Machine Learning Algorithm Random Forest and Bayesian Inference Analyze Biological Dataset

Random Forest is an ensemble learning algorithm that combines the predictions of multiple decision trees to improve accuracy and robustness. Random Forest can identify important features in biological datasets, helping to prioritize factors that contribute to specific outcomes.

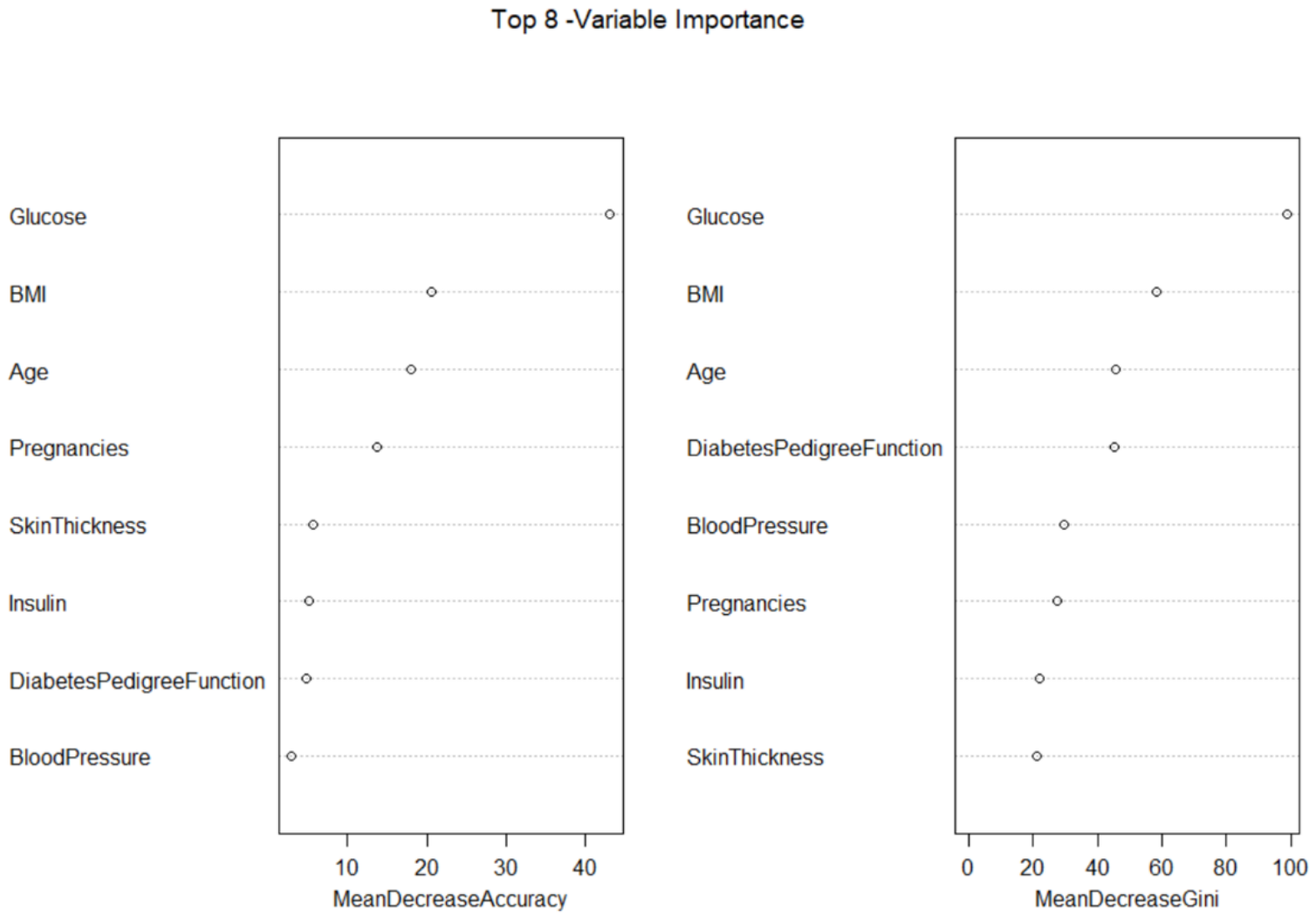
Bayesian inference is a statistical approach that involves updating beliefs about parameters based on both prior knowledge and observed data and calculates the posterior distribution of parameters, which represents the updated knowledge after considering new data.

Bayesian inference is used for estimating parameters in biological models, considering uncertainties and prior knowledge and help provides a formal framework for hypothesis testing in biological experiments.

This study analyzed a diabetic dataset including 9 variables: diabetic outcome, pregnancies, BMI, age, glucose etc.

