**Final Project Bayesian Data Analysis**

**“Predicting Diabetes Case Based on Medical Diagnosis Measurement”**

***Calista Lianardi - 2702325880 || Data Science LF09***

1. **INTRODUCTION**

Bayesian data analysis is a robust statistical approach that combines prior knowledge with observed data to estimate the underlying probability distribution of unknown parameters. Unlike traditional frequentist methods, which focus solely on observed data, the Bayesian framework uses Bayes' theorem to update prior beliefs into posterior distributions as new evidence is incorporated. This allows for a more flexible and probabilistic interpretation of model parameters and predictions.

In this assignment, Bayesian linear regression is employed to **predict the likelihood of diabetes occurrence based on individual health metrics.** Using the **Pima Indians Diabetes dataset**, a widely used benchmark in medical research, this analysis models the binary outcome (presence or absence of diabetes) using predictors such as glucose levels, blood pressure, and body mass index. **The likelihood function is Bernoulli, as the target variable is binary,** and **Gaussian priors** are assigned to the regression coefficients to express initial beliefs about their values.

The goal of this analysis is threefold:

1. To implement Bayesian linear regression using Rjags and interpret the posterior distributions of the model parameters.
2. To evaluate the model's performance using a hold-out test set.
3. To explore the practical benefits of Bayesian inference in capturing parameter uncertainty and enhancing prediction reliability.

This assignment not only demonstrates the applicability of Bayesian statistics in medical decision-making but also highlights its capability to account for uncertainty, making it a powerful tool for modeling and inference in real-world scenarios.

1. **DATASET**

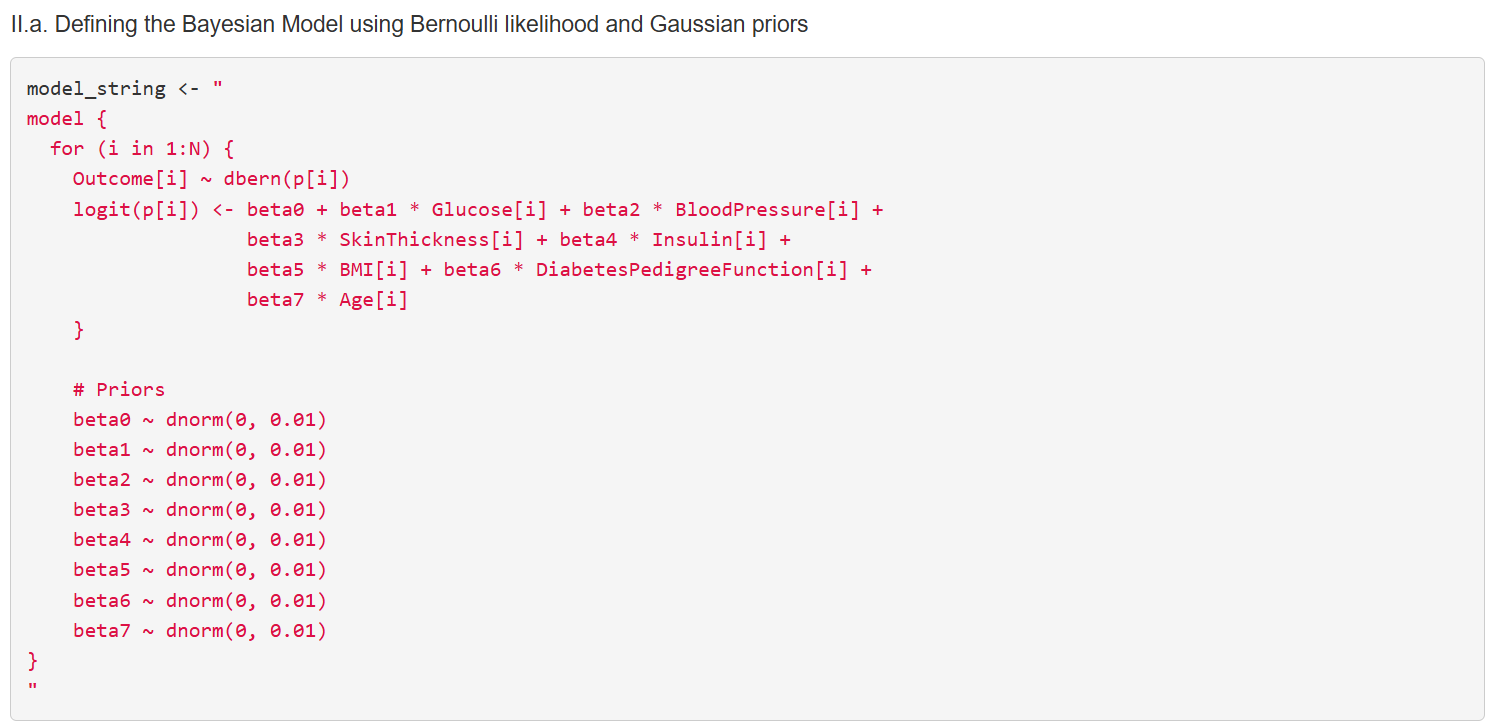
The dataset used in this analysis is the **Pima Indians Diabetes dataset**, a well-known dataset in the medical domain for predicting the presence of diabetes based on various health metrics. It consists of **768 observations with 8 predictor variables and one target variable**.

