Heart Disease Risk Prediction

Using Bayesian Logistic Regression

Angela Valerie Christy, Calvina Adelia Sucipto,

Cheryl Aurellia Valencia, and Sammer Violeta Liu

Bina Nusantara University, Faculty of Computer Science, Jakarta, Indonesia

Bina Nusantara University, Faculty of Computer Science, Jakarta, Indonesia

Bina Nusantara University, Faculty of Computer Science, Jakarta, Indonesia

Bina Nusantara University, Faculty of Computer Science, Jakarta, Indonesia

Date of Submission : Friday, 27 December 2024

ABSTRACT

This research is used to predict the risk of heart disease using a Bayesian logistic regression model. The dataset consists of 1,025 rows and 14 columns. Analysis is done to determine the relationship between several selected variables (X) and the probability of an individual developing cardiovascular disease (Y). Bayesian was chosen because it is considered capable of integrating prior information and handling complex data. The model shows that variables such as chest pain type (cp), ST depression after exercise (oldpeak), and exercise-induced angina (exang) have an impact on the risk of disease. The results suggest that this model can make accurate predictions and provide useful insights, but there are some drawbacks, like the small size of the dataset and the assumption that relationships are linear. This study lays the

groundwork for creating tools that use data to help doctors make better decisions in preventing and diagnosing heart disease.

INTRODUCTION

I. Background

One of the greatest challenges on the society today is cardiovascular disease. According to the WHO (World Health Organization), in 2023 worldwide, the number of people that have heart disease annually was 17.8 million. In Indonesia, 9.025.500 cases were reported in 2023 by the Ministry of Health. In 2023 the death caused by cardiovascular disease is counted around 19.42%. The older you get, the greater risk of heart disease you will have, but it is not impossible for younger people to reach it too.

Cardiovascular disease is caused by stimulation and chemical substances from food or drinks so that it causes plaque on vessels in the blood arterial. This occlusion leads to blood clots in the arteries which interrupts the flow of blood. But there are many types and causes of heart disease. First is coronary artery disease, the most common type that generally causes blood vessels in the heart to be blocked. There are also congenital heart defects, which are a newborn with a heart disease since birth and the cause is still unknown. The other one is heart infection, which is the infection in the innermost lining of the heart and the name of the bacteria that causes it called streptococcus beta-hemolyticus type A. Next is heart failure that occurs when the heart muscle struggles to pump blood effectively is another type that is dispersed throughout the body. Last one, arrhythmia is a type of heart rhythm disorder that causes abnormal heartbeats.

The disease can also be triggered by adopting an unhealthy lifestyle, like smoking, being repeatedly exposed to cigarette smoke, not exercising, eating lots of sugar and fats, stress and lack of sleep. These factors can put you at a higher risk for heart disease. One should have a balanced lifestyle to risk the less. Aside from lifestyle, there is a genetic factor too. People with a family history of coronary artery disease have a greater risk of heart disease. Research has also shown that men are at greater risk of dying from heart disease than women.

The impact of this disease can affect various physical activities, such as chest pain or angina during physical activity. People with heart disease often experience shortness of breath because their heart struggles to deliver oxygen to the body because of the occlusion in the blood vessels. Because of the heart disease you will feel everyday tasks can become much more difficult and tiring. Heart disease can affect anyone, so it is important to have the knowledge on how to prevent it.

II. Statement of the research problem

Heart disease is one of the most important diseases to be aware of and prevent due to the high number of deaths caused by this condition. One fact that needs to be known is that most countries with the highest number of heart disease patients are low- and middle-income countries. This is because people do not receive enough information about the dangers of this disease and are unaware of the preventive steps that can be taken to avoid cardiovascular disease. As a result, many patients with advanced heart disease go undiagnosed, making it more difficult to treat compared to those in the early stages.

In addition, understanding the risk factors for heart disease is crucial. Unhealthy lifestyle habits are one of the main reasons why someone can develop heart

disease. Bad habits such as being sedentary and lacking exercise, combined with poor eating habits, contribute significantly. Sweet foods and junk food are now easily accessible everywhere. These two factors can make a person more susceptible to diabetes and obesity. Smoking, which is difficult to avoid, is another major issue. The problem is that not only active smokers are at risk, but also those who inhale secondhand smoke are affected.

The difficulty in accessing adequate healthcare services is also a major concern regarding why heart disease is better prevented. As mentioned earlier, most countries with high mortality rates due to cardiovascular disease are low-income countries, largely due to limited access to healthcare. People in remote areas, for example, often lack proper healthcare services. Additionally, heart disease treatment is relatively expensive, so many people choose not to seek treatment for their condition.

III. Objectives of the study

The purpose of this project is to identify the risk factors that can cause someone to develop cardiovascular disease. One of the goals is to identify the main risk factors that contribute to the development of cardiovascular disease. Then, the project aims to develop a data-driven model to predict an individual's risk of developing cardiovascular disease. Furthermore, this project is expected to provide data-driven recommendations for more effective public health policies in the prevention and management of cardiovascular disease.