Random Forest Picks Up Important Features/Targets:

Glucose, BMI, Age, DiabetesPedigreeFunction/Pregnancies



Dataframe (768x9), Total 768 samples x 9 variables

```
> str(data)
spc_tbl_ [768 × 9] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Pregnancies             : num [1:768] 6 1 8 1 0 5 3 10 2 8 ...
 $ Glucose                 : num [1:768] 148 85 183 89 137 116 78 115 197 125 ...
 $ BloodPressure           : num [1:768] 72 66 64 66 40 74 50 0 70 96 ...
 $ SkinThickness           : num [1:768] 35 29 0 23 35 0 32 0 45 0 ...
 $ Insulin                 : num [1:768] 0 0 0 94 168 0 88 0 543 0 ...
 $ BMI                     : num [1:768] 33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
 $ DiabetesPedigreeFunction: num [1:768] 0.627 0.351 0.672 0.167 2.288 ...
 $ Age                     : num [1:768] 50 31 32 21 33 30 26 29 53 54 ...
 $ Outcome                 : num [1:768] 1 0 1 0 1 0 1 0 1 1 ...
```

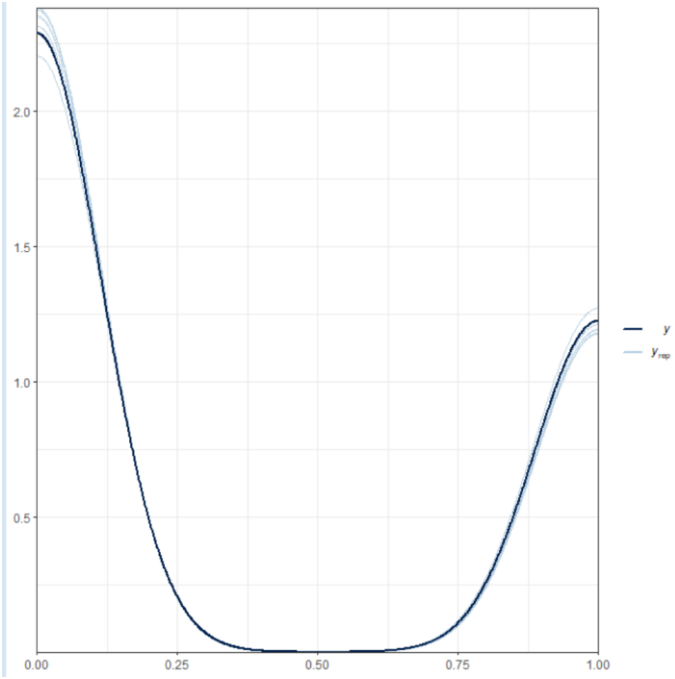
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31.0	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0.0	0.232	54	1

9 variables includes

Pregnancies,
Glucose,
BloodPressure,
SkinThickness,
insulin,
BMI,
DiabetesPedigreeFunciton,
Age,
Outcome

Bayesian Inference Model Established And Verified ('brms_m1')

