ISSN: 3116-3475
International Journal
on Education, Research, and Innovation Perspectives
Volume 2 Issue 2 (February 2026)

Classifying the Cumulative Grade Point Average of the Students Based on the Perceived Illnesses they Experienced: A Categorical Regression Analysis Using Bayesian Additive Regression Trees

Hannan S. Ampuan
Mindanao State University, Marawi City
ampuanhannan1@gmail.com

Date Submitted:
January 10, 2026

Date Accepted:
February 5, 2026

Date Published:
February 17, 2026

DOI:
10.5281/zenodo.18683439

ABSTRACT

The study introduced and utilized Byesian Additive Regression Trees (BART) to classify the Cumulative Grade Point Average (CGPA) of students based on the perceived illnesses they experienced. This addressed the lack of application of BART method in classification predictive modeling. The data was collected from 292 respondents, including 10 illness predictors and the CGPA. First BART model showed that all predictors had higher R^2 values, indicating multicollinearity, which was later confirmed through Variance Inflation Factor (VIF) assessment. After the refinement of the model, only four predictors- stress, anxiety, headache, and stomachache- were retained. Despite adjustments, the classification BART model yielded modest accuracy at approximately 66%. Among CGPA categories, CGPA-C was the most accurately predicted, while CGPA-A showed no correct classifications. Partial dependence plots and posterior predictive checks confirmed that individual perceived illnesses had minimal predictive influence. These findings suggested that although health perceptions were commonly reported among students, they were not strong predictors of academic performance as measured by CGPA. The application of BART highlighted its ability, complexity, and flexibility for statistical predictive modeling in classification setting for consistently employing higher R^2 values to all variable importance plots which determined the genuine relationship between CGPA and perceived illnesses.

Keywords: *Bayesian Additive Regression Trees (BART), Cumulative Grade Point Average (CGPA), Perceived Illnesses, Classification Modeling, Academic Performance, Multicollinearity, Variance Inflation Factor (VIF), Predictive Analytics, Partial Dependence Plots, Student Health*

INTRODUCTION

The performance of graduates from the College of Education at Mindanao State University – Main Campus, Marawi City, in the Licensure Examination for Teachers (LET) reflects the institution's strength in the field of education. In 2024, the passing rate for Elementary Education was 59.55%, while Secondary Education recorded 39.45%, resulting in an overall performance rate of 55.81% (Mindanao Varsitarian, 2024). This achievement continues a trend of academic excellence; for instance, in 2023, a graduate from the College of Education secured the 5th highest rank nationally in the LET with a rating of 91.40% (Mindanao Varsitarian, 2023). In 2022, the passing rates were 59.31% for Elementary and 48.17% for Secondary Education (Mindanao Varsitarian, 2022).

Motivated by this consistent performance, the researcher aimed to classify students' Cumulative Grade Point Average (CGPA) using Bayesian Additive Regression Trees (BART), a method under the umbrella of Bayesian statistics. Bayesian statistics, grounded in Bayes' theorem, offers a probabilistic approach to modeling uncertainty and updating beliefs as new data are introduced (Gelman et al., 2013). One of its key applications lies in predictive modeling—the process of building statistical models that generate forecasts based on historical data (Kuhn & Johnson, 2013).

In this study, predictive modeling was applied to predict the CGPA of College of Education students at MSU – Main Campus based on perceived illnesses using the datasets derived from the study conducted by Ms. Rohainah M. Hassan and Ms. Salihah M. Macarambon, under the advisement of Prof. Sittie Khaironisa M.S. Marohombsar, in their undergraduate study entitled "The Illnesses Experienced, Causes and Remedies as Perceived by College of Education Students". Since CGPA was treated as a categorical variable in the BART model in this study, categorical regression was the appropriate analytical method. Categorical regression models the relationship between a categorical dependent variable and a set of independent variables, whether continuous or categorical. It is particularly useful when outcomes fall into distinct groups rather than continuous values (Agresti, 2018).

To perform this analysis, the researcher used BART—a nonparametric Bayesian approach that combines multiple regression trees to model complex relationships between variables. Unlike traditional regression methods, BART incorporates uncertainty estimation and regularization through Bayesian principles. It has demonstrated high predictive accuracy and is increasingly applied in machine learning, causal inference, and data-driven forecasting. Introduced by Chipman, George, and McCulloch (2010), BART is especially effective in classification tasks and is well-suited for predicting outcomes like CGPA based on multiple influencing factors.

Objectives

This study aimed to introduce Bayesian Additive Regression Trees (BART) method for prediction using the Cumulative Grade Point Average (CGPA) of the College of Education students and the Perceived Illnesses datasets. Specifically, to

1. Provide step-by-step procedure in performing BART for predictive modeling of the Cumulative Grade Point Average (CGPA) of the College of Education students and the Perceived Illness datasets; and

2. Determine the model of prediction for classifying the Cumulative Grade Point Average (CGPA).

Conceptual Framework

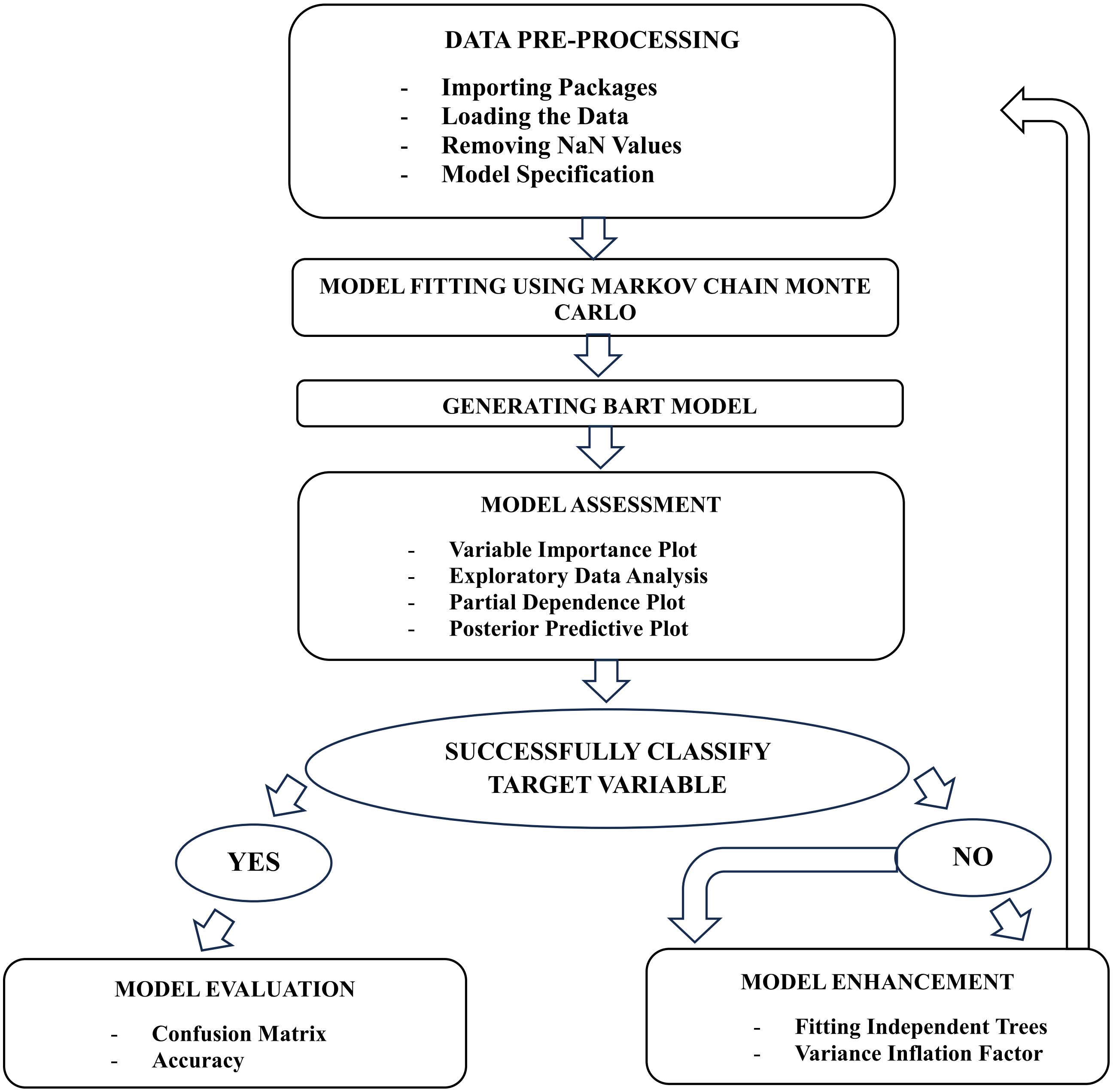

DATA PRE-PROCESSING
- Importing Packages
- Loading the Data
- Removing NaN Values
- Model Specification
MODEL FITTING USING MARKOV CHAIN MONTE CARLO
GENERATING BART MODEL
MODEL ASSESSMENT
- Variable Importance Plot
- Exploratory Data Analysis
- Partial Dependence Plot
- Posterior Predictive Plot
SUCCESSFULLY CLASSIFY TARGET VARIABLE
YES
NO
MODEL EVALUATION
- Confusion Matrix
- Accuracy
MODEL ENHANCEMENT
- Fitting Independent Trees
- Variance Inflation Factor

