Bayesian Methods for Heart Attack Prediction

Lilit Khachatryan, BS in Data Science, American University of Armenia
Supervisor: Tatev Kyosababyan
University of Virginia

***Abstract*—A heart attack is a significant and life-threatening medical emergency that requires early identification and diagnosis for timely prevention and treatment. The objective of this project is to develop a model using methods of Bayesian statistics, which helps estimate and forecast individuals eager for heart attacks. The model estimates the risks of a heart attack using an existing dataset with common factors that can affect the individual's heart condition, potentially leading to an increased risk of the occurrence of a heart attack.**

I. Introduction

Cardiovascular Diseases (CVDs) are the most common causes of death worldwide, taking estimated 17.9 million lives each year [1]. More than four out of five deaths caused by cardiovascular disorders are because of heart attacks among people under the age of 70 [2]. Heart attacks, or myocardial infarctions, occur when a blood clot blocks blood flow. Consequently, oxygen and nutrients are unable to reach the heart muscles, resulting in the subsequent death of the latter. Certain factors like high levels of cholesterol, age, ethnicity, genetic propensity, and diabetes are indicators of risk. To prevent the cause, it's essential to diagnose and check the likelihood of the individual having a heart attack relying on multiple risk factors. Within the scope of this capstone project, different models were created for early diagnosis. The goal is to create a practical and modern implementation of Bayesian classification techniques to analyze the uncertainty in the association of the risk factors with getting a heart attack.

II. Background and Literature review

The main goal of this project is to solve the classification problem of predicting heart attacks, which is done by utilizing Bayesian statistical methods. The Bayesian statistical model uses Bayes' theorem to update the probability of a hypothesis as new evidence becomes available. In Bayesian Analysis, prior knowledge about the parameter is combined with the likelihood of the data to create the following parameter distribution after considering the observed data. The Bayesian method is the best suited for this project, as it's very flexible and can incorporate prior knowledge and uncertainty into analysis and update new information [3]. We used the Highest Density Interval (HDI) statistical tool to estimate the range of the values within which a certain percentage of a probability distribution lies. It's usually used as an alternative to confidence intervals in Bayesian statistics, particularly when dealing with multimodel distributions. HDI also can be used to make probabilistic statements about a population parameter based on a sample from that population. It's a helpful tool for decision-making under uncertainty and for communicating the level of confidence in statistical estimates [4]. No U-Turn Sampling (NUTS) is a Markov Chain Monte Carlo (MCMC) algorithm used in Bayesian statistics. It draws from a posterior distribution, iterating the parameter space and exploring it, changing the step size based on the shape of the distribution, allowing for more efficient and effective sampling compared to the traditional MCMC algorithms. In the project, NUTS is used to obtain posterior samples, and generate reduced models [5]. To compare the models, we used Applicable Information Criterion (WAIC). It is used to compare the predictive accuracy of different models and select the best among them. WAIC considers two factors: how the model fits and the complexity of the model [6]. Several studies explored using the Bayesian statistical model for predicting heart attacks. One is a study by Khan et al. (2018), who developed a Bayesian model using data from the Framingham Heart Study to predict the risk of heart attack in individuals. The studies describe how Bayesian methods outperform traditional methods in terms of accuracy [7]. Another study by Datta et al. (2020) developed a Bayesian model for predicting heart attacks using electronic health record data. The study shows that the methods used were effective despite lacking data conditions [8].

III. Data Analysis

The dataset used for this project is taken from the Kaggle website [9], which was published by the University of California, Irvine [10]. The dataset contains information about 303 individuals with 13 features that can be considered risk factors for further analysis. These predictor variables contain data such as sex, chest pain, fasting blood sugar levels (1 if the levels are greater than 120 mg/dl, 0 otherwise), category of normality of resting electrocardiographic results, the existence of exercise-induced angina, the slope of the peak exercise, etc. The continuous variables, which are age, resting blood pressure, cholesterol levels, maximum achieved heart rate, and ST depression induced by exercise relative to rest, have been standardized. The outcome variable indicates a diagnosis of heart disease, which has binary values of 0 and 1 (0: < 50% diameter narrowing. less chance of heart disease; 1: > 50% diameter narrowing. more chance of heart disease).