METHODOLOGY

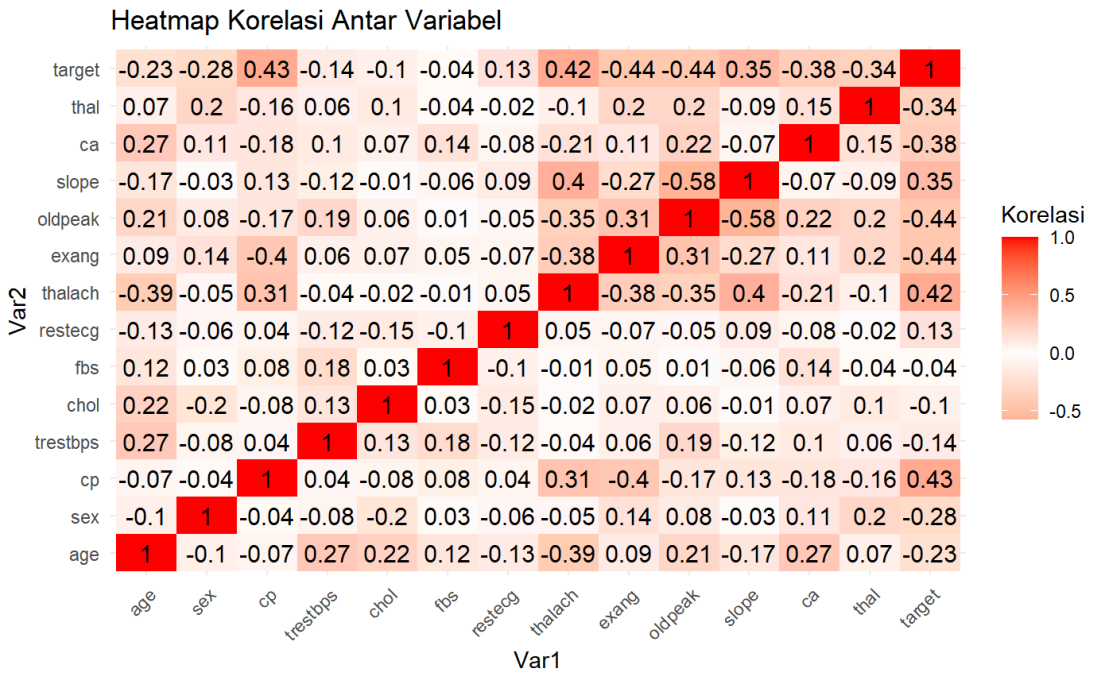
I. Description of the dataset

The dataset used is titled ‘heart.csv’ and is available on Kaggle. We chose this dataset because it contains several predictor variables related to cardiovascular disease risk. This dataset consists of 14 columns and 1025 rows of data.

Here is a list of the variables and their explanations in the dataset:

- **age**: Age of the patient.
- **sex**: Gender (1 = male, 0 = female).
- **cp**: Type of chest pain (0-3, 0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic).
- **trestbps**: Resting blood pressure (mmHg).
- **chol**: Cholesterol level (mg/dL).
- **fbs**: Fasting blood sugar (> 120 mg/dL: 1, others: 0).
- **restecg**: Electrocardiographic results (0-2, categories).
- **thalach**: Maximum heart rate achieved (bpm).
- **exang**: Exercise-induced angina (1 = yes, 0 = no).
- **oldpeak**: depression.
- **slope**: Slope of the ST segment (0-2, categories).
- **ca**: Number of major vessels colored by fluoroscopy (0-3).
- **thal**: Thalassemia type (1 = normal, 2 = fixed defect, 3 = reversible defect).
- **target**: Heart disease indication (1 = positive, 0 = negative).

To explore the relationships between variables, a correlation heatmap has been created.



Based on the heatmap, we observed that the variables exang, oldpeak, and cp show a strong correlation with the target variable and have minimal multicollinearity among them. Therefore, these variables will be used as predictors in this Bayesian regression model.

II. Justification for the chosen Bayesian regression model

The reason we chose the Bayesian Logistic Regression model in this study is based on its ability to handle the data analysis needs required to analyze cardiovascular diseases. Using the Bayesian regression model can help compare existing knowledge or information to identify risk factors, which can enhance the analysis. With a Bayesian model, we can understand which variables have the strongest relationship with the desired outcome. Complex data also requires a flexible model to handle complexities such as managing non-linear relationships between predictor variables and outcomes or handling missing information.

The Bayesian model helps prevent errors due to overly complex data interpretation, such as analyzing many risk factors like age, gender, blood pressure, blood sugar levels, and others. This model can still produce stable results. One of the goals of this project is to provide accurate predictive data analysis, where the Bayesian model predicts the likelihood of the target outcome based on the presented data. The Bayesian regression model also offers various options, allowing us to choose the one we believe is best suited for the dataset we selected.

