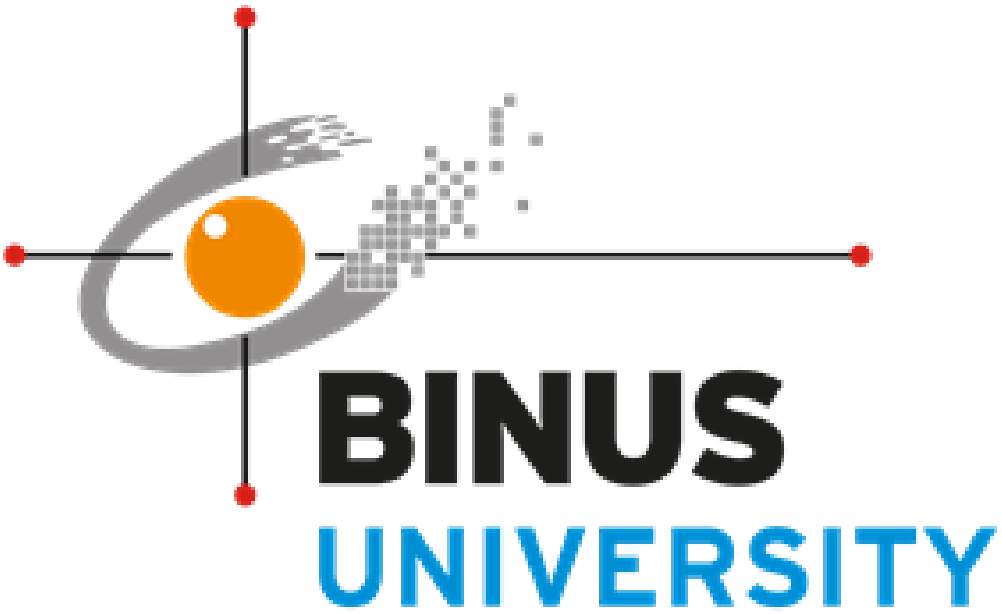
BAYESIAN DATA ANALYSIS FINAL PROJECT REPORT



Aldo Oktavianus - 2702234081
Arieldhipta Tarliman - 2702234636
William - 2702225373

I. INTRODUCTION

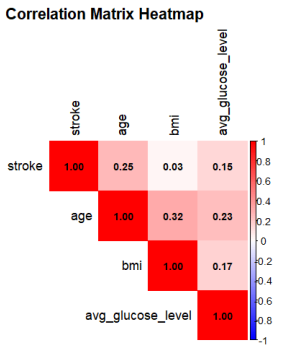
Stroke is a medical condition characterized by a sudden disruption of blood flow to the brain, leading to cell damage. It is a leading cause of death and disability worldwide, with various risk factors such as hypertension, heart disease, age, and lifestyle choices. Early identification of stroke risk factors is crucial for prevention and improving patient outcomes.

This dataset contains information on 5,110 patients, with 12 variables for each. These variables include demographics (age, gender, marital status, residence type), medical history (hypertension, heart disease, smoking status), occupation, and health indicators (glucose level, BMI). The primary goal is to predict stroke risk based on these factors using Bayesian methods.

II. MODELS

The probability distribution of Y follows a Bernoulli distribution Y~Bernoulli(θi), with θ representing the true probability of the individual's stroke status within the interval [0, 1]. Handling Missing Values in continuous variables were imputed with their respective means to retain sample size and avoid biases. Features labeled as "others" or "unknown" were not treated as missing values but included as valid categories. Label encoding was applied to categorical variables instead of one-hot encoding to mitigate the high dimensionality issues that arise from large datasets.

First, our feature selection starts with a correlation matrix heatmap:



Although the correlation coefficients for the numerical variable are weak, it doesn't necessarily imply no relationship. So we decided to make a grouping for those columns. For age, we can use the 10 years interval. For bmi, we can use standard BMI categories: Underweight: <18.5, Normal: 18.5–24.9 Overweight: 25–29.9, Obesity: 30–34.9, Severe Obesity: 35–39.9, Morbid Obesity: 40+. For avg_glucose_level, we can define clinically meaningful ranges: Low: 0–70, Normal: 70–99, Prediabetic: 100–125, Diabetic: 126+.

Feature selection with a chi-square test was used to evaluate the statistical significance of each covariate. Features with high p-values were excluded to focus on predictors with a significant relationship to the stroke probability. The goal was to

identify covariates with the most significant impact on stroke prediction while maintaining computational efficiency.