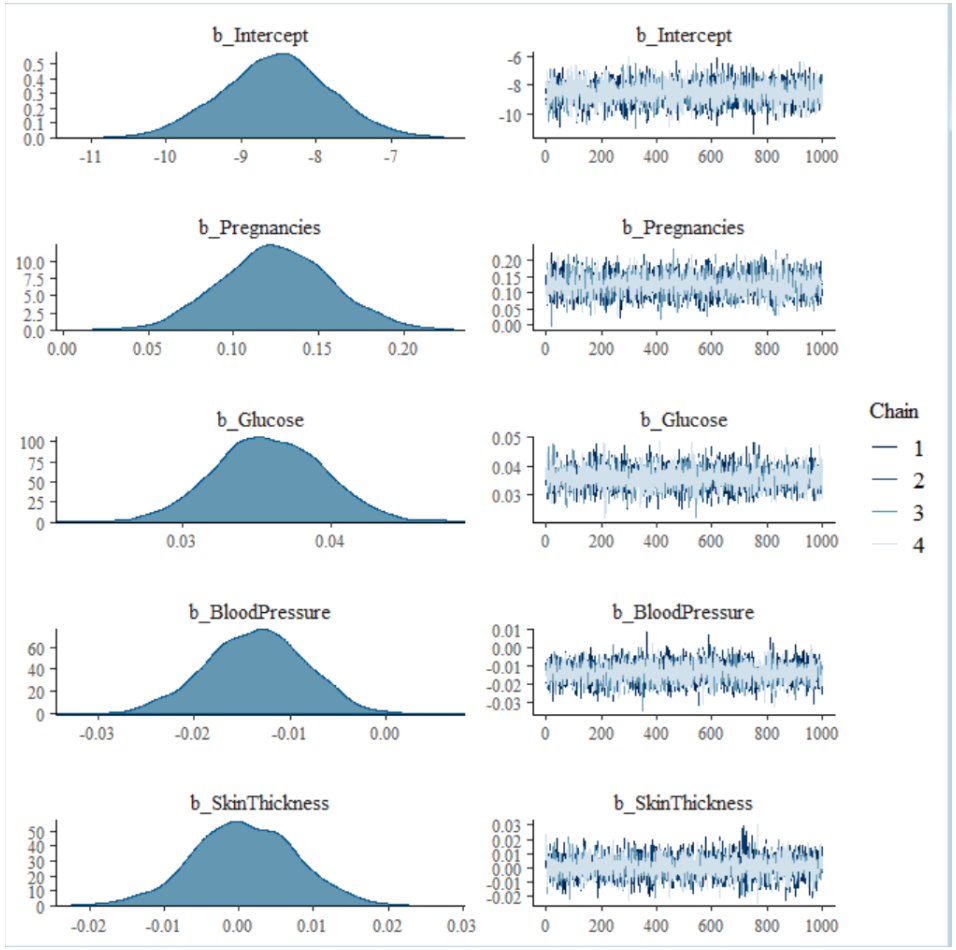
```
> # Display the summary of the model
> summary(brms_m1)
 Family: bernoulli
  Links: mu = logit
Formula: Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness + Insulin + BMI + DiabetesPedigreeF$
   Data: diabetes (Number of observations: 768)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Population-Level Effects:
                         Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept                   -8.53      0.73    -9.97    -7.11 1.00     4958     3331
Pregnancies                  0.13      0.03     0.06     0.19 1.00     3683     3071
Glucose                      0.04      0.00     0.03     0.04 1.00     5421     2827