The Variables are such:

1. **Pregnancies**: Number of pregnancies (int).
2. **Glucose**: Plasma glucose concentration during an oral glucose tolerance test (int).
3. **BloodPressure**: Diastolic blood pressure in mm Hg (int).
4. **SkinThickness**: Triceps skinfold thickness in mm (int).
5. **Insulin**: 2-Hour serum insulin level in μU/ml (int).
6. **BMI**: Body Mass Index, calculated as weight (kg) / height (m²) (float).
7. **DiabetesPedigreeFunction**: A function that scores the likelihood of diabetes based on family history (float).
8. **Age**: Age of the individual (int).
9. **Outcome**: The target variable, where 0 indicates no diabetes and 1 indicates diabetes (binary)

The preprocessing done to this dataset to ensure clean and neat data for modelling are:

1. **Handling Missing Values**: The dataset contains some biologically impossible zero values for variables like **Glucose**, **BloodPressure**, **SkinThickness**, **Insulin**, and **BMI**. These zeros were treated as missing values. Imputation was performed using median values to maintain the integrity of the data. Median is less sensitive to extreme values compared to mean, and typically aligns better with the central tendency of the data.
2. **Normalization**: Continuous predictors (e.g., **Glucose**, **BloodPressure**, **BMI**, etc.) were normalized to scale their values between 0 and 1. This step ensures better convergence during Bayesian inference and prevents numerical instability in the model.
3. **Data Splitting**: The dataset was divided into a training set (80%) and a test set (20%) to train the Bayesian model and evaluate its predictive performance on unseen data.
4. **BAYESIAN MODELING**

****

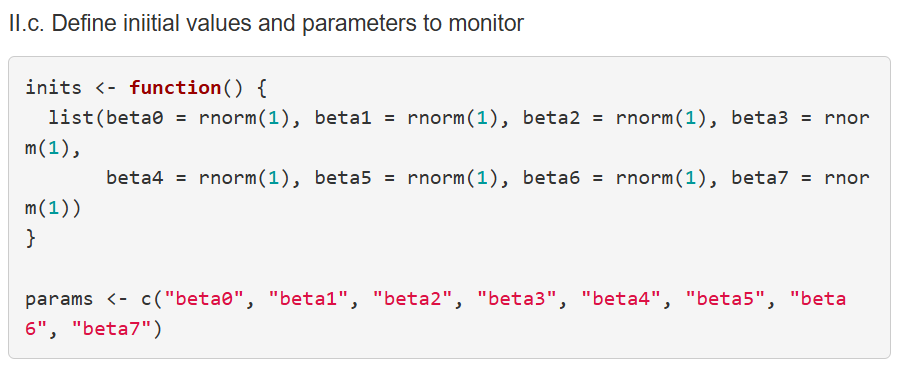
As mentioned above, Outcome data are binary, making the distribution of the dependent variable Outcome[i] follow a Bernoulli Distribution with success probability of p[i], with dbern as the Bernoulli Distribution Function. The Logit transformation logit(p)=log(p/1−p​) converts probabilities into a linear combination of predictors, thus modeling the **Bayesian Linear Regression**. beta0 as intercept and beta1 to beta7 as coefficients describe how each predictor affects the log odds of Outcome[i] being 1.