III. Details of the modelling approach and parameter estimation techniques

1. Modelling Approach

Bayesian logistic regression model is defined as:

$$P(Yi = 1 \mid x_i; \theta) = \frac{1}{1+e^{(-\theta^T xi)}}$$

Atau

$$P(Yi = 0 \mid x_i; \theta) = 1 - \frac{1}{1+e^{(-\theta^T xi)}}$$

Where:

- y_i : Outcome variable (presence/absence of cardiovascular disease).
- x_i: Predictor variables (age, gender, cp, trestbps, etc).
- θ: Model parameters.

The prior distributions are chosen based on existing assumptions, with weakly informative priors selected to incorporate prior knowledge from the data. These priors are designed to reflect prior studies or expert knowledge that indicate strong effects of each factor variable on cardiovascular risk, while still allowing the data to guide the posterior estimates. By using weakly informative priors, the model

ensures that the influence of prior information is not overly restrictive, thus allowing the data to have an appropriate impact on the analysis.

As for the likelihood is chosen based on the type of outcome variable, specifically the target variable. For binary outcomes, a Bernoulli likelihood is used, which models the probability of success or failure. This is paired with a logit function, which links the linear predictors to the probability scale, ensuring that the estimated probabilities remain between 0 and 1. The Bernoulli likelihood is appropriate for modeling situations where the outcome is binary, such as the presence or absence of a disease or event.

2. Parameter Estimation

Bayesian inference is used to estimate model parameters by updating prior beliefs with observed data. The posterior distribution is computed as:

$$P(\theta|data) = \frac{P(data|\theta)\,P(\theta)}{P(data)}$$

Where:

- P(θ | data): Posterior distribution.
- P(data | θ): Likelihood function.
- P(θ): Prior distribution.
- P(data): Marginal likelihood (normalizing constant).

Posterior distributions are usually more complex and cannot be calculated directly. Therefore, numerical methods like Markov Chain Monte Carlo (MCMC) are used. One of the techniques is Gibbs Sampling**.** It simplifies data sampling by breaking it into conditional distributions.

In order to make sure that our MCMC estimations are reliable, Convergence diagnostics are carried out. One of the most used diagnostic tools is Gelman-Rubin Statistic that checks if the chains converge to the same distribution using Gelman Rubin statistic. If the result is 1 then it means the chains have converged. Otherwise, it means that the chains are not converged. The other tool used is Geweke which checks if the chains converge to the same distribution using Geweke. The result must be close to 0 for the chains called converged.

Then, posterior distributions are summarized using Mean or Median and Credible Intervals. Mean is the average value of the parameter while median divides the posterior distribution into two equal halves. On the other hand, Credible Intervals are the confidence intervals in Bayesian context.

3. Model Validation and Comparison

Posterior Predictive Check (PPC) was used as model validation to assess how well the model fits the observed data. Whereas model comparison, WAIC (Watanabe-Akaike Information Criterion), DIC (Deviance Information Criterion), and Bayes Factor were used:

- WAIC (Watanabe-Akaike Information Criterion): Balances model fit and complexity.
- DIC (Deviance Information Criterion): a model selection that used the complexity of a Bayesian model.
- Bayes Factors: Compare the models that we use.

RESULTS

I. Statistical Summary of Posterior

The picture below shows statistical summary of the posterior :

```
Iterations = 2001:12000
Thinning interval = 1
Number of chains = 5
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

         Mean      SD  Naive SE Time-series SE
beta0  0.5912 0.13641 0.0006101      0.0015709
beta1  0.8103 0.08341 0.0003730      0.0008068
beta2 -0.9333 0.08811 0.0003940      0.0008069
beta3 -1.2433 0.18110 0.0008099      0.0014703

2. Quantiles for each variable:

         2.5%     25%     50%     75%   97.5%
beta0  0.3236  0.4990  0.5910  0.6826  0.8584
beta1  0.6483  0.7540  0.8098  0.8661  0.9754
beta2 -1.1086 -0.9922 -0.9318 -0.8736 -0.7644
beta3 -1.6002 -1.3655 -1.2434 -1.1210 -0.8914
```

The results show that the parameter β_0 has a mean value of 0.5912 with a relatively high standard deviation of 0.12641.The naive standard error is 0.0006101, while the time-series standard error, which accounts for autocorrelation in the Markov chain, is 0.0015709. The quantiles for β_0 show that the 2.5th percentile is 0.3236, the 25th percentile is 0.4990, the median is 0.5910, the 75th percentile is 0.6826, and the 97.5th percentile is 0.8584.

Meanwhile, for the parameter β_1, which uses the variable cp, the average result is 0.8103 with a standard deviation of 0.08341. As for the Naive Standard Error and Time Series Standard Error is at 0.0003730. The quantiles indicate that β_1 ranges from 0.6483 at the 2.5th percentile to 0.9754 at the 97.5th percentile, with a median of 0.8098. Additionally, the 25th and 75th percentiles are 0.7540 and 0.8661.

