A Bayesian's approach to predict the probability of getting diabetes

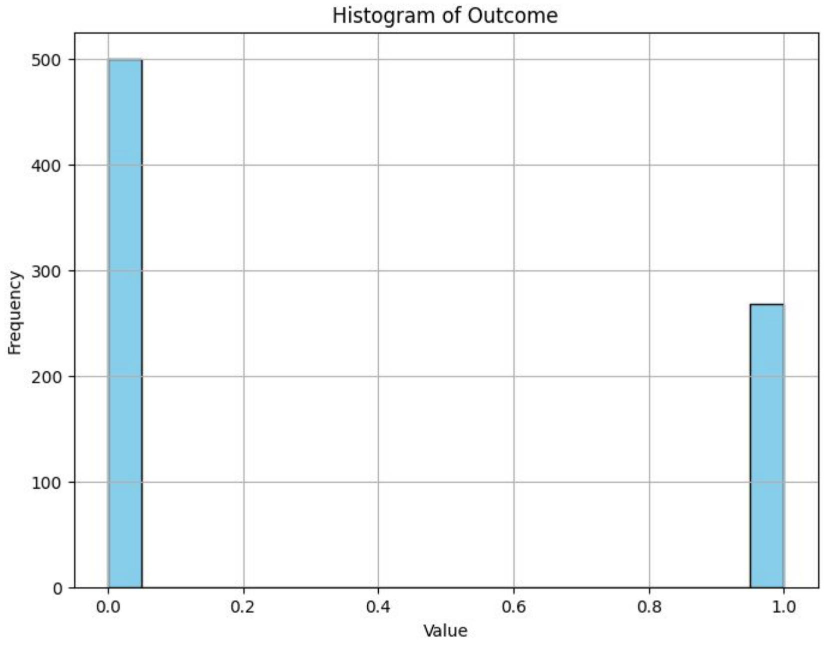
Saksham Malik, Alex Yang, Feifan Liu

Dependent variable - Outcome

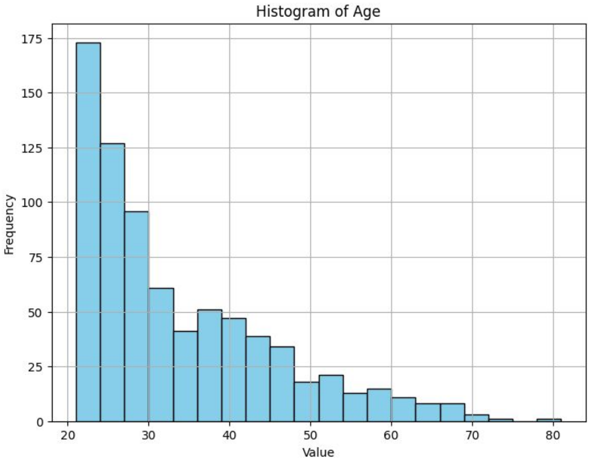
Outcome: Bernoulli RV

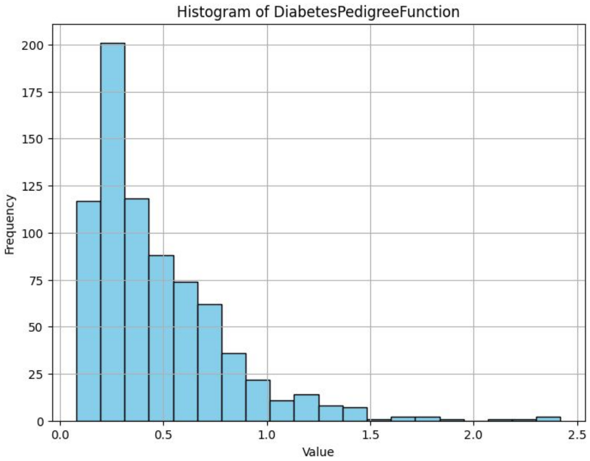
whether the person have a diabetes or not.