As for the **Prior,** Prior specify the initial beliefs about the values of the coefficients beta0 to beta7. Each coefficient is assigned a  **normal distribution prior** using the **dnorm** function with parameter **mean = 0 and Precision = 0.01.** The parameter being **mean** = 0 as it is expected to be centered around 0, which expresses the prior belief that the coefficient could be either positive or negative, with most of its probability mass near 0. Low **precision** equivalent to high variance, allowing wide uncertainty about the values



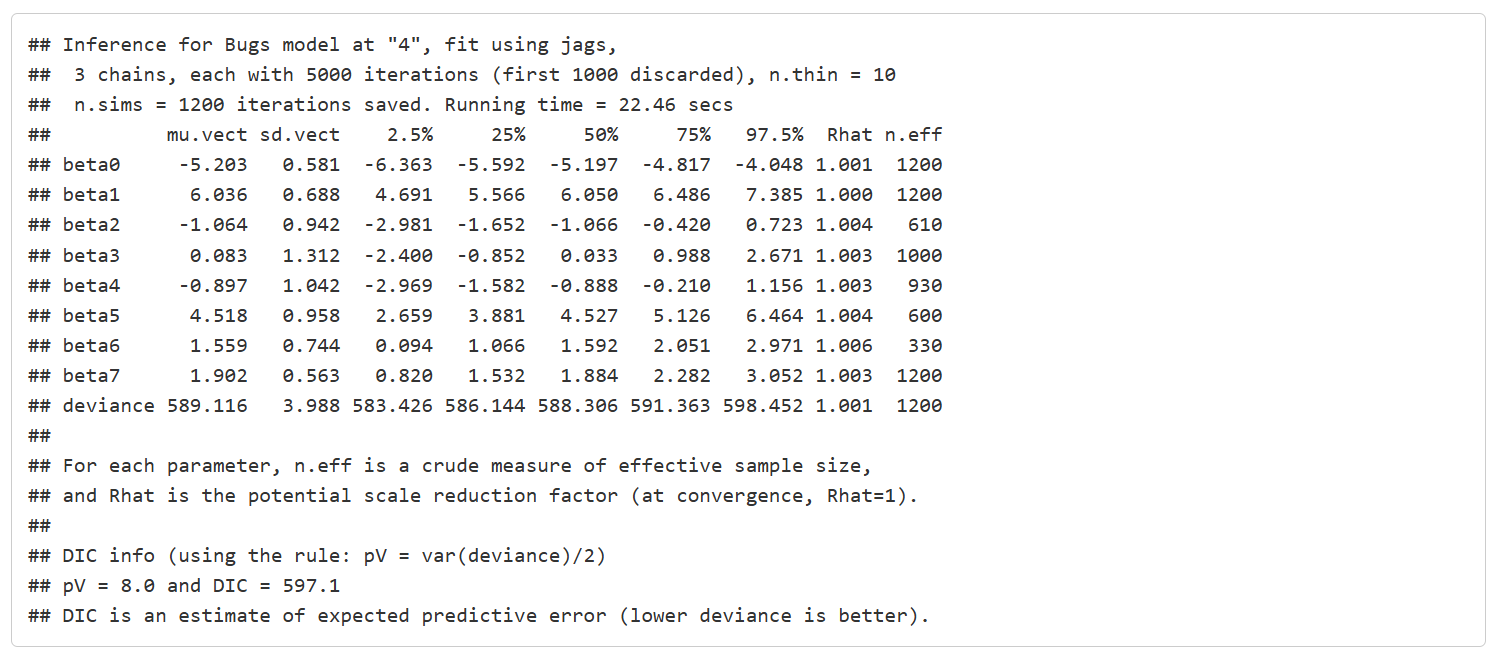
Data preparation for JAGS is a critical steps because JAGS requires data to be in specific format for Bayesian Analysis, **JAGS expects the data to be passed as a list** where each component corresponds to a variable used in the model.