The heatmap (Fig. 1) was created to find correlations between features. The latter shows a moderate association between the outcome variable and features of chest pain type (cp) and the maximum heart rate achieved during exercise (thalach) [12]. Although the relationship is not very strong, there is a significant relationship between the outcome variable

and these two features. The heatmap also shows a weak relationship between the outcome variable and the slope of the peak exercise ST segment (slp). On an electrocardiogram (ECG) ST segment is a waveform that represents the heart's electrical activity, and any change on this slope indicates cardiac issues [13]. The features of the slope of the peak exercise ST segment and the maximum heart rate achieved are also moderately correlated.

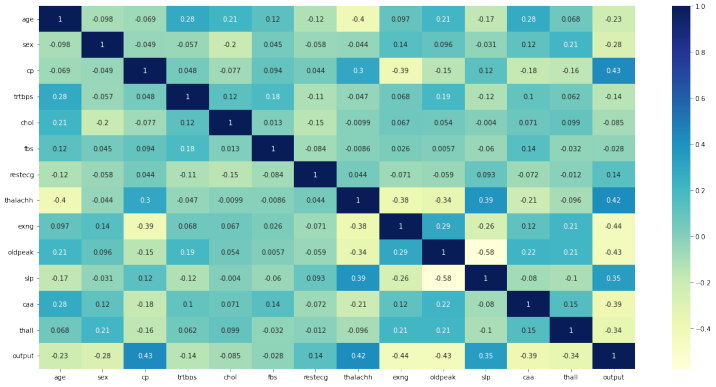


Fig. 1: Heatmap for Correlation Analysis

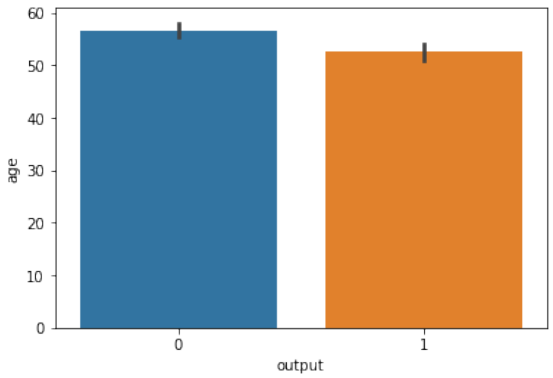


Fig. 2: Distribution of the classes of the outcome variable

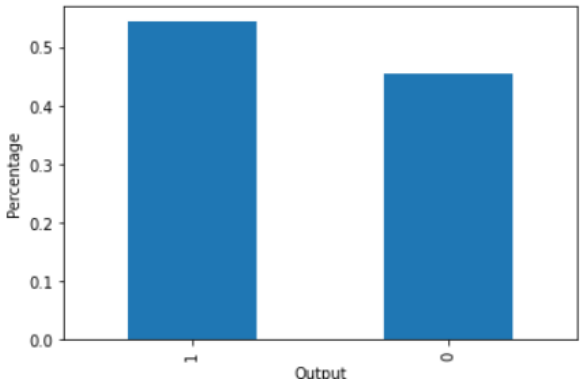


Fig. 3: Percentage of Outputs

Plotting the distribution of the age in regard outcome variable is shown in Fig.2. and Fig. 3 shows percentages of the outputs. The data is rather balanced, as the ratio between classes is 0.54:0.45 (Fig. 3).

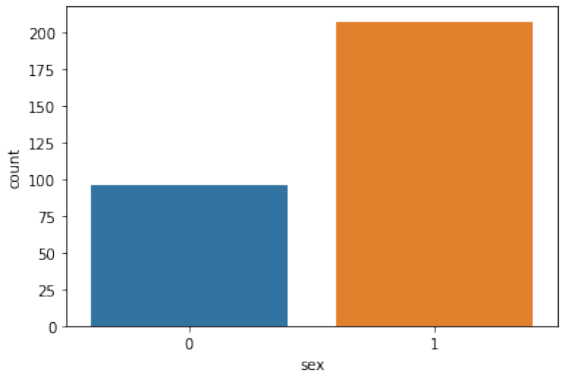


Fig. 4: Distribution of Gender in Heart Attack Patients

The count plot shows the distribution among females (0) and males (1). From the plot, we can see that number of male patients is significantly higher than females. It means that the dataset is skewed toward males. In the case of heart disease, it could indicate that males are more prone to heart disease or that there are more males with heart disease than females in the population from which the data was collected.

