Data Set and Problem Statements

Data set Link: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

Problem Statements:

1. Bayesian Logistic Regression for Diabetes Prediction

Problem Statement:
Predict the probability of diabetes in patients using Bayesian Logistic Regression to estimate model parameters.

Goal:

- Use Bayesian Logistic Regression to predict the binary outcome of diabetes.
- Interpret the posterior distributions of model parameters to quantify uncertainty around the predictions.

Approach:

- Use prior distributions for model parameters (e.g., Normal priors for regression coefficients).
- Perform MCMC (Markov Chain Monte Carlo) sampling or variational inference to obtain posterior estimates.
- Compare the Bayesian model's predictive performance with traditional logistic regression.

Outcome:
Posterior probabilities for each prediction, with credible intervals to express uncertainty.

---

2. Bayesian Inference for Feature Importance

Problem Statement:
Identify the most influential predictors for diabetes using Bayesian analysis.

Goal:

- Use a Bayesian regression framework to estimate posterior distributions of coefficients for each feature.
- Assess which features (e.g., BMI, Glucose, Age) have the most significant impact on diabetes, based on credible intervals.

Approach:

- Set weakly informative priors for the coefficients (e.g., Normal priors centered at 0).
- Analyze the posterior distributions of the coefficients:
    - If the 95% credible interval for a coefficient does not include 0, the feature is likely influential.

Outcome:
A probabilistic ranking of features, showing their relative importance in predicting diabetes.

---

3. Bayesian Risk Estimation for Diabetes

Problem Statement:
Estimate the probability of diabetes for individual patients using Bayesian methods.

Goal:

- Use Bayesian techniques to provide individualized risk scores for diabetes with associated uncertainty.

Approach:

- Fit a Bayesian Logistic Regression model.
- For a new patient with specific features (e.g., Glucose = 140, BMI = 28), compute the posterior predictive probability of diabetes.
- Provide credible intervals for the prediction to quantify uncertainty.

Outcome:
Personalized risk scores with probabilistic confidence intervals.

---

4. Bayesian Model Comparison for Diabetes Prediction

Problem Statement:
Compare different predictive models for diabetes using Bayesian model selection techniques.

Goal:

- Use Bayesian methods (e.g., Bayes Factors, Deviance Information Criterion - DIC, or WAIC) to compare:
    - Bayesian Logistic Regression
    - Bayesian Linear Regression (as an approximation)
    - Hierarchical Bayesian models (if grouped data exists)

Approach:

- Fit multiple models and compute their posterior probabilities or Bayes Factors to determine the most suitable model for diabetes prediction.

Outcome:
Ranked models based on Bayesian criteria with posterior predictive checks for model validation.

---

5. Hierarchical Bayesian Analysis for Grouped Data

Problem Statement:
Analyze the diabetes risk across different subgroups (e.g., age groups or BMI levels) using a Hierarchical Bayesian Model.

Goal:

- Estimate group-level effects (e.g., age, BMI) and individual-level variations in diabetes risk.
- Account for hierarchical structures in the data.

Approach:

- Set up a Hierarchical Bayesian model with group-specific priors:
    - Level 1: Individual predictors (e.g., Glucose, Insulin, BMI).
    - Level 2: Group-specific effects (e.g., Age groups: 20-30, 30-40, etc.).
- Use MCMC sampling to estimate posterior distributions of group effects and individual risk probabilities.

Outcome:
Group-specific diabetes risk estimates with credible intervals.