Figure 1 *Schematic Diagram of the Conceptual Framework of the Study*

The diagram as shown in Figure1.1, illustrated the step-by-step procedure of performing classification using Bayesian Additive Regression Trees (BART). It started with data pre-processing where importing packages, loading the data, removing NaN values, and specifying the model were performed, followed by model fitting using Markov Chain Monte Carlo with a BART classifier PGBART in Python and generate the model. The model is assessed through variable importance plot, partial dependence plot, and posterior predictive, then exploratory data analysis is utilized for justification of the multicollinearity. If the model successfully classified the target variable, then evaluate the model. If not, the model can be enhanced as part of hyperparameter tuning included fitting independent trees and variance inflation factor for the refined model to remove multicollinearity. After enhancing the model, it fitted the model again using Markov Chain Monte Carlo and processed the same performance. Note, even if the model did not classify the target variable, evaluating the model can be performed. Finally, the model's performance was evaluated using metrics such as confusion matrix and accuracy.

Significance of the Study

Future Researchers and Data Scientists. They will benefit from this study as BART provides a powerful approach to modeling complex relationships. This allows researchers to identify which illnesses significantly impact academic performance while also handling uncertainty, ensuring that conclusions are data-driven and accurate.

Educators and Academic Institutions. They can utilize the findings from BART to enhance student support systems. By identifying which illnesses most strongly correlate with lower CGPA, schools and universities can implement more flexible class settings. The ability of BART to uncover hidden patterns ensures that results are based on relevant evidence rather than assumptions.

Health Professionals and Student Counselors. They can use the findings of this study to develop more flexible approaches to student well-being. The application of BART can be used in the interest of health professionals and student counselors to future studies and references.

Policy Makers and Educational Authorities. They will find this study valuable in making evidence-based policies that support student well-being. BART's ability to analyze large datasets and extract meaningful patterns ensures that policy decisions are grounded in robust statistical analysis rather than simplistic correlations.

Students. Students themselves stand to benefit from this research by gaining a better understanding of how their health affects their academic performance. Students in the field of statistics can use the application of BART in this study to explore more variables or possible factors that may affect their CGPA.

Scope and Limitation of the Study

This study used the Cumulative Grade Point Average (CGPA) of the College of Education students and the Perceived Illnesses datasets which include 292 observations and 11 variables such as CGPA and the 10 illnesses. This study utilized Bayesian Additive Regression Trees (BART) to predict the classification of CGPA using the 10 illnesses as the explanatory variables, such as common cold, flu, sore throat, headache, stomachache, cough, back pain, stress, anxiety, and ulcer. The BART model treated the CGPA as categorical variable, and its ordinal nature was not incorporated in the analysis. Python was utilized specifically in Jupyter Notebook for performing the process of BART for classification predictive modeling.

Definition of Terms

Bayesian Additive Regression Trees (BART). BART is a Bayesian nonparametric ensemble method that combines multiple decision trees to model complex relationships in data for both regression and classification tasks (Chipman, H.A., George, E.I., & McCulloch, R.E., (2010)).

Modeling or Model. Model is a simplified representation of complex real-world system, often expressed mathematically, that is used to explain the relationship among variables or predict future outcomes (Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013)).

Prediction. Prediction refers to estimating an unknown outcome based on input features and a statistical model (Shmueli, G. (2010)).

Predictive Modeling. Predictive modeling involves data and statistical techniques to create models that forecast future events based on historical data (Kuhn, M., Johnson, K. (2013)).

Classification Predictive Modeling. Classification Predictive Modeling refers to the task of learning a target function that maps each set of attribute values to one of the predefined class labels (Hastie, Tibshirani, and Friedman (2009)).

Predictors. Predictors, also known as independent variables, are input used in a model to predict the outcome (dependent variable). They can be continuous, categorical, or binary (James, G., Witten, D., Hastie, T., & Tibshiranie, R. (2013)).

Cumulative Grade Point Average (CGPA). CGPA is used as a key indicator to assess students' academic success and can influence academic standing, scholarships, and eligibility for certain academic opportunities (Hughes & Hall, 2007).

Illness. Illness refers to the subjective experience of the patient who perceives a deviation from the state of health, with consequences for personal well-being and functioning (Pender et al., 2015).

Python. Python is a general-purpose programming language widely used for data science, machine learning, and statistical analysis. It is known for its simplicity and extensive libraries for numerical computation (van Rossum, G., & Drake, F.L. (2009)).

Jupyter Notebook. Jupyter notebooks facilitate a seamless integration of code and documentation, making it easier for researchers to share their work and results in a readable and reproducible format (Kluyver et al., 2016).

In this study, Bayesian Additive Regression Trees was utilized as the tool for predictive modeling where it created a model for classification of Cumulative Grade Point Average based on the perceived illnesses of the students using Python programming specifically Jupyter Notebook to execute the codes.

PRELIMINARY CONCEPTS

Bayes' Theorem

Bayes' Theorem is a fundamental concept in probability and statistics that describe how to update the probability of a hypothesis in light of new evidence. It provides a mathematical framework for understanding conditional probabilities, which are probabilities of events that depend on the occurrence of other events.

Bayes' Theorem states that for two events A and B, the probability of A occurring given that B has occurred (written as $P(A|B)$) can be expressed as:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

where:

- $P(A|B)$: **Posterior probability** – the probability of event A given that B has occurred. This is the update probability of the hypothesis A after observing evidence B.

- $P(B|A)$: **Likelihood** – the probability of observing B given that A is true.

- $P(A)$: **Prior probability** – the initial probability A before observing B.

- $P(B)$: **Marginal probability** – the probability of observing B, considering all possible causes. This can be computed as:

$$P(B) = P(B|A)\,P(A) + P(B|A')\,P(A')$$

where A' denotes the complement of A, or "not A."

Regression Trees

Regression trees are a form of recursive partitioning method where data are split into smaller and smaller subsets based on input features to minimize prediction error. A simple tree model is represented as:

$$Y = f(X) + \epsilon$$

where:

- Y is the target variable
- X is the feature vector
- $f(X)$ is the prediction function
- ϵ represents the residual error, usually assumed to follow a normal distribution.

Additive Models in BART

Additive models are an approach where a prediction is obtained by summing the outputs of multiple models. In BART, each component of the additive model is a regression tree, and the overall model is given by:

$$Y = \sum_{j=1}^{m} T_j(X) + \epsilon$$

where:

- m is the number of trees in the ensemble,
- $T_j(X)$ is the prediction of the j-th tree, and
- ϵ is a noise term, often assumed to be normally distributed with mean 0 and variance σ^2.

This additive approach allows BART to capture complex, nonlinear relationships by combining predictions from multiple trees, each representing a simple partition of the data. The concept resembles ensemble methods like Random Forests, though with a Bayesian twist, where each tree is regularized by priors.

Bayesian Framework and Priors

The Bayesian nature of BART introduces a probabilistic approach to modeling. Instead of point estimates for parameters, Bayesian methods estimate parameters, which yields uncertainty quantification in predictions.

- **Tree structure prior**: BART places a prior on the tree structure to control the likelihood of growing deeper trees. This prior typically assigns higher probabilities to shallower trees, helping to regularize the model by penalizing overly complex trees. A common prior might follow a probability of splitting that decreases as depth increases:

$$P(split\ at\ debth, d) = \alpha(1 + d)^{-\beta}$$

where α and β are parameters that control the depth preference.