Variable	Chi-squared Statistic	Degrees of Freedom	p-value	Importance (if p < 0.05)
gender	25.629	1	6.127	Not Significant
hypertension	55.701	1	8.438×10^{-14}	Significant
heart_disease	89.449	1	$< 2.2 \times 10^{-16}$	Highly Significant
ever_married	43.722	1	3.785×10^{-11}	Significant
work_type	38.386	4	9.329×10^{-8}	Significant (with caution due to multiple degrees of freedom)
Residence_type	1.3882	1	2.387	Not Significant
smoking_status	23.615	3	3.005×10^{-5}	Significant

The logistic regression model was specified using the chosen features. Below is the mathematical representation of the model:

$$logit(p_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$$

The selected text explains a decision made during feature selection in the analysis. It states that gender and residence type were excluded as features, likely because they were not deemed significant or relevant in predicting stroke probability. However, the feature "work type" was retained, as previous studies and cases indicated that different types of occupations are associated with varying stroke risk levels.

We also do a SMOTE to our dataset because of the imbalance of our target value(‘stroke’) that before SMOTE the proportion of stroke and not stroke are (5% and 95%) meanwhile after SMOTE it became 76% and 24%. Thus, the data become more balanced allowing the model to better detect for the minority class.

Our model differences are based on the feature selection, so to better understand the initial values of each model we use Maximum Likelihood Estimation (MLE). MLE is a fundamental statistical method used to estimate the parameters of a probability distribution or statistical model. The choice of initial values can influence whether the algorithm converges to a global maximum of the likelihood function or gets trapped in a local maximum.

Both of our model algorithms are the same, we use a Bayesian logistic regression model using JAGS(Just Another Gibbs Sampler) to analyze binary outcome data. In Bayesian logistic regression, the **likelihood function** quantifies the probability of observing the given data under specific parameter values. For each observation, the outcome Yi is modeled as a Bernoulli random variable with a success probability πi.

The **prior distributions** in Bayesian analysis encapsulate our beliefs about the parameters before observing the data. In the provided model, both the intercept α\alphaα and the regression coefficients βjare assigned normal priors with mean 0 and precision 0.01. These priors are considered weakly informative, suggesting that, prior to data observation, we have no strong preference for any particular parameter values. The choice of a normal distribution with a small precision (large variance) reflects a belief that the parameters are likely to be near zero but allows for a wide range of possible values. This approach prevents the priors from unduly influencing the posterior estimates, allowing the data to play a significant role in shaping the posterior distributions.

III. RESULTS

Four independent Markov chains were initialized with distinct starting values to ensure diverse exploration of the parameter space. Each chain was run for 8,000 iterations, with the first 1,000 iterations discarded as a burn-in period to allow the sampler to reach the target distribution. Additionally, step sizes were tuned during the warm-up phase to improve sampling efficiency and ensure robust posterior estimates.

The diagnostic convergence was assessed using the Geweke diagnostic and Gelman-Rubin diagnostics. These methods compared the means of the early and late sections of each chain. For the Model 1, the first and fourth chains showed good convergence, with most parameters having z-scores close to 0, indicating no significant differences between the early and late segments of the chains. However, the second and third chains exhibited non-convergence for specific parameters, such as β[2], β[9], and β[10]. This suggests potential initialization or sampling issues. The Gelman-Rubin diagnostic showed that all parameters achieved a Potential Scale Reduction Factor (PSRF) of 1.00, suggesting excellent overall convergence across chains. Discrepancies between the Geweke results and Gelman-Rubin diagnostics imply that while certain individual chains (e.g., Chain 2) faced issues, the overall convergence was satisfactory. The average Effective Sample Size (ESS) was greater than 1,000, reflecting good mixing and sufficient sampling for robust posterior inference.

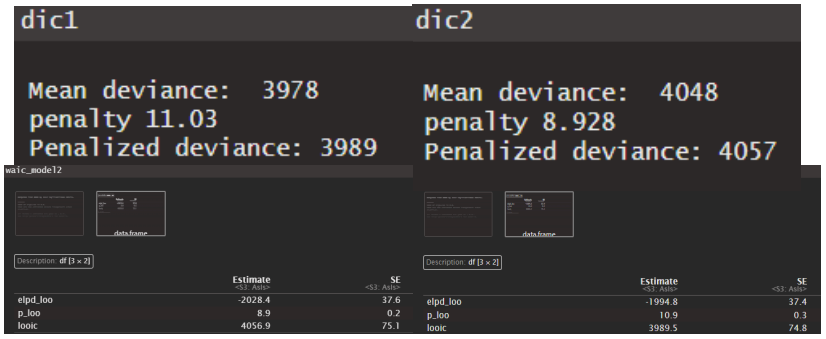
For the Model 2, the Geweke diagnostic showed that the first and third chains demonstrated reasonable convergence, but the second and fourth chains displayed poor convergence for parameters such as α and β[8]. These issues may be linked to