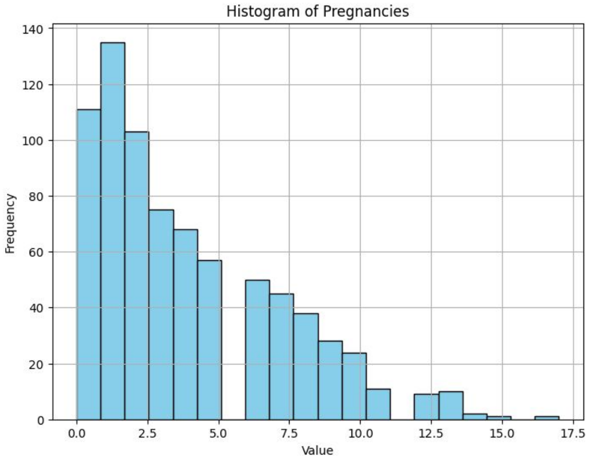
$P \approx 0.34$



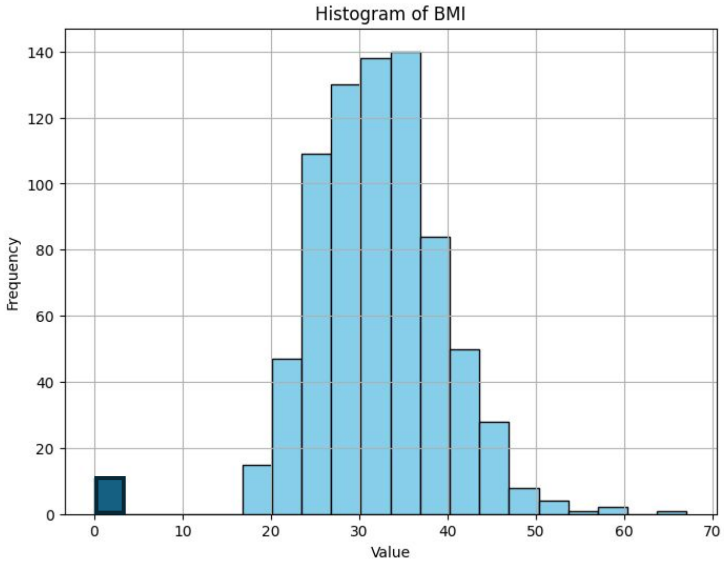
Features


Histogram of Age
Frequency
Value
175
150
125
100
75
50
25
0
20
30
40
50
60
70
80

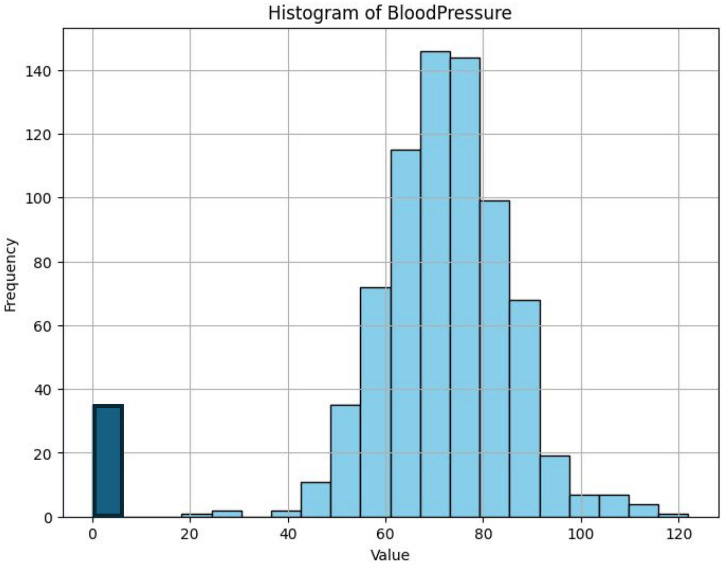

Histogram of DiabetesPedigreeFunction
Frequency
Value
200
175
150
125
100
75
50
25
0
0.0
0.5
1.0
1.5
2.0
2.5


Histogram of Pregnancies
Frequency
Value
140
120
100
80
60
40
20
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