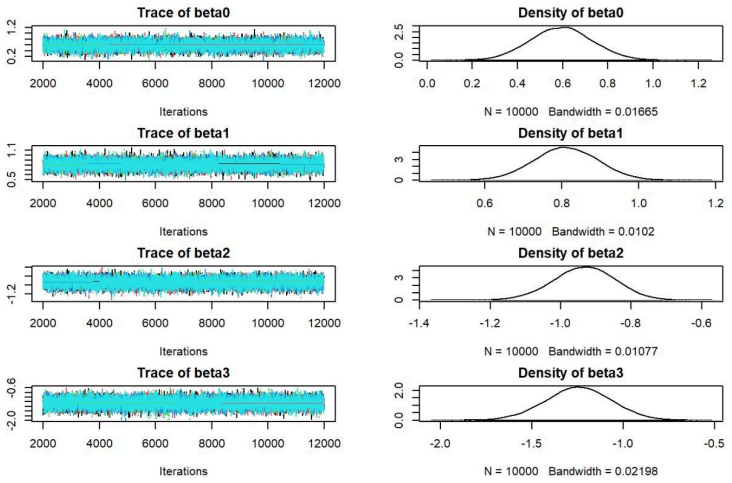
As for parameter β_2 that uses the variable oldpeak has -0.9333 for the mean value and 0.08811 for the standard deviation. The naive standard error is 0.0003940, while the time-series standard error is 0.0008069. Quantile analysis reveals that β_2 has a 2.5th percentile of −1.1086, a median of −0.9318, and a 97.5th

percentile of −0.7644 with −0.9922 as the 25th percentiles and −0.8736 and as the 75th percentiles.

Lastly, The mean estimate for β_3 which uses the variable exang is −1.2433 with a standard deviation of 0.18110. The naive standard error is 0.0008099, while the time-series standard error is 0.0014703. The 2.5th percentile of β_3 is −1.6002, the median is −1.2434, and the 97.5th percentile is −0.8914. The 25th and 75th percentiles are −1.3655 and −1.1210.

II. Trace and Density Plot

The picture below shows the trace plots and posterior densities of the parameters sampled in the MCMC process :

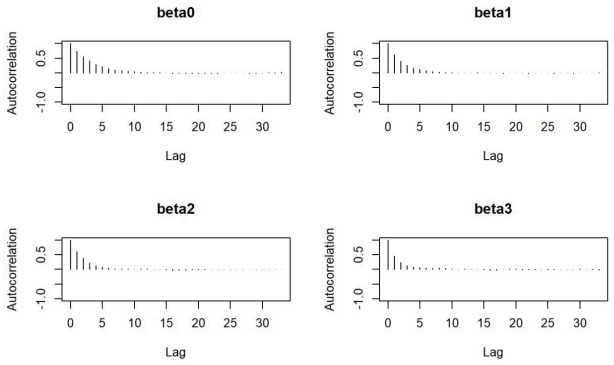


The trace plot for β0, β1, β2, and β3 shows stable increases or decreases without drastic changes throughout the iterations, with the parameter values fluctuating around a specific number. In the density plot, the distribution of each parameter appears smooth and bell-shaped. For example, β0 is around 0.59, β1 is around 0.81, β2 is around -0.93, and β3 is around -1.25. The spread is relatively tight, indicating consistent parameter estimates. The bandwidth for each parameter is small,

such as 0.01665 for β0, 0.0102 for β1, 0.01077 for β2, and 0.02198 for β3, suggesting that the results are quite good.

III. Autocorrelation Plot

The picture below shows the autocorrelation behavior of the parameters (β0, β1, β2, β3) in the MCMC sampling process:



The autocorrelation plots for parameters β0, β1, β2, and β3 show that the inter-sample relationship decreases as the lags increase. Initially, the correlation is quite high, but eventually decreases rapidly to near zero after a few lags. This is seen for all parameters.

IV. Convergence Diagnostic

1. Geweke Diagnostic

```
[[1]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

  beta0   beta1   beta2   beta3
 0.7531 -0.8539  0.4963 -1.4046


[[2]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

  beta0   beta1   beta2   beta3
-0.2514  0.5930 -0.2860 -0.6033


[[3]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

  beta0   beta1   beta2   beta3
-0.6249  0.2648  0.2263  1.1662


[[4]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

  beta0   beta1   beta2   beta3
 0.1777  0.2879 -0.6577 -0.3497


[[5]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

  beta0   beta1   beta2   beta3
 1.9221 -1.5251 -2.1261 -0.3147
```