poor initialization or insufficient mixing, warranting further investigation and potential re-sampling. The Gelman-Rubin diagnostic indicated that most parameters had PSRF values close to 1.00, with slight deviations for α and β[6]. These results suggest general convergence but highlight localized issues for specific parameters. Additionally, the average ESS for most parameters was well below 1,000, indicating that the MCMC chains had a high degree of autocorrelation, resulting in fewer effectively independent samples.

A. Convergence Diagnostics and two information criteria

Model 1 demonstrated superior convergence and predictive performance compared to Model 2. Both Geweke and Gelman-Rubin diagnostics confirmed Model 1's stability, with key covariates like β[5] and β[6] showing robust posterior estimates. In contrast, while Model 2 passed the Gelman-Rubin diagnostic, Geweke identified issues in Chains 2 and 4, suggesting potential biases in posterior estimates for parameters such as alpha and β[8]. Significant predictors in Model 1, including age, blood pressure, and lifestyle factors, highlighted its reliability in stroke prediction.

Model comparison using DIC and WAIC consistently favored Model 1. Model 1 had a lower penalized deviance (3989 vs. 4057) and mean deviance (3978 vs. 4048) than Model 2, indicating better fit despite slightly higher complexity. WAIC results reinforced this, with Model 1 achieving a higher elpd_loo (-1994.8 vs. -2028.4) and lower loo-ic (3989.5 vs. 4056.9). Although Model 2 was simpler (p_loo = 8.9), this came at the cost of reduced predictive accuracy.



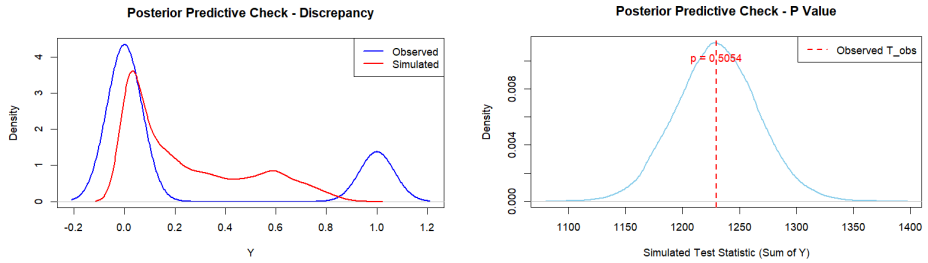
Overall, Model 1's superior fit, convergence, and predictive reliability make it the preferred model for balancing complexity and accuracy in this analysis.

B. Posterior Predictive Checks

The posterior predictive checks indicate that the Bayesian logistic regression model generally fits the data well, with a Bayesian p-value of **0.5054**, suggesting the observed data aligns closely with the simulated predictions. This reflects strong performance in predicting the majority class (Y=0). However, the model underperforms for the minority class (Y=1), as

evidenced by the underestimation of Y=1 in the simulated data. Although SMOTE was applied, the remaining class imbalance (3882:1230 post-SMOTE) and overlapping feature distributions limit the model's ability to accurately predict strokes.

The mean discrepancy of **0.1827** further emphasizes this issue, indicating that while the model captures Y=0 patterns well, it struggles to confidently predict Y=1. The broader spread in simulated predictions reflects increased uncertainty for the minority class.While the overall fit is reasonable, improving sensitivity for the minority class may require enhanced resampling, feature engineering, or adjustments to the model architecture.



IV. CONCLUSION

Model 1 outperformed Model 2 in convergence and predictive reliability, with the Gelman-Rubin diagnostic confirming convergence, though Geweke's diagnostic flagged issues in Model 2. Key predictors in Model 1—age, blood pressure, and lifestyle factors—proved essential for stroke prediction. Despite better convergence, Model 1 underestimates the minority class (Y=1). The Bayesian p-value of 0.5054 suggests overall model fit, but the mean discrepancy of 0.1827 highlights misalignment in minority class predictions. Even after SMOTE, residual class imbalance (3882:1230) and feature overlap limit predictive accuracy. To enhance performance, further feature engineering, stronger priors, or alternative models are needed. Addressing initialization, tuning, and extending sampling periods may also improve sensitivity to the minority class.