Posterior Inference via Markov Chain Monte Claro (MCMC)

In Bayesian inference, we aim to find the posterior distribution over all unknown quantities, which here includes tree structures, node parameters, and the noise variance σ^2. The posterior distribution is given by

$$P\left(\{T_j\}, \{\theta_{jl}\}, \sigma^2 | data\right)$$

where:

- $\{T_j\}$ represents the set of tree structures,
- $\{\theta_{jl}\}$ are the leaf node values for each tree, and
- σ^2 is the noise variance.

BART typically uses MCMC techniques to draw samples from this posterior distribution.

- **Tree structure sampling**: This step involves updating the structure of each tree, including which features to split on and the corresponding threshold values, conditioned on the data and current values of other parameters.
- **Variance sampling**: The residual variance σ^2 is also updated in each iteration, allowing the model to adjust uncertainty as it learns from data.

Sum of Trees Model

The sum of trees model allows each tree to contribute incrementally to the overall prediction, rather than relying heavily on any single tree. This collective approach offers two key advantages:

- **Robustness**: By summing the contributions of many shallow trees, BART is less prone to overfitting compared to a single complex tree.
- **Non-linearity**: Although each tree captures only simple patterns (due to shallowness), their combination can model highly complex, non-linear interactions.

Mathematically, the target variable Y for a given input X can be represented as

$$Y = \sum_{j=1}^{m} T_j\left(X; \theta_j\right) + \epsilon$$

where $T_j\left(X; \theta_j\right)$ denotes the prediction from tree j with parameters θ_j.

Prediction Uncertainty

A major strength of BART, due to its Bayesian foundation, is its ability to capture uncertainty in predictions. By sampling from the posterior distribution of the model parameters, BART generates an ensemble of predictions that reflect uncertainty where point estimates are often taken as the means of posterior samples and credible intervals can be derived from the spread of the posterior predictive samples, providing a natural measure of uncertainty.

Hyperparameters in BART

Hyperparameters in BART influence the model's complexity and performance such as the number of trees m. Generally, more trees provide a better approximation of the response but can increase computational cost.

LITERATURE REVIEW

In view of rare cases of categorical regression study using Bayesian Additive Regression Trees (BART), this portion included only related studies of the introduction and application of the BART in various fields and in continuous response variable with a high-dimensional setting.

The Development of Bayesian Additive Regression Trees

Bayesian Additive Regression Trees is a strong non-parametric Bayesian approach for predictive modeling. The study of Chipman et al. (2010) entitled "Bayesian Additive Regression Trees" which focuses on developing a nonparametric approach for regression and classification that balances flexibility and interpretability, particularly in high-dimensional and nonlinear data scenarios. It introduces a 'sum-of-trees' model where each tree is a weak learner, utilizes a Bayesian backfitting Markov Chain Monte Carlo (MCMC) algorithm for fitting and inference, and employs regularization priors to prevent overfitting by constraining tree complexity. The study has found that BART performs well in high-dimensional datasets, capturing nonlinearities and interactions effectively and it provides posterior distributions, offering uncertainty quantification, which is a significant advantage over traditional machine learning methods.

The Enhancement of Bayesian Additive Regression Trees

High-Dimensional Setting

In today's big data world, a high-dimensional setting presents unique challenges and opportunities, requiring advanced techniques to extract meaningful patterns from complex datasets. The study of Boatman et al. (2021) entitled "Co-Data Learning for Bayesian Additive Regression Trees (BART)" which aimed to improve BART's efficiency in high-dimensional datasets by combining empirical Bayes (EB) with a co-data model for adaptive learning. It introduces EB co-BART, a method where EB estimates hyperparameters of BART, assisted by external co-data which allows for more accurate models while reducing dimensionality. This study implied that the approach improved BART's performance in high-dimensional environments by efficiently guiding variable selection and reducing overfitting.

Detecting Interactions

Detecting Interactions reveals how variables influence one another, uncovering patters that are often missed by examining factors in controlled. The study of Sparapani et al. (2021) entitled "Detecting Interactions with Bayesian Additive Regression Trees (BART)" which utilize BART for detecting high interactions in large datasets. It employs BART's flexibility to model complex, non-linear relationships, including variable selection for interaction terms in high-dimensional data. The study concluded that BART proves highly effective in identifying interactions between variables, outperforming traditional regression models in capturing complex relationships in high-dimensional settings.

Predictive Accuracy

Predictive accuracy is crucial for evaluating how well a model generalizes to a new data, directly impact the reliability of its outcome in real-world applications. The study of Hill & He (2015) entitled "Bayesian Additive Regression Trees for Predicting High-Dimensional Environments" which explores the application of BART in complex datasets, focusing on predictive accuracy and interpretability in high-dimensional settings. It applies BART to various high-dimensional classification problems, analyzes the impact of hyperparameter tuning and regularization priors on model performance, and conducts comparative analyses with alternative machine learning methods. The study concluded that BART exhibits strong performance in both regression and classification tasks, with significant advantages in interpretability and uncertainty quantification and highlights the importance of hyperparameter calibration for optimizing predictive power.

The Application of Bayesian Additive Regression Trees

Modeling

Modeling is essential for capturing the underlying structure of real-world systems to analyze and stimulate complex phenomena. The study of Murray, J.S., et al. (2021) entitled "Bayesian Regression Trees in High-Dimensional Environmental Modeling" which focuses on the application of Bayesian Additive Regression Trees (BART) for environmental data modeling, particularly in high-dimensional settings with numerous predictors. This study demonstrated BART's capacity for handling complex interactions among variables, providing accurate and interpretable predictions in environmental studies, especially where data is heterogeneous and high-dimensional.

Prediction

Prediction enables the anticipation of future outcomes based on current or historical data, which is crucial for decision-making and planning. The study of Liu et al. (2021) entitled "Application Study: Predicting Asthma Risk Using Bayesian Additive Regression Trees (BART)" which aim to predict childhood asthma incidence in high-risk urban areas using BART, incorporating environmental, genetic and socioeconomic factors. It integrated environmental pollution data, genetic, predispositions, and neighborhood socioeconomic metrics, used BART to model complex linear interactions and handle high-dimensional predictors, and cross-validation assesses predictive performance and variable importance. The study concluded that BART outperformed logistic regression achieving 85% accuracy, identified significant contributors like air pollution, income levels, and genetic makers and highlighted BART's capability for precise and interpretable health risk modeling.

Estimation

Estimation plays a key role in quantifying unknown variables, especially when direct measurement is difficult. The study of Zhang et al. (2020) entitled "Application of Bayesian Additive Regression Trees (BART) for Estimating Daily Concentrations of PM2.5 Components" that aimed to apply BART for predicting daily concentrations of fine particulate matter (PM2.5) components, including elemental carbon, organic carbon, nitrate, and sulfate across California from 2005 to 2014. It utilized predictors such as meteorological data, land-use variables, satellite-derived parameters, and chemical transport model simulations, performed cross-validation to evaluate out-of-sample prediction performance, and tuned BART to assess its ability to handle high-dimensional predictors and quantify prediction uncertainty. The study concluded that BART demonstrated strong predictive accuracy (R^2 values of $0.62 - 0.73$) and robust uncertainty quantification, and the model effectively estimated PM2.5 components at unmonitored locations, highlighting its utility for environmental and health impact studies.

METHODS

Setting Up the Data in Jupyter Notebook

Importing Packages and Data Loading

Importing important packages in performing BART was essential for the execution of the codes. The necessary packages were arviz, matplotlib, numpy, pandas, pymc, pymc_bart, seaborn, and sklearn. There are functions within a certain package where access is granted only if the package is imported. Loading the proper location of the data was important for be able to accurately prepare the data in analysis.

Model Specification

In this section, specification of model means declaring which variables in the datasets were the predictor variables and the response variable for the readability and interpretability of the model.

Response Variable (also known as the dependent variable or outcome variable): is the variable being studied and predicted. It is the variable that researchers are trying to predict or explain. In statistical models like linear regression, the dependent variable is typically denoted as (Y).

Predictor Variable (also known as independent variable, explanatory variable): is a variable that is manipulated or controlled by the researcher. It is the variable that is hypothesized to have an effect on the dependent variable. In statistical models, independent variables are typically denoted as (X).