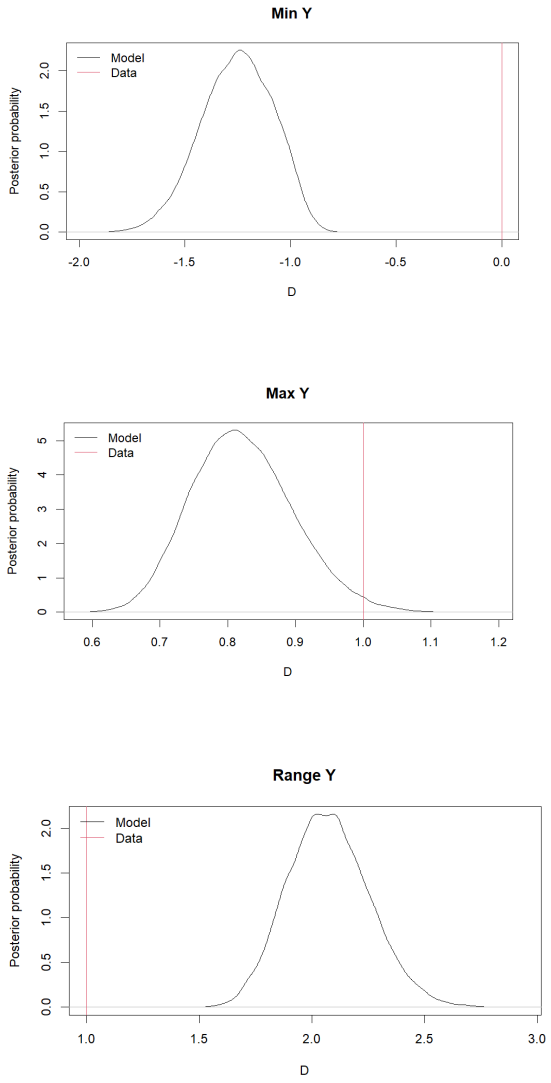
The Geweke diagnostic results show that the convergence of each chain varies. Chain 2 and 4 shows an indication of full convergence, with all parameters close to zero. Chains 1, 3, and 5 have some parameters that have not converged, such as β3 (-1.4046) in Chain 1, β3 (1.1662) in Chain 3, β0 (1.9221), β1 (-1.5251), and β2 (-2.1261) in Chain 5.

2. Gelman-Rubin Diagnostic

```
Potential scale reduction factors:

      Point est. Upper C.I.
beta0          1          1
beta1          1          1
beta2          1          1
beta3          1          1

Multivariate psrf

1
```

Based on the Gelman-Rubin diagnostic results, the model has achieved convergence for all parameters. This gives confidence that the posterior results are reliable, although there was previously a small indication from the Geweke diagnostic that there may be partial convergence on some chains.

V. PPC



The figure above shows a graph of the Posterior Predictive Check (PPC) results. The PPC results compare the posterior distribution of the model (shown with a black line) with the actual data (marked with a red vertical line) for three statistics: Min Y, Max Y, and Range Y. In the Min Y graph, the actual data values (red line) are

around 0, while the model posterior distribution is mostly around -1.5 to -0.5, showing the difference between the model and the data for the minimum value. In the Max Y graph, the actual data values are around 1.0, and the model posterior distribution is mostly scattered around 0.8, with the peak of the distribution close to the actual value. For the Range Y graph, the actual data value is around 2.5, while the model posterior distribution is scattered around 2.0 with the peak close to this range. Overall, the model distributions show that the model is quite capable of describing some statistical aspects of the actual data, although there are still some discrepancies in some metrics.

VI. Model Comparison

1. Bayes Factor

```
Iterations = 2001:12000
Thinning interval = 1
Number of chains = 5
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean             SD       Naive SE Time-series SE
     0.0531814      0.0617155      0.0002760      0.0003423

2. Quantiles for each variable:

    2.5%      25%      50%      75%    97.5%
-0.06903  0.01172  0.05301  0.09495  0.17438
```

```
[1] 1.00064
```

The results show that the empirical mean of the parameters is 0.053814 with a standard deviation of 0.0617155. The standard error for the mean (naive SE) is 0.0002760, while the time-series SE is 0.0003423. In the quantile distribution, the value at the 2.5% percentile is -0.06903, the 25% percentile is 0.01172, the 50% percentile (median) is 0.05301, the 75% percentile is 0.09495, and the 97.5% percentile is 0.17438. The result on the right shows a bayes factor value of 1.00064. This value expresses a comparison of how strongly the data supports the full model compared to the null model.

2. DIC

```
Rata-rata DIC untuk Model Logistik: 0.953033
Rata-rata DIC untuk Model Binomial: 0.9530761
```

The average DIC for the logistic model is 0.953033, while the average DIC for the binomial model is 0.9530761.