Features

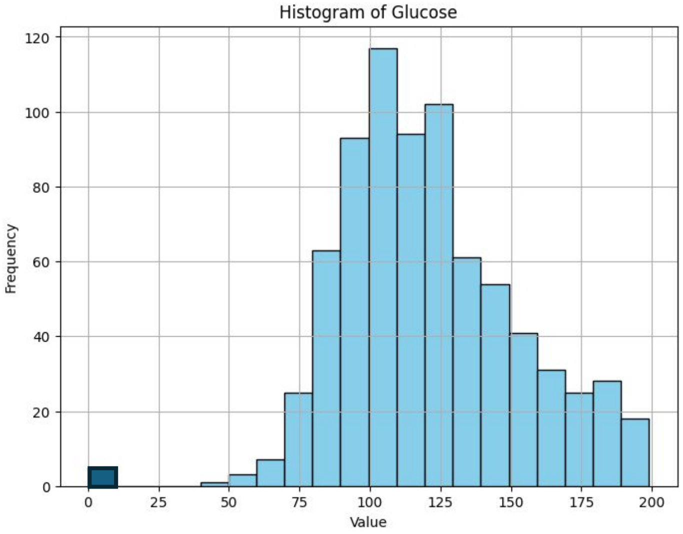


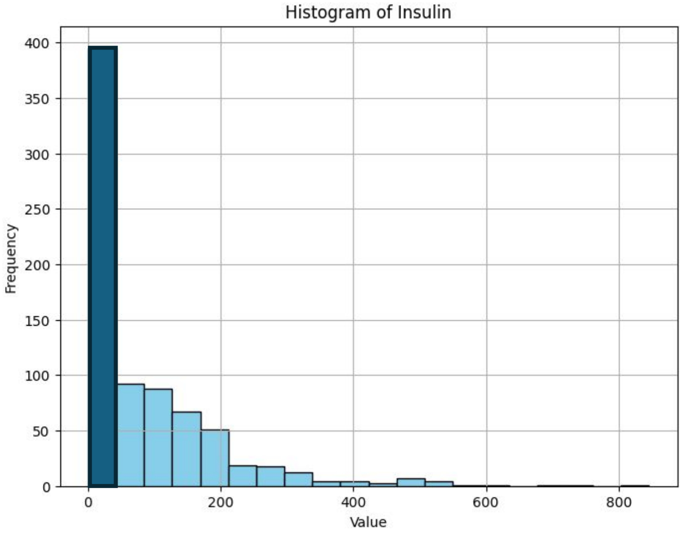
Histogram of BMI
Frequency
Value
0
20
40
60
80
100
120
140
0
10
20
30
40
50
60
70



Histogram of BloodPressure
Frequency
Value
0
20
40
60
80
100
120
140
0
20
40
60
80
100
120

Features


Histogram of Glucose
Frequency
Value
0
20
40
60
80
100
120
0
25
50
75
100
125
150
175
200


Histogram of Insulin
Frequency
Value
0
50
100
150
200
250
300
350
400
0
200
400
600
800

Data Imputation

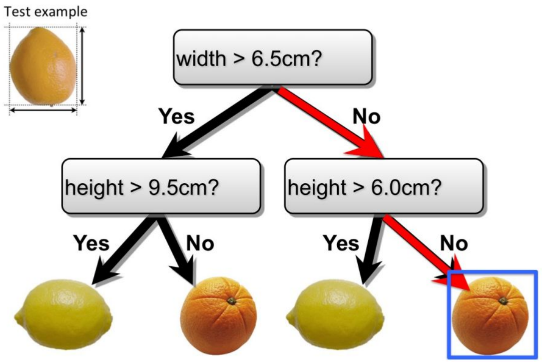
Multiple Imputation by Chained Equations (MICE) algorithm

1. Replace missing values with the column mean
2. Fix a column with missing values and fit a regression model on the remaining columns to predict the fixed column values
3. Predict missing value (which was set to mean initially) based on other columns and replace it.
4. Repeat for other columns
5. Repeat steps 2-4 for k iterations (often use k=5 in practice)

Assumption: Data is missing at random (MAR) or missing completely at random (MCAR). Missing not at random (MNAR) can only be imputed properly via Bayesian methods (equivalent algorithm is BICE)

Bayesian Additive Regressive Trees (BART) model

- Non-parametric model regression approach
- Performs bayesian model averaging (ensembling) on a large number of shallow(low depth) and sparse(low number of splits per level) decision trees.
- The hyperparameters α and β parametrize the probability that a node at depth $d(=0,1,2,\ldots)$ is non-terminal, given by $\alpha(1+d)^{-\beta}$.
- The default values $\alpha=0.95$ and $\beta=2$ ensure the trees are shallow.



