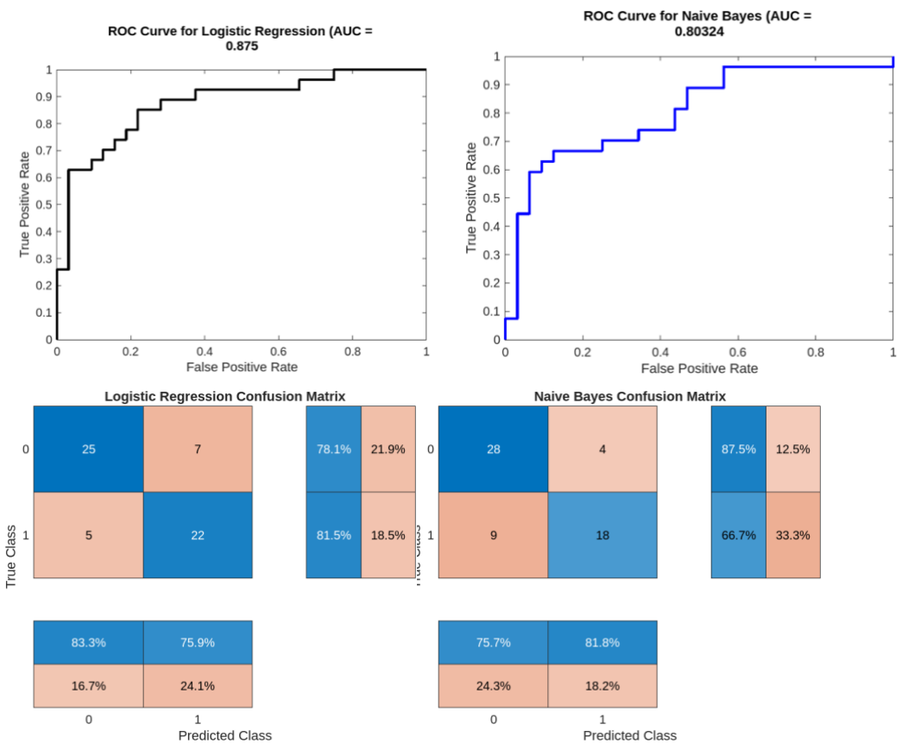


Figure 5: ROC Curves

Figure 6: Confusion Matrices

References:

- Simplilearn.com. (n.d.). *An Introduction to Logistic Regression in Python*. [online] Available at: https://www.simplilearn.com/tutorials/machine-learning-tutorial/logistic-regression-in-python#what_is_logistic_regression.
- Geeks for Geeks. (2024). *Naive Bayes vs Logistic Regression in Machine Learning*. [online] Available at: https://www.geeksforgeeks.org/naive-bayes-vs-logistic-regression-in-machine-learning/.
- Learning Classifiers based on Bayes Rule. (2015). Available at: https://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf.
- uk.mathworks.com. (n.d.). *Choosing the Best Machine Learning Classification Model and Avoiding Overfitting*. [online] Available at: https://uk.mathworks.com/campaigns/offers/next/choosing-the-best-machine-learning-classification-model-and-avoiding-overfitting.html.
- Kumar Bhowmik, T. (2015). Naive Bayes vs Logistic Regression: Theory, Implementation and Experimental Validation. *INTELIGENCIA ARTIFICIAL*, 18(56), pp.14–30. doi: https://doi.org/10.4114/ia.v18i56.1113.
- Narkhede, S. (2018). *Understanding AUC - ROC Curve*. [online] Medium. Available at: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.