3. WAIC

```
WAIC: 981.0214

Iterations = 2001:12000
Thinning interval = 1
Number of chains = 5
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

           Mean      SD  Naive SE Time-series SE
beta[1]  0.5924 0.13862 0.0006199      0.0016190
beta[2]  0.8103 0.08375 0.0003745      0.0008104
beta[3] -0.9347 0.08888 0.0003975      0.0008234
beta[4] -1.2434 0.18118 0.0008103      0.0015339

2. Quantiles for each variable:

           2.5%     25%     50%     75%   97.5%
beta[1]  0.3175  0.4995  0.5923  0.6852  0.8648
beta[2]  0.6471  0.7538  0.8093  0.8657  0.9773
beta[3] -1.1131 -0.9938 -0.9329 -0.8742 -0.7649
beta[4] -1.6021 -1.3639 -1.2433 -1.1207 -0.8879
```

The WAIC (Watanabe-Akaike Information Criterion) value was 981.0214. Simulations were conducted with 5 chains, each having 10,000 samples, resulting in total iterations from 2001 to 12000. Parameters β1, β2, β3, and β4 have posterior means of 0.5924, 0.8103, -0.9347, and -1.2434, respectively, with standard deviations of 0.13862, 0.08375, 0.08888, and 0.18118. The posterior 95% confidence intervals (2.5% to 97.5%) for β1 are [0.3175, 0.8648], β2 are [0.6471, 0.9773], β3 are [-1.131, -0.7649], and β4 are [-1.6021, -0.8879].

DISCUSSION

I. Interpretation of the result

1. Summary of Posterior

The picture below shows statistical summary of the posterior:

```
Iterations = 2001:12000
Thinning interval = 1
Number of chains = 5
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

         Mean      SD  Naive SE Time-series SE
beta0  0.5912 0.13641 0.0006101      0.0015709
beta1  0.8103 0.08341 0.0003730      0.0008068
beta2 -0.9333 0.08811 0.0003940      0.0008069
beta3 -1.2433 0.18110 0.0008099      0.0014703

2. Quantiles for each variable:

         2.5%     25%     50%     75%   97.5%
beta0  0.3236  0.4990  0.5910  0.6826  0.8584
beta1  0.6483  0.7540  0.8098  0.8661  0.9754
beta2 -1.1086 -0.9922 -0.9318 -0.8736 -0.7644
beta3 -1.6002 -1.3655 -1.2434 -1.1210 -0.8914
```

Parameter with positive mean value like β1 indicates a positive relationship between the variable with the log odds response variable and otherwise. This means that the increase of variable cp as the β1 will also increase the probability of having heart disease and decrease of variable oldpeak as β2 and exang as β3 will also decrease the probability of having heart disease. Parameter β1 has a smaller standard deviation and credible interval that indicates the variable cp produces better precision in the estimation. This is also supported by the result of the standard error that displays the smallest value among all of the parameters used. Parameter β3 which is exang has the largest standard deviation and credible interval that shows the highest variability and uncertainty in its distribution.

2. Trace and Density Plot

The image on the left is a trace plot, which is used to see if the sampling process using MCMC (Markov Chain Monte Carlo) goes well. From this plot, it can be seen that all parameters (β0, β1, β2, β3) move randomly without a clear up

or down pattern. This indicates that the sampling process has run smoothly, is not stuck at a certain value, and is able to explore all possible parameter values. In other words, the model has reached convergence, so the results can be trusted.

The image on the right is a density plot showing the posterior distribution of each parameter. The distributions look smooth and close to normal, which means that the model has fairly strong confidence about the parameter values. Based on the results, β0 (intercept) has a mean of about 0.59. This indicates that when all other variables are zero, the probability of the target occurring is around that value (on a log-odds scale). For β1, the average value is about 0.81, meaning that the variable x1 has a positive influence on the target chance (the larger x1 is), the larger the chance. In contrast, β2 has an average of about -0.93, which means x2 has a negative influence (the larger x2 is), the smaller the chance of the target occurring. The same goes for β3, with an average of about -1.25, which shows the negative effect of x3 is stronger than x2. Overall, these results give a clear picture of how each variable affects the target. The model has evaluated the data well and produced reliable conclusions.

3. Autocorrelation Plot

The autocorrelation illustrates the relationship between samples over various iterations in the Markov chain. As the iterations increase, the correlation between samples will decrease, indicating that the samples become more independent after a few iterations. The result of the autocorrelation plot shows that the correlation between samples decreases after a few iterations, which means the samples in the Markov chain are quite independent of each other. Since each sample is less impacted by the previous sample, this shows that the sampling

process is effective. The rapid decrease in correlation also suggests that the chain works very well in exploring parameters without getting stuck on any single value. Overall, the samples are independent enough to be analyzed further, which supports the quality of the sampling process in the MCMC model being used.

4. Convergence Diagnostic

Geweke Diagnostic

The results of the Geweke diagnostic for parameters β0, β1, β2, and β3 show varying values for each chain. This diagnostic evaluates whether the distribution of sample values from the beginning and end of the chain are similar, with values approaching zero indicating no significant difference and suggesting convergence. In Chain 1, parameters such as β0 (0.7531) , β1(-0.8539), and β2 (0.4963) have values close to zero, indicating that these parameters have reached convergence. However, β3 (-1.4046) has a value suggesting potential non-convergence.