Procedure of Performing Bayesian Additive Regression Trees (BART) in Building a Model for Prediction

Bayesian Additive Regression Trees (BART) Overview

BART models the relationship between input features (predictors) and the target outcome using an ensemble of trees. Each tree in the ensemble is a regression model, but in classification, it predicts probabilities of the class.

BART assumes that the data can be explained by a sum of trees which is given by

$$f(x) = \sum_{t=1}^{T} g_t\,(x)$$

where each tree $g_t(x)$ captures a small part of the data structure.

Prior Distribution for Trees

In BART, a prior is specified over the structure of the trees and their parameters. The priors are designed to regularize the model, preventing overfitting.

- The **tree structure prior** controls the growth of the trees, typically assuming that trees are shallow, meaning they do not grow too deep.
- The **parameter prior** typically assumes that the tree parameters (such as the coefficients) have normal distributions with small variances, reflecting a weak prior belief about the magnitude of the effects.

Model Fitting using Markov Chain Monte Carlo (MCMC)

Sampling the posterior is where the model parameters (tree structures and coefficients) are sampled from the posterior distribution using MCMC methods. This process involves Gibbs sampling to update the tree structures and the parameters iteratively, tree resampling where every iteration tree structure is altered, and parameter update based on the current tree structure. The goal is to estimate the posterior distribution of the model parameters, which is done over many iterations to ensure convergence.

Posterior Prediction

After fitting the model, predictions are made by averaging the outputs of all the trees in the ensemble, weighted by their posterior probabilities. In the classification context, this involves predicting class probabilities by averaging over all trees

$$\hat{P}(y = 1|x) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{1 + exp(-f_m(x))}$$

where M is the number of trees in the ensemble and $f_m(x)$ is the output of the m-th tree.

Model Interpretation

- **Variable Importance**

After fitting the model, examine the feature importance to understand which variables contribute most to the predictions. This can be done by looking at how often each feature is used to split the data across the trees in the ensemble.

- **Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is a fundamental step in statistical analysis that involves summarizing, visualizing, and interpreting datasets to uncover patterns. This is essential for understanding the nature of the data in identifying the potential predictors (features) for the response variable. This analysis generates a plot where the distribution of each feature is displayed according to the response variable

- **Partial Dependence Plots (PDPs)**

For each feature, examine how the predictions change as that feature’s value varies, while keeping other features constant. This can provide insights into the relationships between individual features and the target variable.

Model Uncertainty

i. Posterior Distribution

One of the strengths of BART is the ability to quantify uncertainty in predictions. Assess the variability in predictions by examining the posterior distribution of the model’s output. This can help gauge how confident the model is in its predictions.

ii. Credible Intervals

For each prediction, calculate the credible intervals (the Bayesian equivalent of confidence intervals), which gives a range of plausible values for the prediction.

iii. Variance of Predictions

Evaluate how much the model's predictions vary across the ensemble of trees, which gives an indication of model uncertainty.

iv. Posterior Predictive Checks

This involves checking if the predicted class probabilities align with the observed outcomes, which can be used for model validation using posterior predictive plot. If there is a presence of uncertainty, Variance Inflation Factor (VIF) should be utilized.

Variance Inflation Factor (VIF)

This is a statistical measure used to detect multicollinearity in a multiple regression model. It quantifies how much the variance of a regression coefficient is inflated due to correlation with other independent variables. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated, leading to unreliable coefficient estimates.

For a given predictor variable of X_j, the VIF is calculated as:

$$VIF_j = \frac{1}{1 - R_j^2}$$

where

- R_j^2 is the coefficient of determination (R-squared) obtained by regressing X_j on all other predictor variables.
- VIF_j represents the degree to which the variance of X_j is increased due to multicollinearity.

Model Evaluation

Split the Data

It involves splitting the data into training for model fitting and test for evaluation.

Fit the BART Model

- **Model Fitting**

Using the training data, fit the BART model, which involves sampling from the posterior distribution of the tree structures and their parameters using Markov Chain Monte Carlo (MCMC) methods.

- **Hyperparameter Tuning**

If necessary, adjust hyperparameters such as the number of trees, tree depth, or the number of MCMC iterations to optimize the model.

Evaluation

Use the trained BART model to predict the outcomes for the test set. Compute the appropriate metrics to assess the model's predictive performance on the test set.

i. **Accuracy**. The proportion of correct predictions.

ii. **Confusion Matrix**. To compute true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). This is useful for calculating precision, recall, and F1-score.

RESULTS AND DISCUSSION

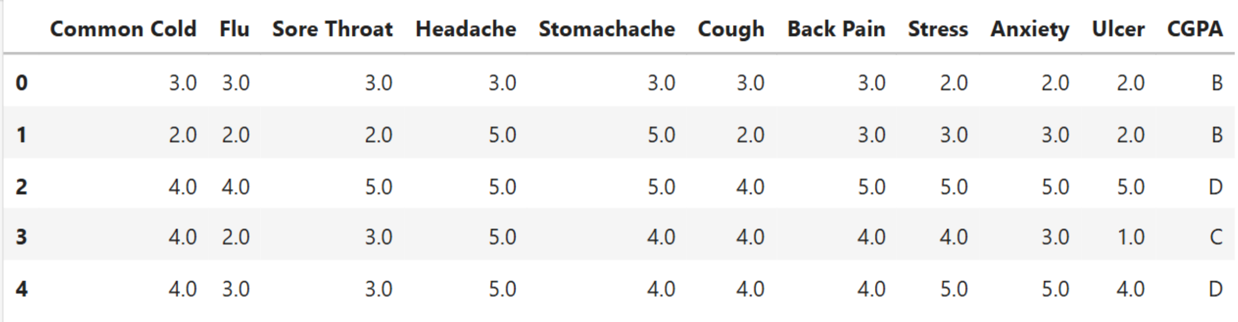
Step-by-step Procedure of Performing Bayesian Additive Regression Trees

Data Preparation

a. Importing Packages and Loading the Data using Jupyter Notebook

Packages in Python were essential to be extracted and imported for easy application of functions in the modules within a certain package. The packages needed to use prior to performing BART were arviz, matplotlib, NumPy, pandas, pymc, pymc_bart, and seaborn.

After successfully importing the packages, the researcher settled the format using random seed for reproducibility, consistency, and readability of the results in performing BART. The use of random seed ensured that the results remain the same after several times running the codes by controlling the random processes involved, such as MCMC sampling. Also, the researcher standardized the visualization of the results using az.style.use("arviz-darkgrid"). Loading the data and eliminating the NaN values was performed to clean the data using the dropna() function. Table 4.4 is the head of the data which included common cold, flu, sore throat, headache, stomachache, cough, back pain, stress, anxiety, ulcer, and CGPA.



	Common Cold	Flu	Sore Throat	Headache	Stomachache	Cough	Back Pain	Stress	Anxiety	Ulcer	CGPA
0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	2.0	2.0	2.0	B
1	2.0	2.0	2.0	5.0	5.0	2.0	3.0	3.0	3.0	2.0	B
2	4.0	4.0	5.0	5.0	5.0	4.0	5.0	5.0	5.0	5.0	D
3	4.0	2.0	3.0	5.0	4.0	4.0	4.0	4.0	3.0	1.0	C
4	4.0	3.0	3.0	5.0	4.0	4.0	4.0	5.0	5.0	4.0	D

Figure 2 Head of the Used Dataset

The variable CGPA was categorized with A, B, C, and D where it has true value of 1.00 – 1.25, 1.26 – 1.45, 1.46 – 1.75 and 1.76 – 2.00, respectively. The perceived illness variables have values from 5-point Likert scale responses where 1-Never, 2-Rarely, 3-Sometimes, 4-Often, and 5-Always. These were the responses of the students as to how they often experienced a certain illness that interfered with their studies.

b. Model Specification

For the preparation of the model, declared CGPA as the response variable, and the 10 perceived illnesses as the predictor variables. The CGPA data was declared categorical variable using pd.factorize() function. Defining coordinate level was essential for Bayesian analysis where it handled the data setting "predictors (n_obs)" as row labels, and the "CGPA" as the column labels for the data.

In setting up the model, the pm.math.softmax() function was utilized for μ from pmb.BART(), because it guaranteed that the vector sums to 1 along the axis=0 in the case. As shown in Table 4.2, the diagram was a Bayesian model using BART to predict a categorical variable CGPA across 292 observations where μ represented as latent parameter predicted by the BART model, $BART\_\mu$ which means that the parameter μ is modeled using a BART prior, that is, a non-parametric approach to capture complex relationships, θ representeds a deterministic transformation of μ, likely involving a softmax function to convert BART's outputs into probabilities for each category of CGPA, y.