Example of decision tree

BART algorithm

1. Recursive partitioning:
   a. Iterate through each feature (X) at each step.
   b. Choose the split point that minimizes an impurity measure (like variance) for the target variable (y). This creates two child nodes.
   c. Repeat splitting on the child nodes until a stopping criteria is met (e.g., minimum number of data points in a node). This creates a single decision tree.
2. Assign a constant value (average target variable) to each terminal node (leaf) of the tree.
3. Repeat Steps 1-2 ***m*** times (often use 50, 100, 200 in practice)
4. Apply regularization based on prior node probabilities
5. Prediction for new data point:
   a. Obtain a prediction from each tree (the value assigned to the terminal node where the data point lands).
   b. Use BMA on the predictions from all trees in the ensemble to get the final BART prediction for the new data point.

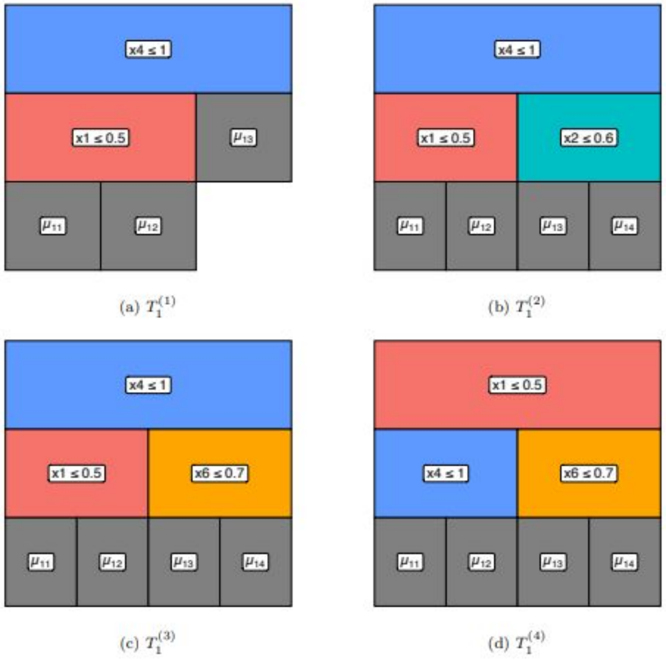
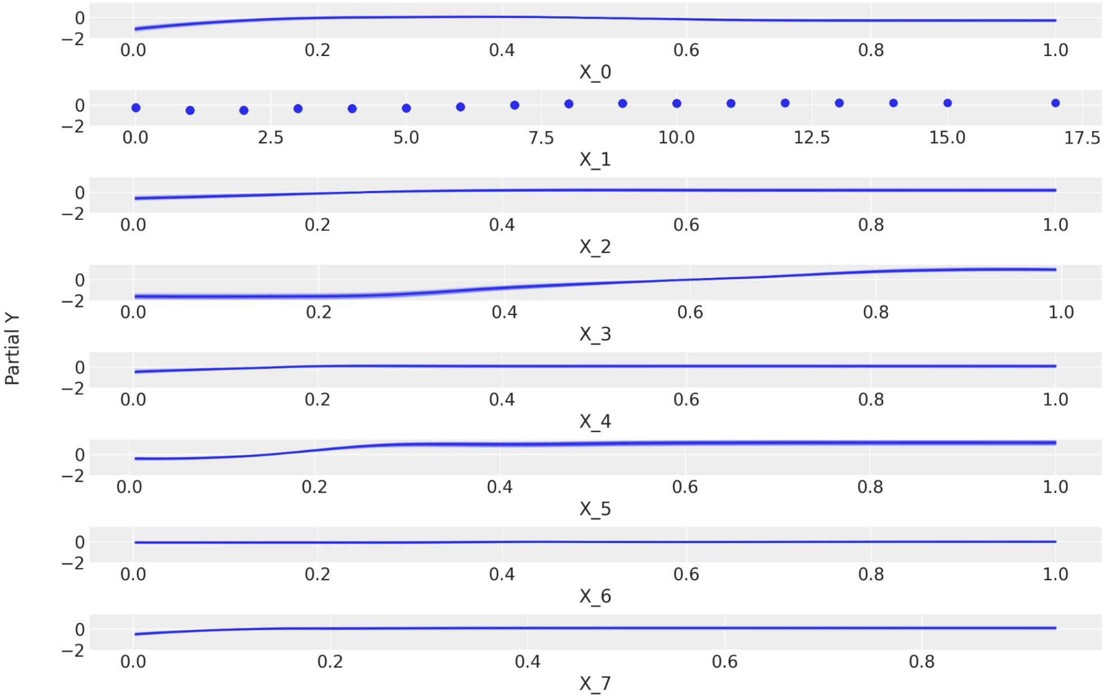