BloodPressure               -0.01      0.01    -0.02    -0.00 1.00     5248     3106
SkinThickness                0.00      0.01    -0.01     0.01 1.00     4425     3076
Insulin                     -0.00      0.00    -0.00     0.00 1.00     4381     3560
BMI                          0.09      0.02     0.06     0.12 1.00     4126     2995
DiabetesPedigreeFunction     0.97      0.30     0.40     1.56 1.00     3800     3104
Age                          0.01      0.01    -0.00     0.03 1.00     3816     3019

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

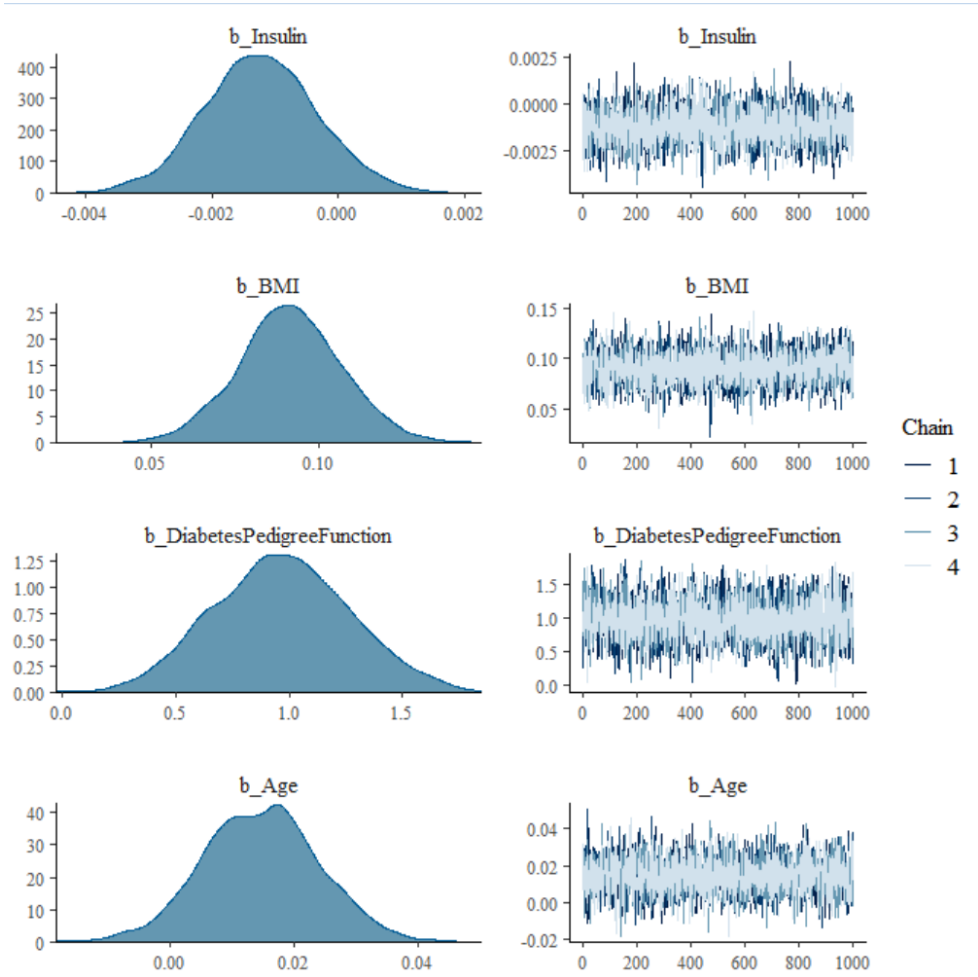


Rhat=1, Bulk_ESS and Tail_ESS are bigger, the model is good