The count plot displays the distribution of chest pain types among the patients in the dataset, where 1 represents typical angina, 2 represents atypical angina, 3 represents non-anginal pain, and 0 represents asymptomatic patients. The y-axis represents the count of patients. From the plot, we can see that the most common chest pain type among the patients is asymptomatic. Typical angina is the second most common chest pain type. Atypical angina and non-anginal pain are less common chest pain types.

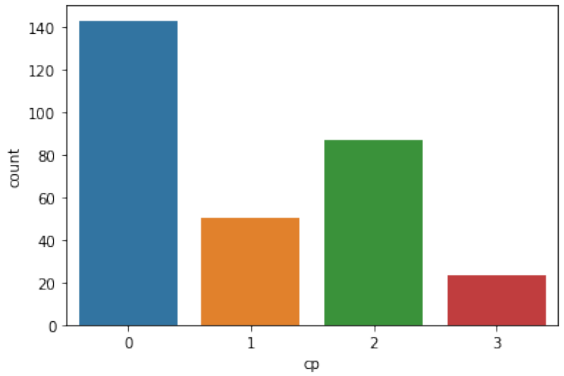


Fig. 5: Distribution of Chest Pain Types

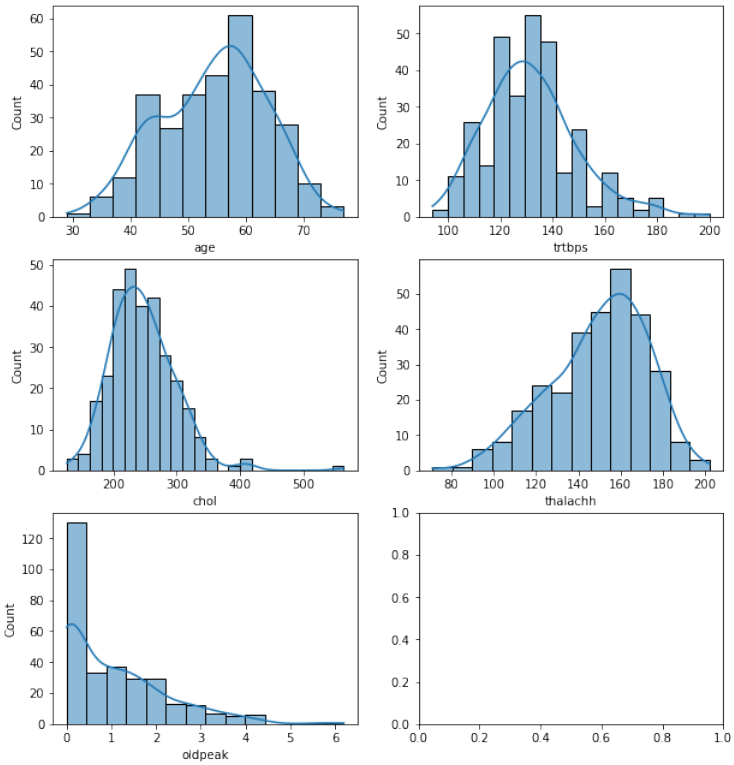


Fig. 6: Distribution of Numeric Variables

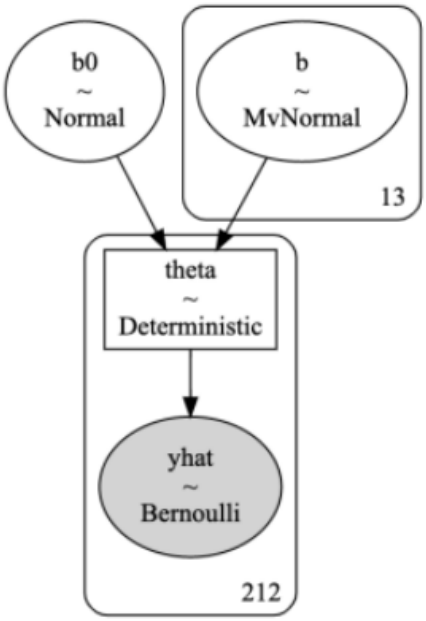


Fig. 7: Probability model

From the multiple numeric distributions in Fig. 4, we can see that the distribution of ages is relatively uniform; therefore, we can state that the least skewed variable is age. For this classification problem, we apply Bayesian Logistic regression, which allows us to use prior knowledge and update it with new information. With this method, we predict the likelihood of a heart attack based on the patient's risk factors.