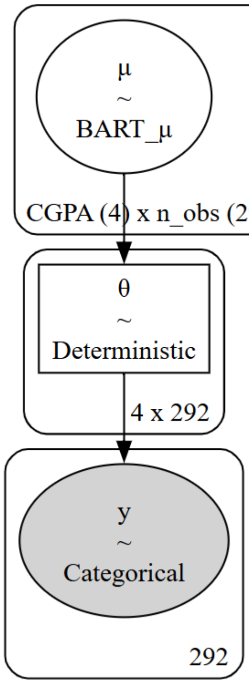


Figure 3 Diagram the Model

Model Building

a. Fit the Model

This included fitting the model and getting samples from posterior. The execution of the codes took 2 hours, 15 minutes and 1 second since the data used is large and reflects the complexity of BART. The sampling progress for a Bayesian model used PyMC with the PGBART (Probabilistic Gradient-Based Bayesian Additive Regression Trees) prior to the parameter μ. The model used 4 separate chains running in 4 jobs to explore the space of parameter μ. Since the parameter μ was modeled using a BART prior, the analysis used PGBART. The 0 divergence value indicated that the model was well-behaved and implied that the sampler has accurately explored the posterior distribution. Having zero divergence was a good sign that the model was stable and there was an accurate posterior estimation. The sample has been adjusted for accuracy and efficiency since it has tuned 1000 iterations, and the 1000 draws implied collecting samples after tuning to estimate the posterior distribution. Now, the model has 4000 tuning samples and 4000 posterior samples where there was 1000 per chain. After estimating μ using BART, model sampled the categorical variable y which only involved evaluating the likelihood based on the already sampled θ (probability distributions derived from μ).

b. Model Assessmet

i. Variable Importance

It may be that some of the illnesses were not informative for classifying CGPA, so in the interest of this study and in reducing the computational cost of model estimation, it was useful to quantify the importance of each variable in the dataset and performed variable importance. PyMC-BART provided the function plot_variable_importance(), which generated a plot that has showns its x-axis the number of covariables and on the y-axis, R^2 (the square of the Pearson correlation coefficient) between the predictions made for the full model where all variables are included.

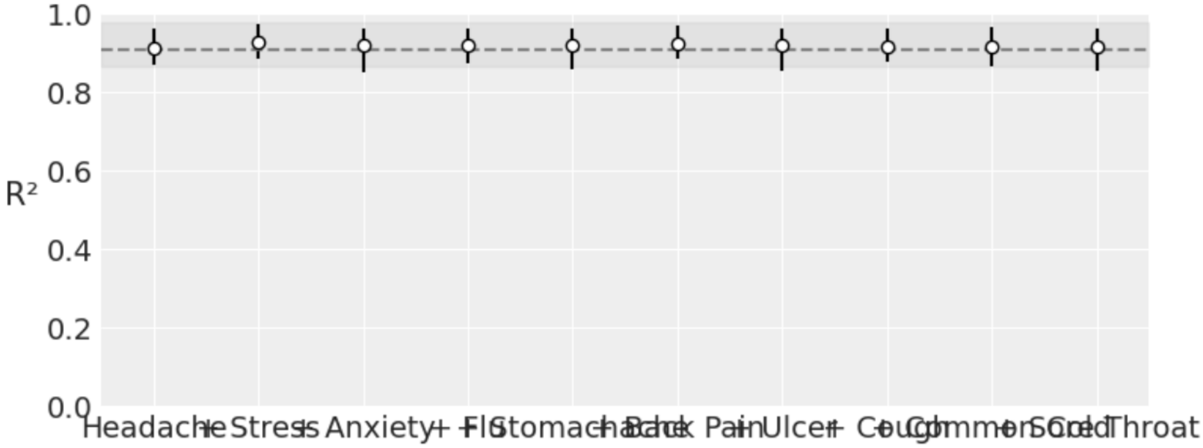


Figure 4 Variable Importance Results

As shown in Figure 4, the code in performing variable importance plot computed the significant importance of the data where idata was an inference data which contained posterior samples, and the method used was VI for the variable importance. In the plot, all R^2 values were higher (approximately from 0.9 to 1.0 which implied that each predictor (perceived illnesses) significantly contributed to the CGPA. This indicated multicollinearity where there was strong dependence among predictors. The error bars were relatively small (narrow confidence intervals), which suggested stability in the variable importance estimates. Moreso, there were no single predictors that stand out differently among the other predictors, indicating a potential redundancy among predictors.

ii. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was utilized for understanding, visualizing, and summarizing the datasets using sns.pairplot() function from the seaborn package. This helped identify patterns that will significantly impact model performance. As shown in Figure 5, it visualized the relationship between the perceived illnesses of students they experienced and their CGPA. The different colors were the different categories of CGPA, which suggested how the perceived illnesses impacteds the academic performance of the students based on their CGPA. The distribution of the common cold was right skewed, which implied that the respondents seldomly experienced it. The scatter plots indicated no strong correlations of common cold to other perceived illnesses, and the CGPA categories overlapped significantly with one another, suggesting that students experienced common cold may not have an impact on their CGPA. The distribution of flu revealed the same as the common cold, and how the flu may not solely impact the CGPA, while the scatter plots showed a slight correlation of flu to sore throat and headache. The distribution of sore throat showed the same as the common cold and flu, and it was unlikely to be a sole factor to impact CGPA, while the scatter plots showed a moderate correlation of sore throat to flu and headache.

Experiencing headache, stomachache, and back pain were of the same distribution with the common cold, flu, and sore throat having a right-skewed graph. While headache, stomachache, and back pain showed correlation to stress and anxiety in scatter plots, these three did not entirely differentiate the categories of CGPA. The distribution of cough has the same distribution as with common cold and flu and has a moderate correlation to sore throat. However, this perceived illness did not imply that it was a sole factor that may impact CGPA because of overlapping levels of CGPA. The perceived illness that showed unique distribution was stress, which has a strong correlation to headache, back pain, and anxiety since it has a more pronounced spread of data. The CGPA levels were somewhat mixed, and it implied that stress may have an impact on it. Anxiety has the same distribution as stress and was strongly associated with it, however, the overlapping categories of CGPA implied that even if anxiety was closely correlated with stress, it may have no impact on the CGPA. Moreso, the ulcer was right skewed and showed moderate correlation to stress and stomachache. According to the graph of CGPA, categories remained mixed which provided a hint that experiencing ulcer may have a minimal impact on the CGPA.

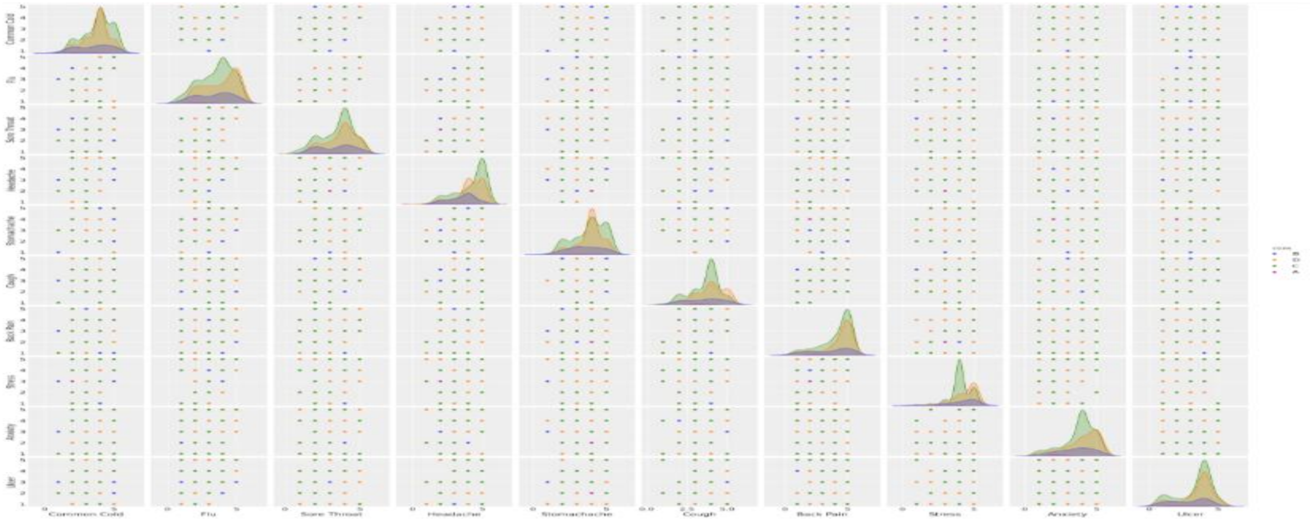


Figure 5 Exploratory Data Analysis

On the other hand, while the perceived illnesses showed interrelations such as stress, anxiety, stomachache and headache, the CGPA categories were intertwined across all the plots, suggesting that these perceived illnesses alone may not significantly predict CGPA. This evidence implied that there was a case of multicollinearity.

iii. Partial Dependence Plot

A Partial Dependence Plot (PDP) was used to understand the effect of a specific feature on the predicted outcome of the model while keeping other features constant.