Model diagnostics: fit

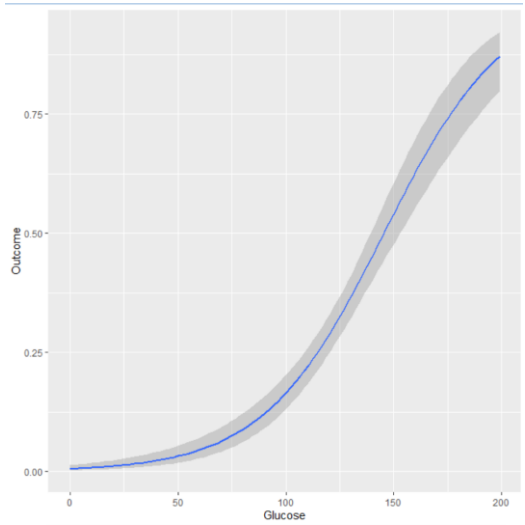
Posterior Probability Distribution of Coefficients for Each Variable

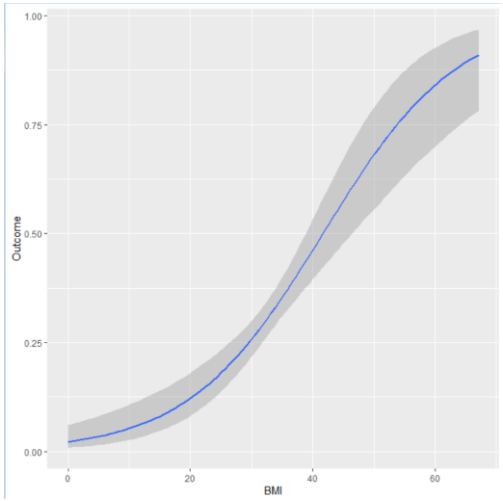


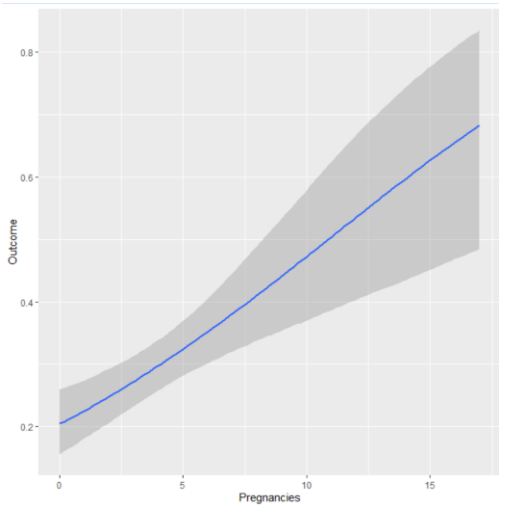


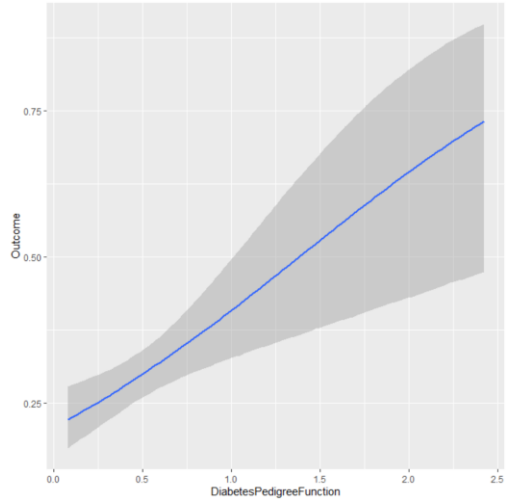
The variables positively correlate with diabetic outcome : glucose, BMI, pregnancies and diabetes pedigree function and age

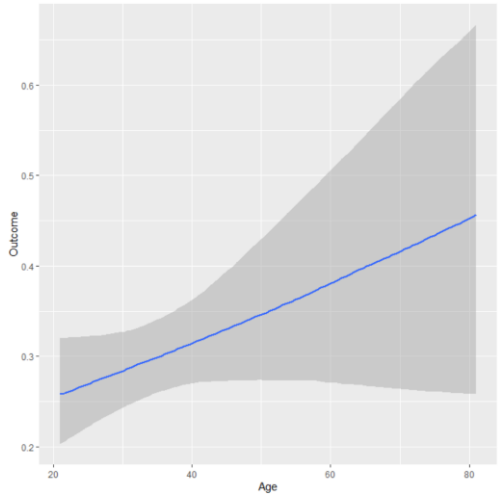
Variables Positively Correlate With Diabetic Outcome From Raw Data







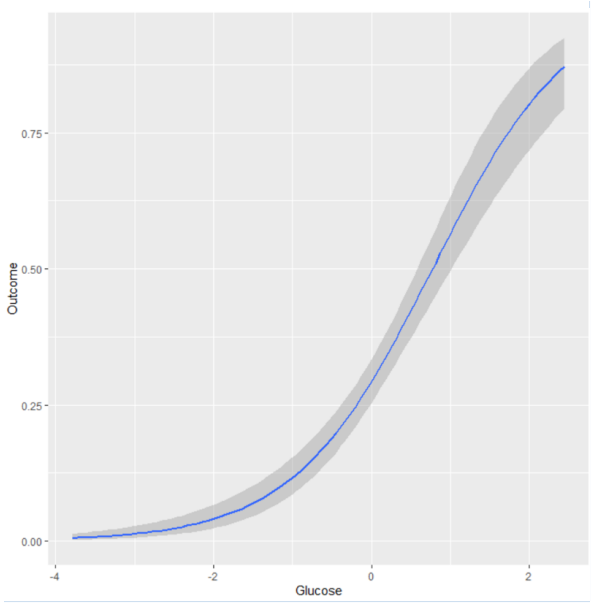


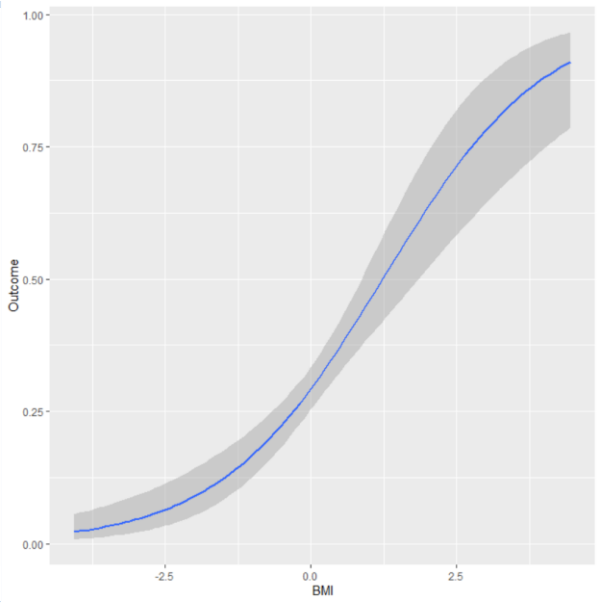


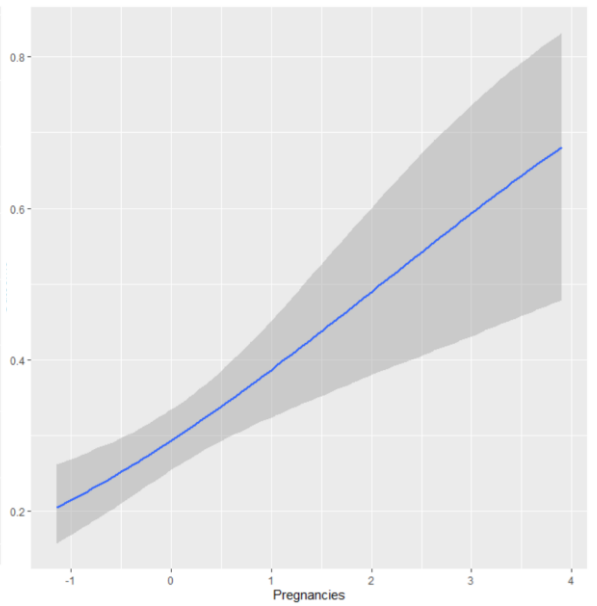
Variables as

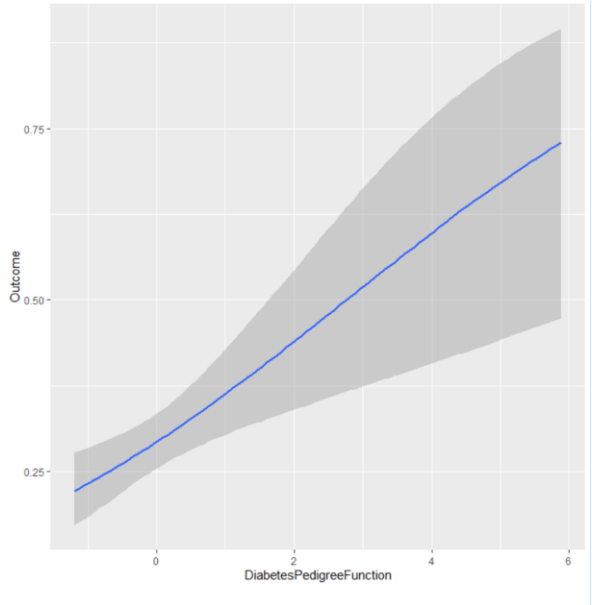