IV. Probability Model

The predictor variables are assumed to be normally distributed. Relying on this assumption, the joint distribution of the predictor variable can be modeled using a Multivariate-Gaussian distribution. The structure problem described in Fig. 7 represents a graphical model showing the relationship between the predictor variable, outcome variable, and relevant variables in the problem. Calculation of the response variable was done based on the values of the predictor variables and a deterministic function. We assume the Bernoulli probability distribution for the outcome variable, which models binary events.

Obtaining the full model built on a dataset can be seen in HDI (Highest Density Interval) plot in Fig. 8, which includes all variables in the dataset. We use the plot to assess the probability intervals for the predictors representing fasting blood sugar, the maximum heart rate achieved, and the ST depression factor induced by exercise relative to rest are wider than the others. This indicates that there is more uncertainty in the mentioned predictor variables than in the others. Removing variables that contain zero, from the confidence intervals, we can find that there were only 6 variables that we need to take into account: the resting blood pressure, cholesterol levels, fasting blood sugar levels, resting electrocardiographic results, ST depression factor induced by exercise relative to rest, and the number of the major vessels.

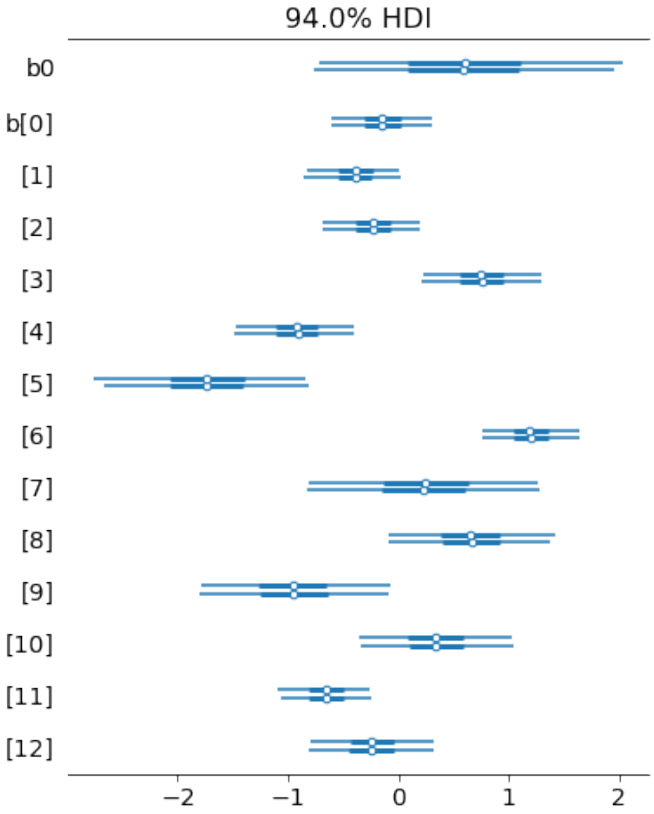


Fig. 8: HDI for the Full Model

Based on scientific research, we built a third model based on the known factors that were identified by the Centers for Disease Control and Prevention (CDC) [14]. According to the CDC, the most influential risk factors are cholesterol levels, fasting blood sugar levels, and blood pressure. However, according to Mayo Clinic, age, and sex are also essential characteristics to consider while predicting the risks of heart attack [15]. Therefore, we incorporated these vivid factors from both of these resources to construct the second model based on the mentioned variables and prior information. The models were compared with each other and combined. To improve the accuracy of approximations No U-Turn Sampling (NUTS) was used. NUTS is a part of the Bayesian approach, which is a type of Hamiltonian Monte Carlo Sampling. To inspect the convergence of the Markov Chain Monte Carlo (MCMC) algorithm, we use traceplots of the full model. Fig. 9 shows a good convergence; therefore, we can infer that the MCMC algorithm works correctly and produces reliable results.