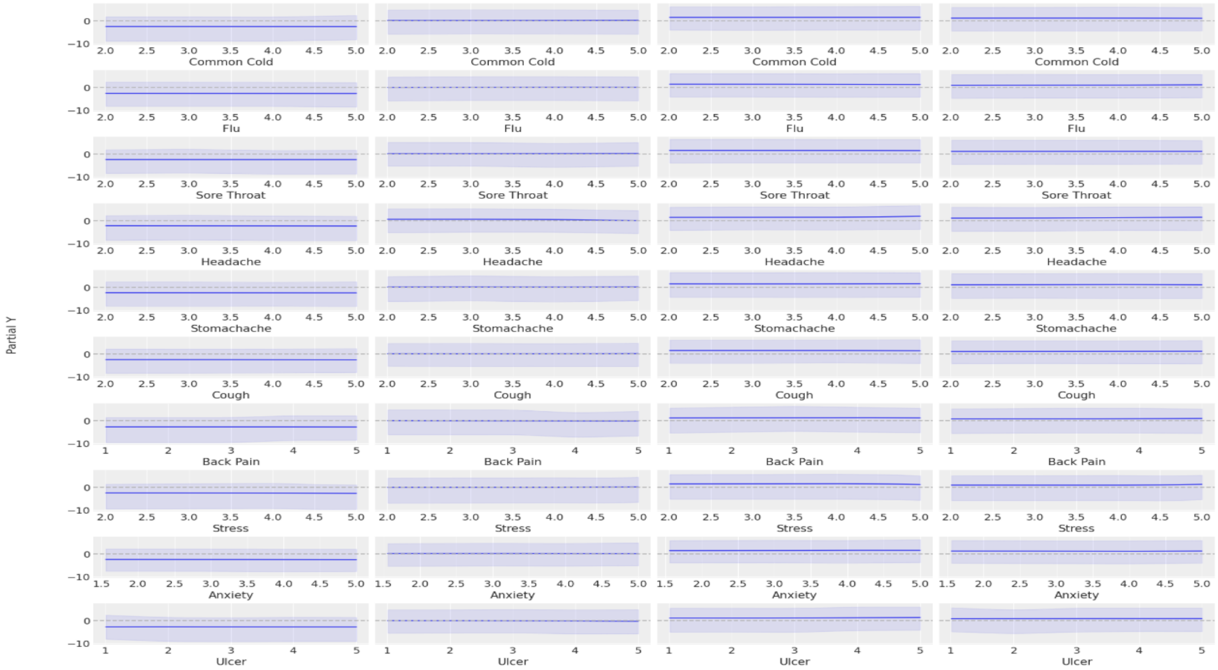


Figure 6 Partial Dependence Plot Results

As shown in Figure 6, the Partial Dependence Plot (PDP) visualized the relationship between multiple predictor variables (such as Common Cold, Flu, Sore Throat, Headache, Stomachache, Cough, Back Pain, Stress, Anxiety, and Ulcer) and the response variable (Partial Y, CGPA). The plot showed flat or near zero patterns that most predictors have little to no variation in their dependence values, which suggested that none of the variables have a strong sole impact on the outcome when considered controlled. Also, there were wide confidence intervals (blue-shaded bands) per predictor which indicated uncertainty in the model's estimates that potentially suggest collinearity among predictors (illnesses). With the same implication as Figure 5, there was a tendency to have redundant variables resulting in having a mostly linear or constant plot.

iv. Posterior Predictive Plot

The posterior predictive plot visualized the posterior predictive distribution of CGPA categories (A, B, C, D) based on a Bayesian model which determined how well the model captures the observed data and generates new data points based on estimated parameters.

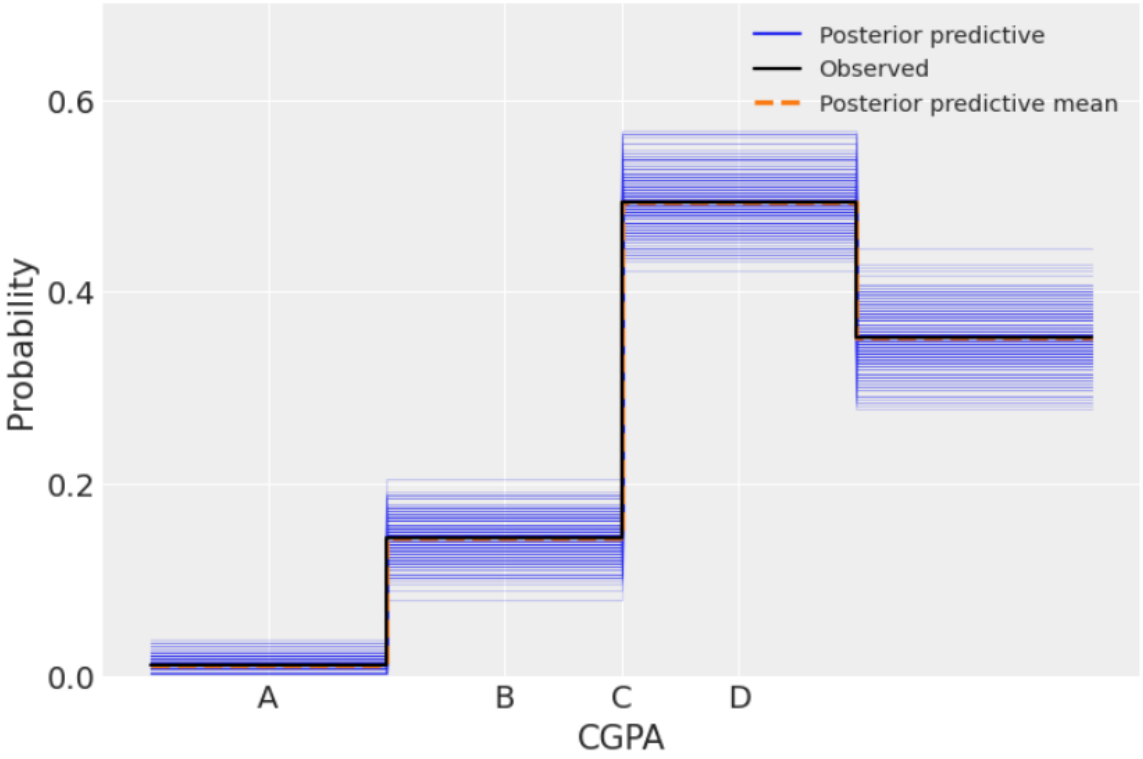


Figure 7 Posterior Predictive Plot

As shown in Figure 7, the plot showed that Category C has the highest probability (~0.6) and category D has the next higher probability (~0.4) while category A and B have lower probabilities, implying that fewer respondents have achieved higher CGPA. The observed data aligned well with the posterior predictive mean which implied that the model's predictions were consistent with the real data.

As observed data (black-colored line) is closely aligned with the posterior predictive mean (orange-colored dashed line), indicated that the Bayesian model effectively captured the CGPA probabilities distribution. In the plot, there was some uncertainty in the probability estimates since the posterior predictive samples showed some variability (blue-colored line) perhaps due to the possible collinearity as found previously in Figure 5 and Figure 6. However, there was a narrow range of uncertainty which means that despite this the model remained good for predictions.

In the previous assessment of the model, it has showed that it was a good fit model, however, the data was concerning because of the previous evidence of variability, possible collinearity among predictors, and overlapping categories of CGPA that indicated predictors may not be able to effectively classify CGPA. The obtained higher R^2 values made the whole result suspicious, and checking for multicollinearity is performed. However, to reduce variability shown in the previous partial dependence plots, adjusting the independent trees may be able to change the results (although it did not imply that it could change significantly).