Glucose,
BMI
pregnancies
DPF(DiabetesPedegreeFunciton)
Age

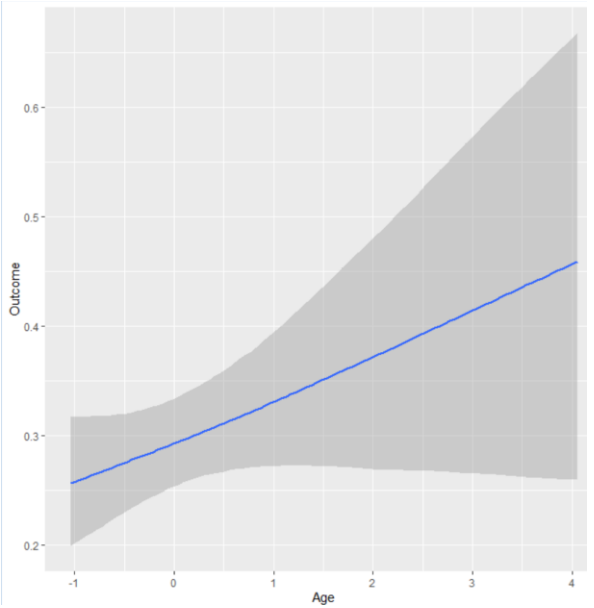
Ranking Each Variable Contributing To Diabetic Outcome











The data was standardized for raw data, the contribution of diabetic outcome was for each variable based on each unit change causing change of the outcome.

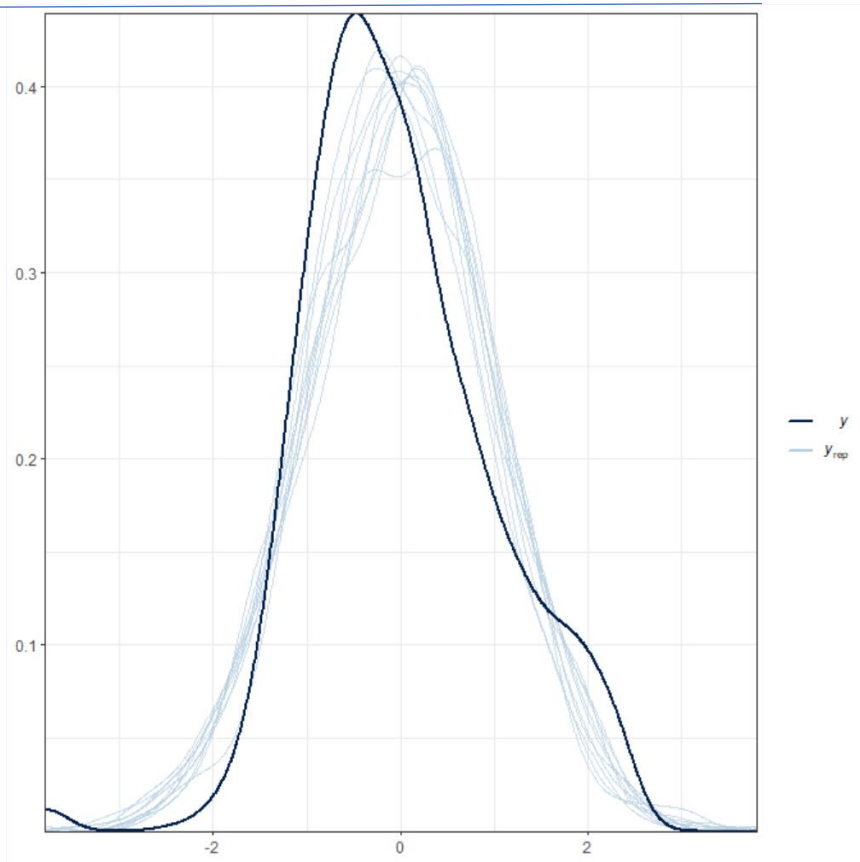
Glucose 0.26/unit,
BMI 0.15/unit,
pregnancies 0.10/unit,
DPF 0.07/unit,
Age 0.04/unit

Results show that

BMI >Pregancies >DPF >Age

Bayesian Inference Model ‘brms_m2’

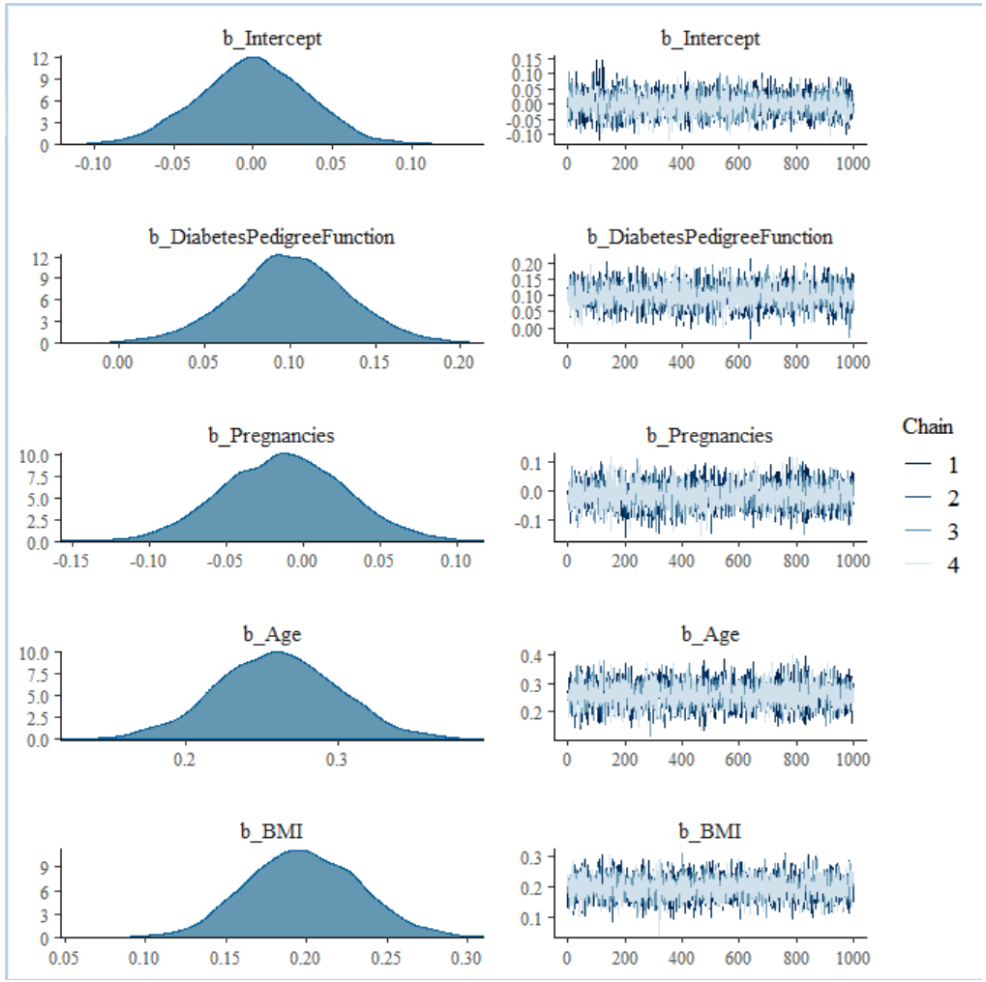
```
> summary(brms_m2)
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: Glucose ~ DiabetesPedigreeFunction + Pregnancies + Age + BMI
   Data: diabetes_s (Number of observations: 768)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Population-Level Effects:
                         Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept                    0.00      0.03    -0.07     0.07 1.00     5442     2812
DiabetesPedigreeFunction     0.10      0.03     0.03     0.17 1.00     4903     2989
Pregnancies                 -0.01      0.04    -0.09     0.07 1.00     4278     3057
Age                          0.26      0.04     0.18     0.34 1.00     4078     3110
BMI                          0.20      0.03     0.13     0.27 1.00     4360     2836