In Chain 2, parameters β0 (-0.2514), β1 (0.5930), β2 (-0.2860), and β3 (-0.6033) show values close to zero, indicating convergence for this chain. In Chain 3, parameter β3 (1.1662) has values far from zero, indicating that this chain has not fully converged for most parameters. In Chain 4, parameters β0 (0.1777), β1 (0.2879), β2 (-0.6577), and β3 (-0.3497) show values close to zero, indicating convergence for this chain. Chain 5, parameter β3 (-0.3147) have values close to zero, which indicate convergence. However, other parameters, such as β0 (1.9221), β1 (-1.5251), and β2 (-2.1261) show significant values, indicating non-convergence for this chain. Overall, most parameters show signs of progress toward convergence, although some significant values still suggest instability.

Gelman-Rubin Diagnostic

The Gelman-Rubin diagnostic is used to evaluate convergence in the MCMC process. Convergence means that all chains in MCMC are running properly, resulting in a consistent parameter distribution. The Gelman-Rubin diagnostic serves as a tool to ensure that the MCMC results are valid and ready for further analysis. If the potential scale reduction factor (PSRF) is significantly different from 1, it suggests that the sampling process needs some adjustments.

According to the Gelman-Rubin diagnostic results, the point estimate, upper confidence interval (c.i.), and multivariate PSRF values are all equal to 1. This shows that the Markov chain has converged effectively, with all chains operating correctly and yielding consistent parameter values. As a result, the MCMC sampling process has been successful, and the estimated results for each parameter can be considered trustworthy for future analysis.

5. PPC

Min Y	Max Y	Range Y
0.0000	0.0146	1.0000

Min Y of 0.0000 meaning that the minimum value the model predicted was 0. For such data, the model predicts that the event under consideration is really rare or nearly impossible. The best value obtained for Max Y is 0.0146 meaning that the highest value predicted by the model is 0.0146. For this model it indicates that for some data the chance that the event already occurs is quite high, but not too far. The Y Range of 1.0000 is the difference between the maximum and the minimum value predicted by the model o The high temporal variance (1), which means that the model predicts the likelihoods with high variability from close to 0 (0) to slightly higher (0.0146). In general, these findings indicate that

the model has a narrow range of predicting occurrence probability (Y), and predicted values are mostly small.

6. Model Comparison

Bayes Factor

```
Iterations = 2001:12000
Thinning interval = 1
Number of chains = 5
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean             SD       Naive SE Time-series SE
     0.0531814      0.0617155      0.0002760      0.0003423

2. Quantiles for each variable:

    2.5%      25%      50%      75%    97.5%
-0.06903  0.01172  0.05301  0.09495  0.17438
```

```
[1] 1.00064
```

These results show that the mean of the tested parameters is close to zero with a relatively narrow distribution, as evident from the small standard deviation and time-series SE. The 2.5% to 97.5% quantiles, which fall between -0.06903 and 0.17438, indicate that the parameter values can range from negative to positive values. The positive mean and nearly equal median indicate moderate potential evidence in favor of the alternative model, although the range of quantiles indicates considerable uncertainty. The Bayes Factor value of 1.00064 means that the data provides almost equal support for both the full model and the null model. There is no significant difference in the data support for the two models.

DIC

```
Rata-rata DIC untuk Model Logistik: 0.953033
Rata-rata DIC untuk Model Binomial: 0.9530761
```

The results show that the average DIC values for the Logistic Model (0.953033) and Binomial Model (0.9530761) do not have too much difference.

This indicates that both models have almost the same ability to fit the data. Since the difference is very small, the two models can be considered equally good in terms of the balance between complexity and model fit. These results suggest that there is no significantly superior model, so model selection can be based on other needs or preferences.

WAIC

The WAIC value of 981.0214 suggests that this model is suitable for comparing its effectiveness against other models using the same dataset. The posterior means for all parameters show consistent average estimates with low standard deviations, which means there's not much variability. Furthermore, the 95% confidence intervals for each parameter exclude zero, signifying the statistical significance of all parameters. The positive coefficients for β1 and β2 imply a positive association with the response variable, whereas the negative coefficients for β3 and β4 indicate a negative association. Collectively, these findings demonstrate that all parameters significantly contribute to the model and can be interpreted in terms of the nature of their relationships.

II. Comparison with existing literature

According to the Journal of Medical Sciences from the Zainoel Abidin General Hospital, Banda Aceh, chest pain is one of the most common reasons people visit the emergency department. This is because chest pain can be caused by either acute coronary issues or other non-coronary causes. However, most patients with acute coronary problems experience chest pain due to reduced blood flow to the heart. Immediate treatment is necessary to reduce the mortality rate from cardiovascular diseases

The Framingham Study in the United States reported that 1% of men aged 30-62 years without initial symptoms showed signs of coronary heart disease in follow-up examinations, with 38% experiencing stable angina and 7% experiencing unstable angina. According to the National Library of Medicine, a case was observed where a 56-year-old man came to the hospital after experiencing chest tightness during physical activity for two weeks. The symptoms occurred while he was riding his bike in the morning. It was concluded that chest pain during exercise in individuals with heart disease could be caused by vasospastic angina triggered by physical activity. Although rare, this condition can be fatal if not treated properly.