Model Enhancement

a. Fitting Independent Trees

The option to fit independent trees with pymc_bart was set with the parameter pmb.BART(…, separate_trees = True, …). Fitting independent trees did not give a big difference in predictions but helped reduce the variability. The execution of the codes ran the same model and analysis as previous in Figure 5, 6, and 7, but fitting independent trees. However, the time of execution took longer than the recent run. This also utilized MCMC sampling for fitting the independent trees in the model.

b. Model Assessment

i. **Variable Importance of Adjusted Independent Trees**

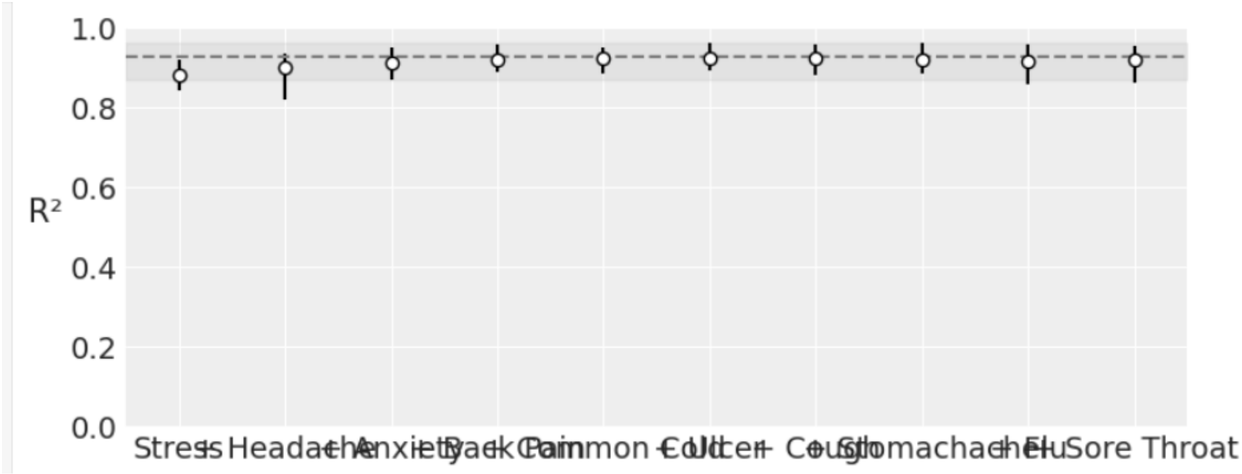


Figure 8 Variable Importance Plots After Adjusting Independent Trees

The plot in Figure 8 showed the same result and implication as Figure 4, where all the predictors have obtained higher R^2 consistently above 0.85. There was a slight variability in confidence intervals, which suggested that some predictors (e.g., Headache) may have fewer stable contributions, but in general, the predictors as one were strong predictors for CGPA.

ii. **Partial Dependence Plot of Adjusted Independent Trees**

As shown in Figure 9, this Partial Dependence Plot suggested that most perceived illnesses have a minimal predictive impact on the outcome variable, with only slight negative trends observed for stress, anxiety, headache, and back pain. The flat nature of many plots implied that changes in severity levels did not significantly altered the model's predictions, indicating that these perceived illnesses may not be strong predictors in the given dataset, which was of the same implication as Figure 6 although there were slight differences.

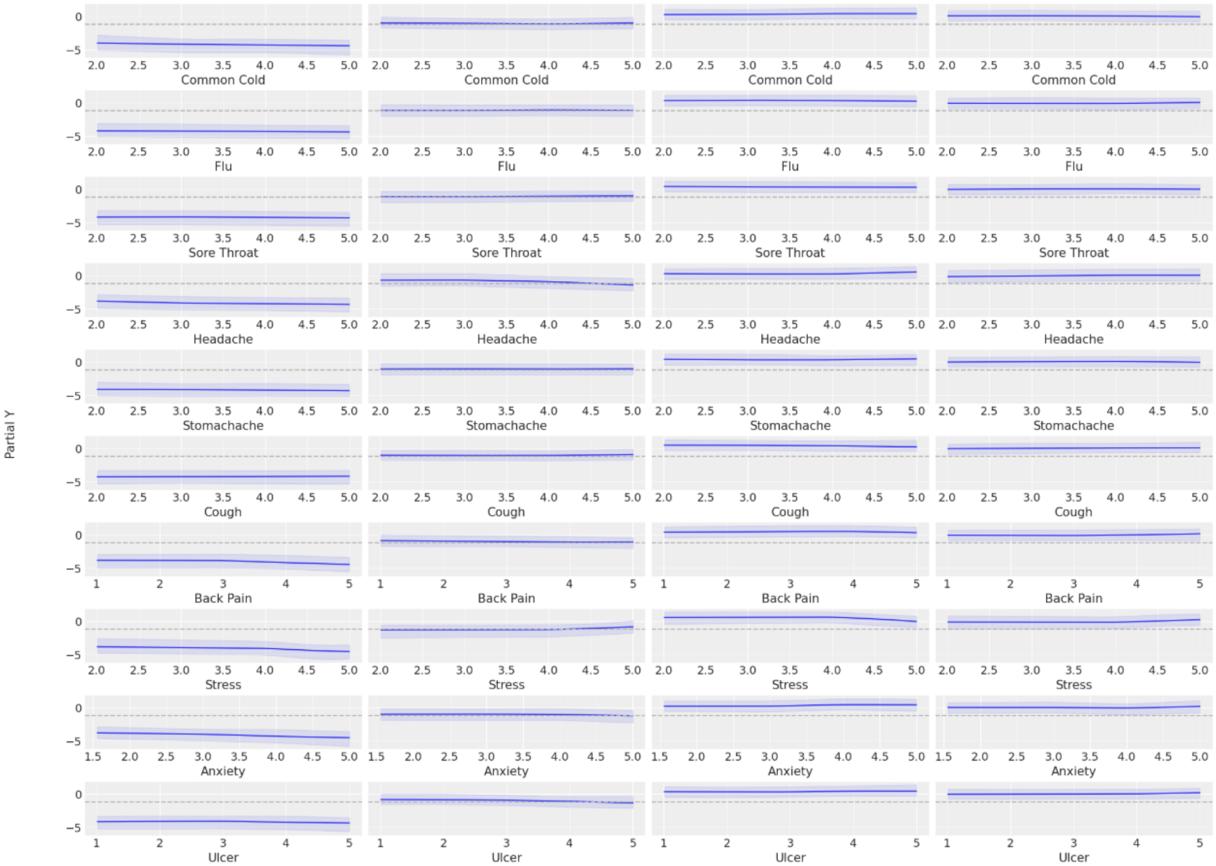


Figure 9 Partial Dependence Plots After Adjusting Independent Trees

iii. Posterior Predictive Plot of Adjusted Independent Trees

As shown in Figure 10, the plot was identical to Figure 7, which implied the same implications. In general, results indicated that perceived illnesses did not affect the CGPA of the students, but there was no single perceived illness that has a strong impact solely on CGPA.

The obtained high R^2 values suggested that experiencing the perceived illnesses played a role in the CGPA of the students, but the partial dependence plot confirmed that individual perceived illnesses have a small direct effect on the CGPA where the step-like pattern in the predictive posterior plot showed that multiple factors together influence CGPA.