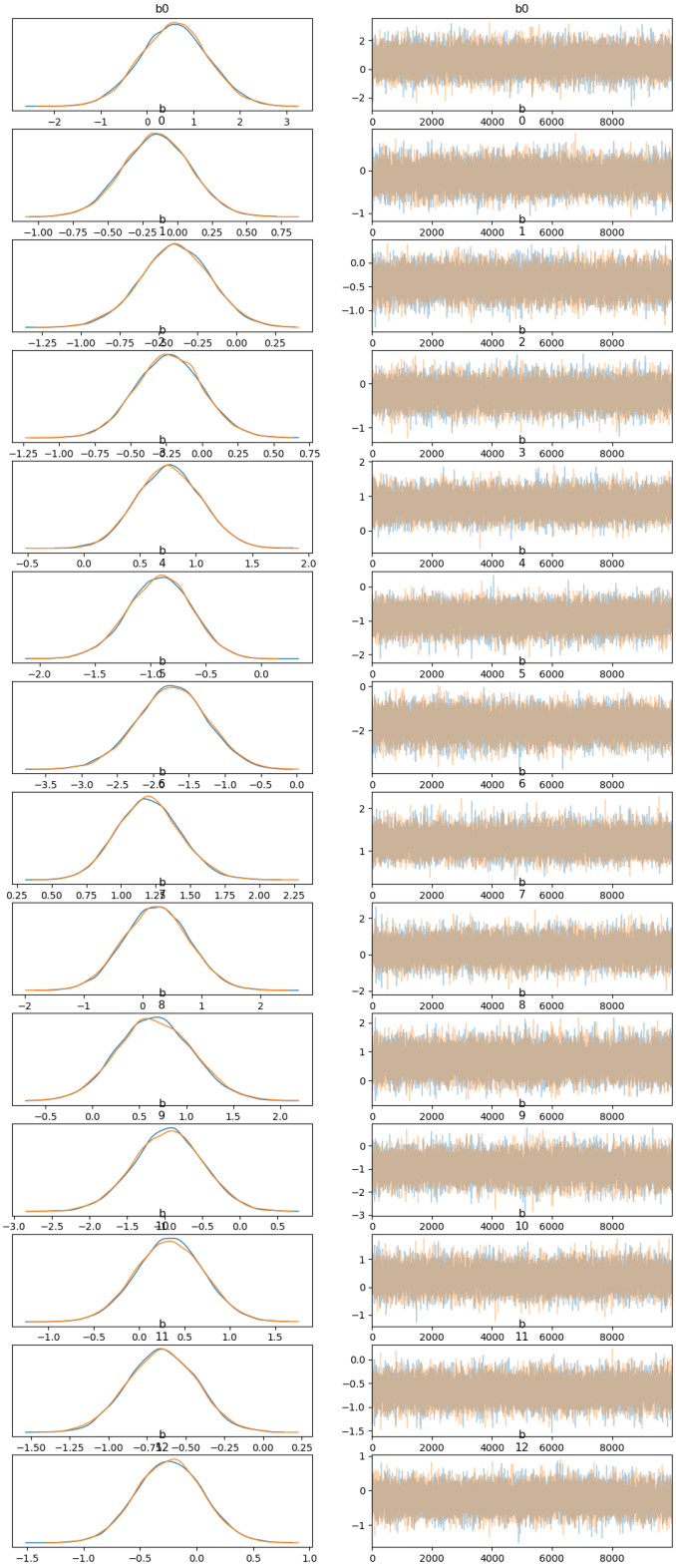


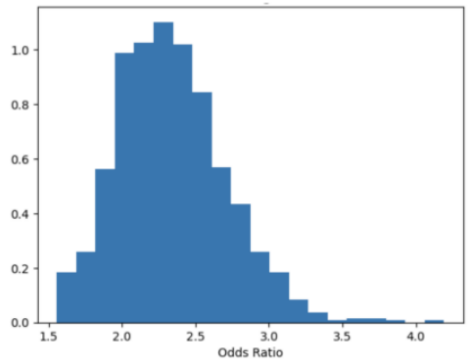
Fig. 9: Full Model Traceplots

To compare the models, we used the Widely Applicable Information Criterion (WAIC); as one of the commonly used leave-one-out cross-validation methods, it is used to estimate out-of-sample prediction accuracy.