The response variable (Outcome) and the predictor variables (features) must be clearly separated.



The initial value are needed as it will serve as the starting point for the MCMC algorithm for each parameter to begin sampling; the initial value can help the chains converge faster to the posterior distribution. **Rnorm** function is used as it generates random values from a **normal distribution** which is often used for initial values. The param variable are to store tha parameter to be monitored

1. **OUTPUT**

****

As each parameter in the model is estimated with a posterior distribution, giving mu.vect and sd.vect as mean and standard deviation, and other credible quantile for each parameter.

* **beta0 (Intercept)**:
  + Mean = -5.220, indicating a negative intercept, which suggests that when all other variables are zero, the log-odds of the outcome (diabetes) is negative.
  + The credible interval (-6.351, -4.170) suggests this negative relationship is likely true with high probability.
* **beta1 (Glucose)**:
  + Mean = 6.052, which suggests a positive relationship between glucose levels and the likelihood of diabetes **(higher glucose levels increase the likelihood of diabetes).**
  + The credible interval (4.779, 7.324) is positive and suggests that **glucose is a significant predictor.**
* **beta2 (Blood Pressure)**:
  + Mean = -1.056, indicating a negative relationship between blood pressure and the likelihood of diabetes (although this effect is weaker).
  + The credible interval (-2.989, 0.778) includes zero, indicating that blood pressure may not be a strong predictor in this model.
* **beta3 (Skin Thickness)**:
  + Mean = 0.156, suggesting a weak positive relationship between skin thickness and the likelihood of diabetes.
  + The credible interval (-2.377, 2.746) is quite wide, indicating high uncertainty about this predictor’s effect.
* **beta4 (Insulin)**:
  + Mean = -0.892, indicating a weak negative relationship with diabetes.
  + The credible interval (-2.920, 1.081) includes zero, suggesting this predictor may not have a significant effect.
* **beta5 (BMI)**:
  + Mean = 4.511, indicating a positive relationship between BMI and the likelihood of diabetes **(higher BMI increases likelihood)**.
  + The credible interval (2.656, 6.355) **is positive and supports this relationship.**
* **beta6 (Diabetes Pedigree Function)**:
  + Mean = 1.527, suggesting a positive relationship between this variable and the likelihood of diabetes.
  + The credible interval (0.037, 2.954) includes zero, but the mean is quite high, indicating a **significant influence**.
* **beta7 (Age)**:
  + Mean = 1.853, indicating a positive relationship with diabetes likelihood **(older age increases the likelihood).**
  + The credible interval (0.755, 2.963) supports this **positive relationship.**

The Deviance being 589.023, measures the goodness-of-fit of the model. A lower deviance indicates a better fit to the data. The range of the deviance (583.407 to 599.259) gives an idea of the **uncertainty in the model's fit**.

1. **MODEL EVALUATION**

The model shows that **Glucose, BMI, Diabetes Pedigree Function, and Age** are the strongest predictors of diabetes in this dataset, with positive relationships. The model has converged well, as indicated by Rhat values close to 1 and high effective sample sizes. The Deviance and DIC values suggest the model fits reasonably well but can be improved with further model tuning or by incorporating other predictors.