Furthermore, individuals with depression can experience sudden heart failure, which may lead to death from cardiac arrest. This is supported by a Cross-Sectional Study, which previously showed that 19 to 66 percent of patients who had a heart attack also had mental health disorders, particularly depression and anxiety. Some studies report that 17 to 44 percent of patients with coronary heart disease are diagnosed with major depression. Additionally, about 27 percent of patients undergoing heart bypass surgery experience depression after the procedure. Based on the research we have conducted and the literature mentioned above, it is stated that individuals with cardiovascular diseases are likely to exhibit symptoms such as those mentioned above.

III. Implications of the findings

From the research we have conducted, there are several important implications that need to be considered:

1. Chest pain, in general, can be caused by various conditions, ranging from heart disease to non-heart-related issues. However, chest pain caused by heart disease is very dangerous and can threaten the patient's life. Therefore, it is essential to have

knowledge about the potential dangers that chest pain can pose to cardiovascular patients, so chest pain should never be underestimated.

2. Chest pain that occurs after exercise, known as stable angina, is caused by the narrowing of coronary arteries. This indicates that the heart muscle is not receiving enough oxygen during physical activity. Therefore, individuals with heart disease must adjust their exercise routines accordingly.
3. Depression and anxiety are common in individuals with cardiovascular diseases, increasing the risk of sudden heart failure and cardiac arrest. It emphasizes the importance of addressing mental health alongside physical health in cardiovascular patients. The findings also suggest that post operative mental health support is crucial for patients undergoing heart bypass surgery.

CONCLUSION

I. Summary of the study

This study aims to predict the indication of heart disease using patient clinical data with the logistic regression method. The dataset used consists of 1,025 rows of data with 14 columns. Logistic regression was chosen due to its ability to model binary target variables, with a Bernoulli distribution used for the likelihood. The analysis results show that the model parameters have a significant effect on the target probability. The β0 intercept is around 0.59 when all variables are zero. β1 (cp) has a strong positive relationship with the target probability, with an average of 0.81, while β2 (oldpeak) and β3 (exang) show negative relationships of -0.93 and -1.25, respectively. The trace plot and density plot test for good convergence and stable posterior distribution. The autocorrelation plot indicates sampling efficiency with sufficiently independent samples. The Geweke and Gelman-Rubin diagnostics mostly support convergence. The Posterior Predictive Check shows a narrow prediction

range, with overall low probabilities. Model comparison through Bayes Factor, DIC, and WAIC indicates similar performance between the full and null models, with stable and significant results. From the series of studies we have conducted, we conclude that individuals with heart disease can be identified based on the three variables mentioned exang, oldpeak, and cp.

II. Limitations

One of the main limitations of this study is the relatively small dataset. Some parameters, such as β2 and β3, show indications of instability. This can reduce the interpretation of the model's results, and therefore, further observation is needed to ensure the model reaches convergence. This study also lacks sufficient resources as references, making it difficult to understand some of the variables involved.

III. Suggestion for future research

For future research, it is recommended to use a larger and more diverse dataset in order to analyze more factors that cause heart disease so that the results can be more accurate and specific. Additional factors that could be considered include family history, diet, physical activity level and stress. In addition, future research can also consider geographical factors to broaden the context and increase the precision of the analysis results. Comparing different Bayesian models is also recommended to find the most accurate model that captures more complex relationships between risk factors and heart disease.

REFERENCES

Humas UB. (2023). *World Heart Day 2023: Use Heart Know Heart» Prasetya UB*. Ub.ac.id. https://prasetya.ub.ac.id/world-heart-day-2023-use-heart-know-heart/

Kementrian kesehatan. (2023, September 25). *Cegah Penyakit Jantung dengan Menerapkan Perilaku CERDIK dan PATUH*. Sehat Negeriku. https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20230925/4943963/cegah-penyakit-jantung-dengan-menerapkan-perilaku-cerdik-dan-patuh/

Khawaja, I. S., Westermeyer, J. J., Prashant Gajwani, & Feinstein, R. E. (2009). *Depression and Coronary Artery Disease: The Association, Mechanisms, and Therapeutic Implications*. Psychiatry (Edgmont), 6(1), 38. https://pmc.ncbi.nlm.nih.gov/articles/PMC2719442/

Pohan, T. A. (2024, October 31). *RS Pondok Indah*. RS Pondok Indah Group. https://www.rspondokindah.co.id/id/news/mengenal-penyakit-jantung-koroner

Ridwan, M., Yusni, & Nurkhalis. (2020). *Analisis Karakteristik Nyeri Dada pada Pasien Sindroma Koroner Akut di Rumah Sakit Umum Daerah dr. Zainoel Abidin Banda Aceh*. *Journal of Medical Science , 1*(1), 20–26. https://rsudza.acehprov.go.id/publikasi/index.php/JMS/article/view/5

Tamura, A., Nagao, K., Inada, T., & Tanaka, M. (2018). *Exercise-induced vasospastic angina with prominent ST elevation: a case report.* European Heart Journal - Case Reports, 2(4). https://doi.org/10.1093/ehjcr/yty141