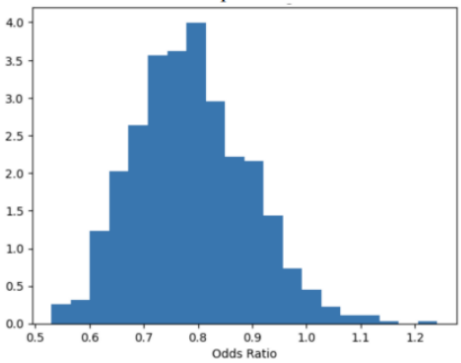
	rank	elpd_waic	p_waic	elpd_diff	weight	se	dse	warning	scale
full_model	0	161.224968	12.025636	0.000000	0.882057	18.107211	0.000000	True	deviance
reduced_model	1	212.842547	4.323012	51.617578	0.016631	17.072588	14.431683	False	deviance
med_prior_model	2	215.423147	6.884594	54.198179	0.101312	16.095409	16.327499	True	deviance

Fig. 10: Model Comparison Table

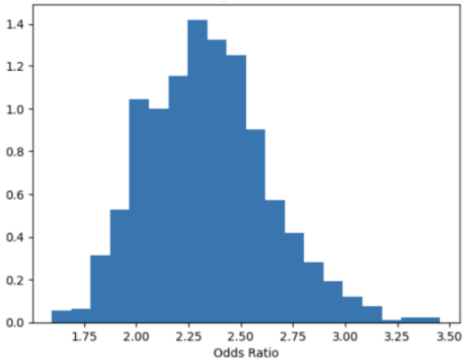
The full model is significantly better than the reduced model (Fig. 10). But we can see that the reduced model still has some value; therefore, we used the pseudo-Bayesian mode averaging method to combine models using the weights that were generated by WAIC. The combined model shows the weighted average of each model and provides a more precise prediction.



(a) Full Model Age Odds Ratio



(b) Prior Information Age Odds Ratio



(c) BMA Model Age Odds Ratio

Fig. 11: Odds Ratios for the models

Age was the only variable that was less skewed and more likely to be normally distributed, that's why we created odd ratio plots to compare the effect of this variable on having a heart disease given a certain age. Fig. 11 (a), (b) and (c) show the odds ratios for the full model, the first reduced model, and the second reduced model, respectively. Plots show that age has a positive effect on the likelihood of having a heart attack, stating that age is a well-known risk factor. The other variable depending on age has varying effects. Important to mention that the full model was pushing the reduced model odds ratio higher, and the former is closer to the Bayesian model Averaging (BMA) model due to its weight.

V. Results

Based on the WAIC score results among the three models full model has the lowest results, which means that it is the best-performing model in terms of predictive accuracy. The score of 161.22, the full model, indicates that it is a better fit for the data than the reduced model. Besides, the full model has the highest weight model, which states that it is the most precise model and should be considered in future predictions. The reduced model is ranked as the second-best fitting model as it performs reasonably well after the full model. The model was trained using 70% of the original data and was tested using the rest 30% to make inferences (Fig. 12).

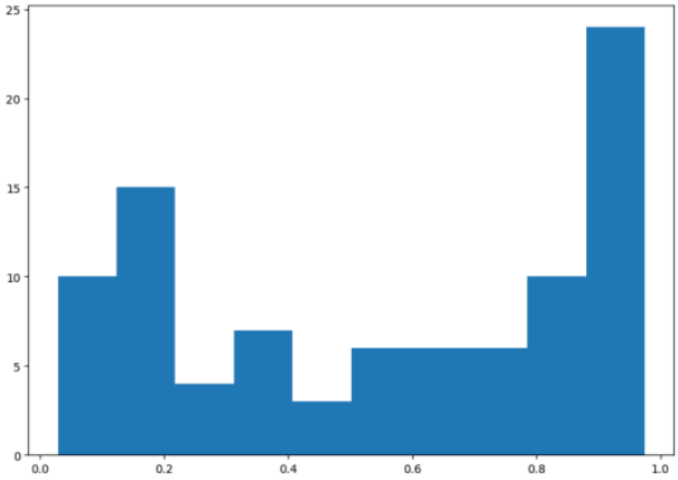


Fig. 12: Distribution of Predictions

The thresholds and accuracy scores were defined considering the medical nature of the problem and the cost of misclassification. As the problem includes real medical cases and life factors, we define the cost of false negatives as more expensive than the cost of false positive results. The reason is the misclassification of a patient under the risk and has more significant consequences than a healthy patient diagnosed with positive results.