Figure1. - Inglis, A., Parnell, A., & Hurley, C. (2022). Visualizations for Bayesian Additive Regression Trees [arXiv:2208.08966]

Individual Conditional Expectance Plots



Age:X_0,

Pregnancies:X_1,

DiabetesPedigreeFunction:X_2,

Glucose:X_3,

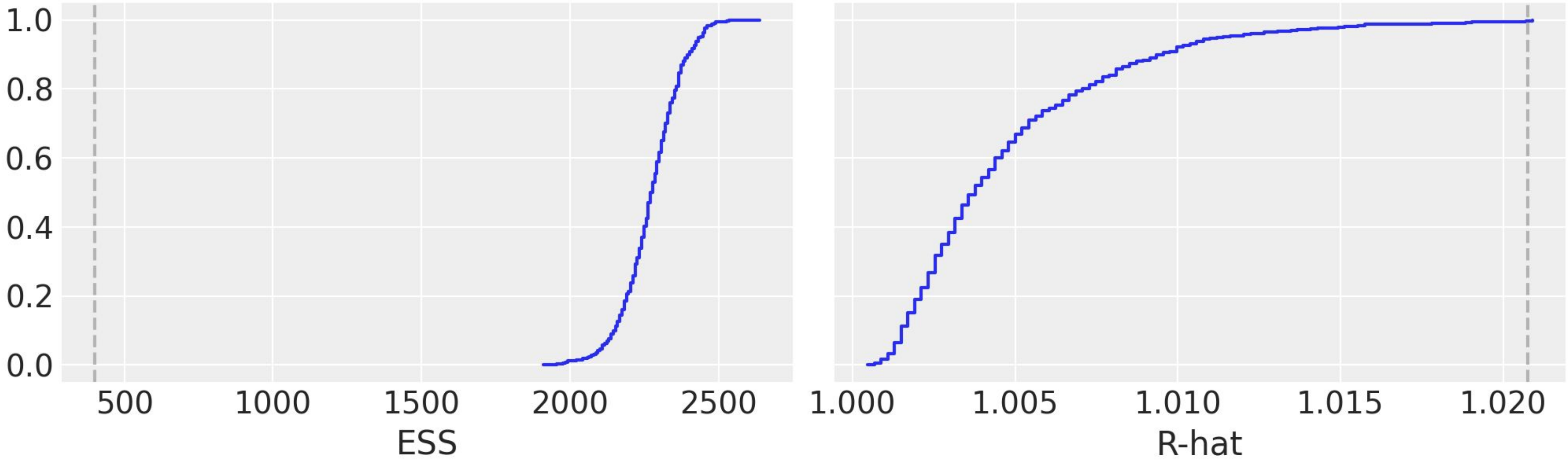
Insulin:X_4,

BMI:X_5,

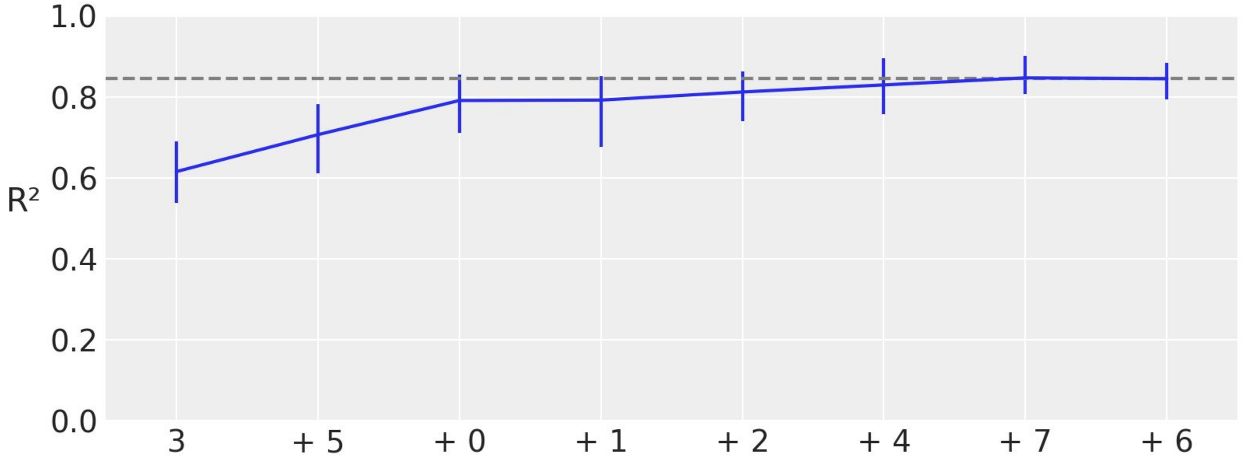
BloodPressure:X_6,

Glucose_Insulin_Product:X_7

Convergence diagnostics



Variable Importance Plots



Age:X_0,

Pregnancies:X_1,

DiabetesPedigreeFunction:X_2,

Glucose:X_3,

Insulin:X_4,

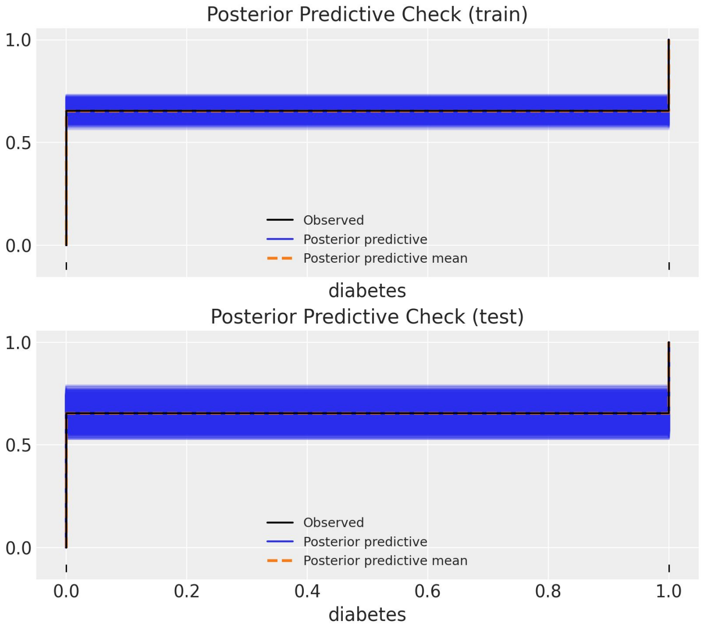
BMI:X_5,

BloodPressure:X_6,

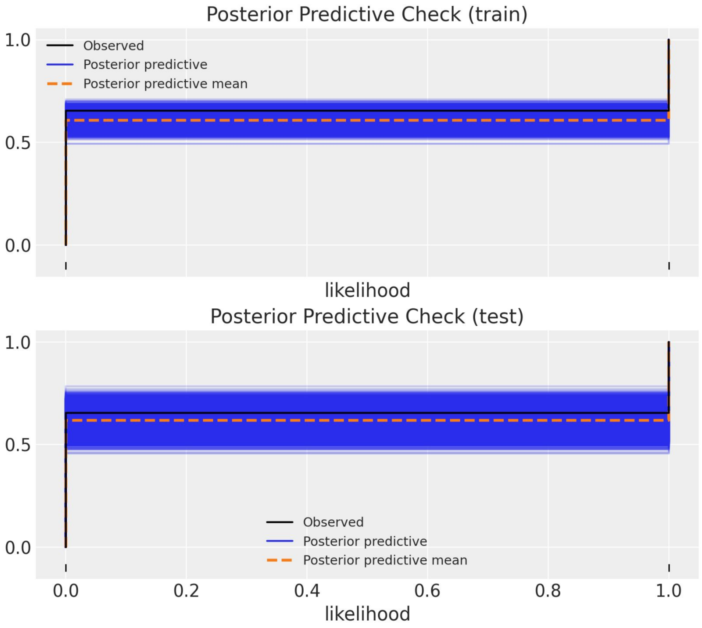
Glucose_Insulin_Product:X_7

Out of Sample Prediction

Bart Model



Regular Logistic Model



References

- PIMA Indians Diabetes Database. (2016, October 6). Kaggle. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
- Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). “BART: Bayesian additive regression trees.” The Annals of Applied Statistics, 4(1): 266–298 https://arxiv.org/pdf/0806.3286.pdf
- Inglis, A., Parnell, A., & Hurley, C. (2022). Visualizations for Bayesian Additive Regression Trees https://arxiv.org/pdf/2208.08966.pdf
- Decision Tree example image -Chris J Maddison STA314 Lecture 2 2021 https://www.cs.toronto.edu/~cmaddis/courses/sta314_f21/slides/lec02.pdf

Thanks For Listening