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.94      0.02     0.89     0.99 1.00     5121     3022

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```



Rhat=1, Bulk_ESS and Tail_ESS are bigger, the model is good

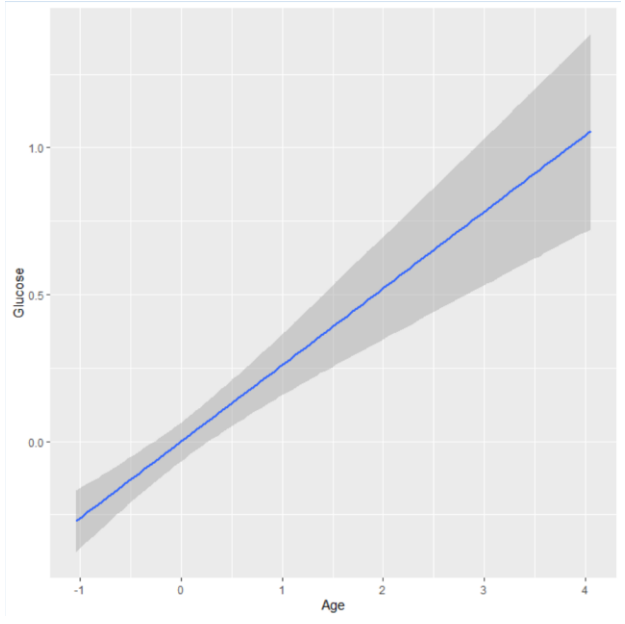
Model diagnostics: fit

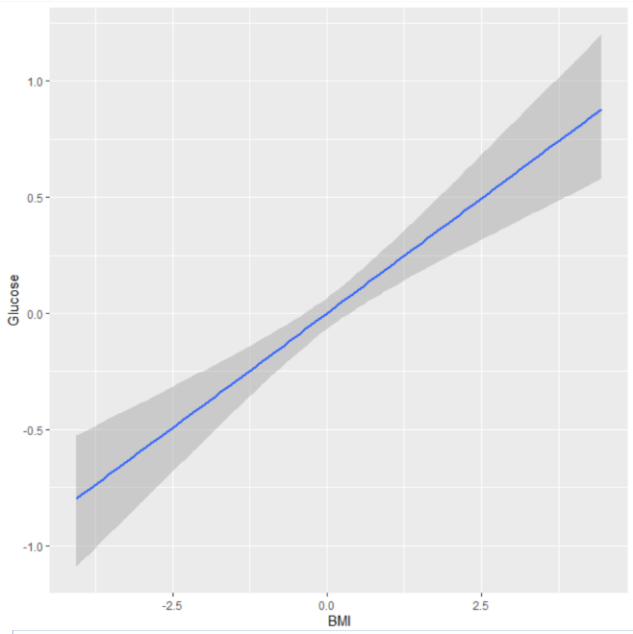
Posterior Probability Distribution of Coefficients for Each Variable

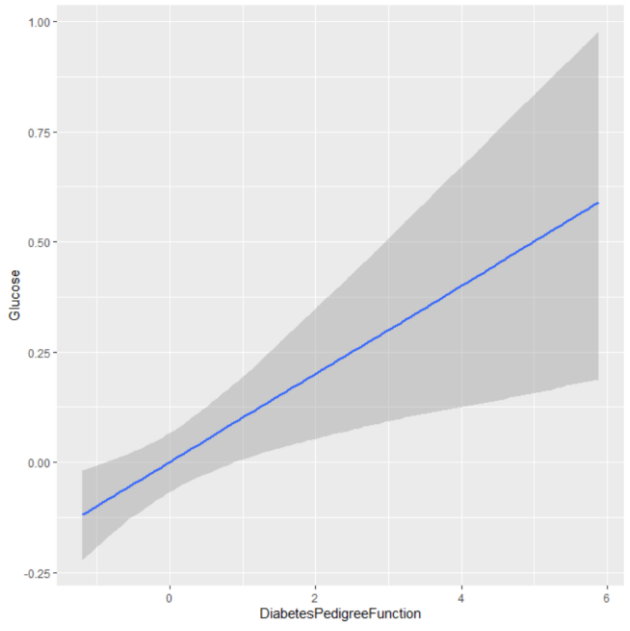


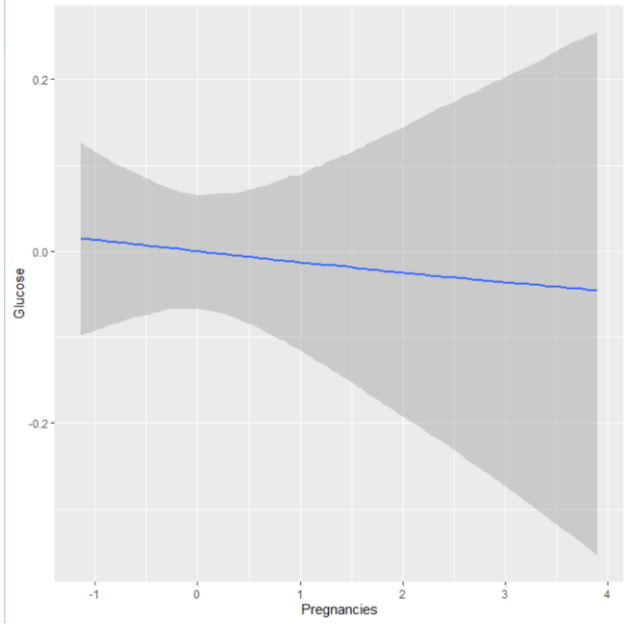
The model‘brms_m2’ explore the relationship of glucose level vs the other variables as BMI, Age, Pregnancies and DPF

Ranking Each Variable Contributing To Glucose Level









Age 0.25/unit,
BMI 0.20/unit,
DPF 0.10/unit,
Pregnancies -0.02/unit

Results show that

Age >BMI >DPF > Pregnancies

Pregnancies contributes nothing to increase glucose level, although pregnancies positively correlate with diagnostic diabetic outcome

Summary

This study use machine learning algorithm Random Forest and Bayesian inference analyze the diabetic dataset.

The results demonstrate that age, BMI, DPF and pregnancies positively correlate with diabetic outcome.

Pregnancies have no contribution to glucose level, although contributing to diabetic outcome.

New hypothesis should be formulated regarding pregnancies affecting diabetic outcome