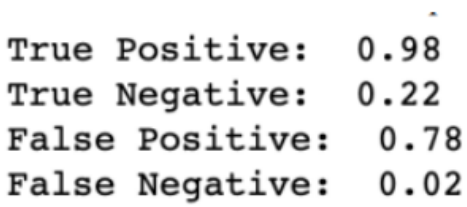


Fig. 13: Percentages of the Results

The accuracy score of the model is 63.73%, indicating that the model accurately predicted 63.73% of cases regarding the cost of false positive being one and false negative being 8 (Fig. 13).

VI. Conclusion

The project's objective was to understand how the different features were associated with the risk of getting a heart attack for each individual, which we could identify with the help of the forest plots. The developed model indicated an accuracy score of 63.73%, considering the intentionally implemented threshold for minimizing false negative outcomes. The limitations of the work are mainly focused on the binary interpretation, the number of categorical features, as well as the size of the data. In the future, with the help of variational inference, it would be possible to quantify the uncertainty of the variables.

VII. Guidelines

The code and README file of instructions for this project can be found at https://github.com/lilit07/capstone.git Github Repository.

VIII. References

1. Johns Hopkins Medicine. (n.d.). Heart Attack. https://www.hopkinsmedicine.org/health/conditions-and-diseases/heart-attack

2. World Health Organization. (2021). Cardiovascular Diseases. https://www.who.int/health-topics/cardiovascular-diseasestab=tab$_1$

3. Shariful Islam, S. M., Tabassum, S., Sarker, M. A. R., Rawal, L. B. (2021). Applying Bayesian methods for heart attack prediction: A review. Statistical Methods in Medical Research, 30(5), 1555–1576. https://doi.org/10.1080/13683500.2021.1896486

4. Kruschke, J. K. (2015). "BEST: Bayesian estimation supersedes the t test". Journal of Experimental Psychology: General, 144(2), 573-603. https://doi.org/10.1037/xge0000100

5. Hoffman, M. D., Gelman, A. (2014). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". Journal of Machine Learning Research, 15(1), 1593-1623. http://www.jmlr.org/papers/volume15/hoffman14a/hoffman14a.pdf

6. Watanabe, S. (2010). "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." Journal of Machine Learning Research, 11, 3571-3594. http://www.jmlr.org/papers/volume11/watanabe10a/watanabe10a.pdf.

7. Rathore, S. S., Curtis, J. P., Wang, Y. (2011). Cardiovascular Disease in Chronic Kidney Disease. In S. S. Rathore (Ed.), Cardiovascular Complications in Chronic Kidney Disease (pp. 1–22). Springer New York. https://doi.org/10.1007/978-1-4419-7185-2$_1$

8. Al-Najafi, R. E., Saatchi, R., Burke, D. (2020). A Data-Driven Machine Learning Algorithms for Heart Attack Prediction. Bioengineering, 7(4), 122. https://doi.org/10.3390/bioengineering7040122

9. Rahman, M. R., Pritom, R. (2021). Heart Attack Analysis Prediction Dataset. Kaggle. https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset?select=heart.csv 10. Dua, D., Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

11. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K. H. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64(5), 304–310. https://pubmed.ncbi.nlm.nih.gov/2756873/

12. Boston University School of Public Health. (n.d.). Correlation and Regression. https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module9-Correlation-Regression/PH717-Module9-Correlation-Regression4.html

13. National Center for Biotechnology Information. (2019). The Genetic Landscape of Diabetes. In N. R. Gough (Ed.), Endotext [Internet]. MDText.com, Inc. http://www.ncbi.nlm.nih.gov/books/NBK459364/

14. Centers for Disease Control and Prevention. (2021). Heart Disease Facts. https://www.cdc.gov/chronicdisease/resources/publications/factsheets/heart disease-stroke.htm

15. Mayo Clinic. (2021). Heart Attack. https://www.mayoclinic.org/diseases-conditions/heart-attack/symptoms-causes/syc-