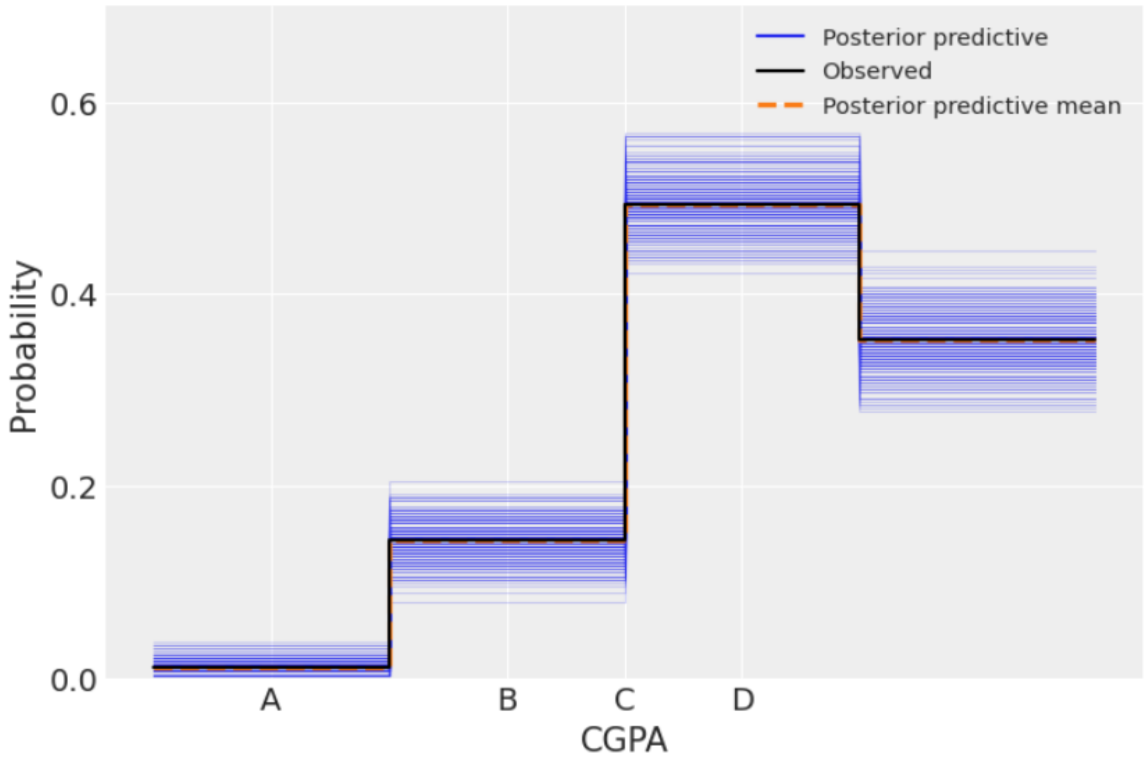
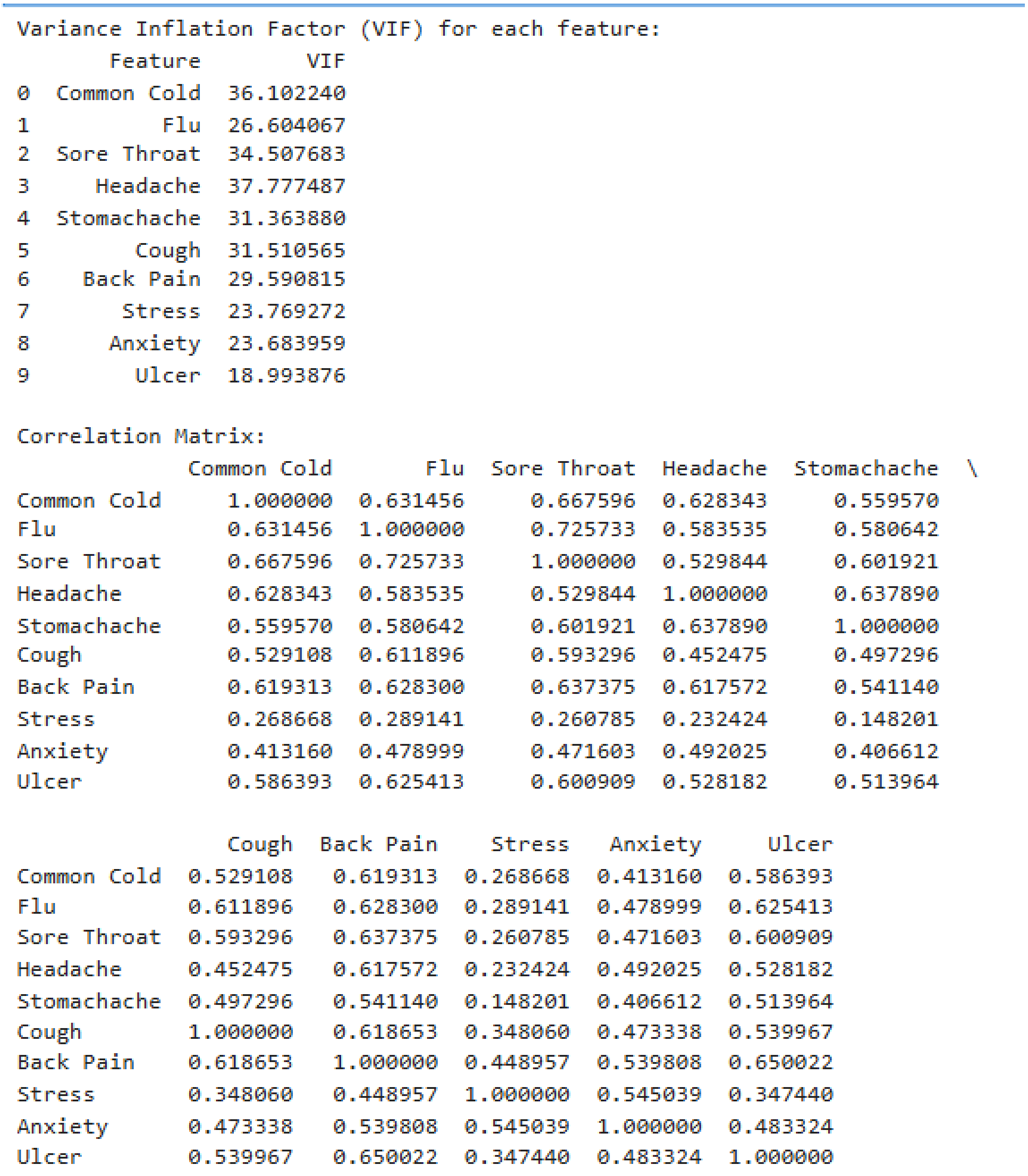


Figure 10 Posterior Predictive Plot After Adjusting Independent Trees

Overall, while the perceived illnesses students experienced, CGPA showed to have been impacted by the predictors together but not individually, making it hard to predict with complete certainty. With the cases of multicollinearity and uncertainty, Variance Inflation Factor was performed to refine the model.

Checking for Multicollinearity using the Variance Inflation Factor (Selecting Potential Predictors)

Performing the VIF of the predictors and correlation matrix determined what predictors were highly correlated to each other, then discarded them and identified potential predictors. As shown in Figure 11, the variance inflation factors were all higher, indicated that there was a severe multicollinearity, which suggested that the predictors shared a lot of information with one another. To determine which predictors were correlated with other predictors, correlation matrix was utilized. In the printed output, the predictors that were highly correlated having $r > 0.6$ were common cold and sore throat, flu and sore throat, back pain and ulcer, and cough and flu. These strongly correlated predictors confirm the evidence of the multicollinearity occurrence in the data, and to the previous analysis and results.



```
Variance Inflation Factor (VIF) for each feature:
       Feature        VIF
0  Common Cold  36.102240
1          Flu  26.604067
2  Sore Throat  34.507683
3     Headache  37.777487
4  Stomachache  31.363880
5        Cough  31.510565
6    Back Pain  29.590815
7       Stress  23.769272
8      Anxiety  23.683959
9        Ulcer  18.993876

Correlation Matrix:
             Common Cold       Flu  Sore Throat  Headache  Stomachache  \
Common Cold     1.000000  0.631456     0.667596  0.628343     0.559570   
Flu             0.631456  1.000000     0.725733  0.583535     0.580642   
Sore Throat     0.667596  0.725733     1.000000  0.529844     0.601921   
Headache        0.628343  0.583535     0.529844  1.000000     0.637890   
Stomachache     0.559570  0.580642     0.601921  0.637890     1.000000   
Cough           0.529108  0.611896     0.593296  0.452475     0.497296   
Back Pain       0.619313  0.628300     0.637375  0.617572     0.541140   
Stress          0.268668  0.289141     0.260785  0.232424     0.148201   
Anxiety         0.413160  0.478999     0.471603  0.492025     0.406612   
Ulcer           0.586393  0.625413     0.600909  0.528182     0.513964   

                Cough  Back Pain    Stress   Anxiety     Ulcer  
Common Cold  0.529108   0.619313  0.268668  0.413160  0.586393  
Flu          0.611896   0.628300  0.289141  0.478999  0.625413  
Sore Throat  0.593296   0.637375  0.260785  0.471603  0.600909  
Headache     0.452475   0.617572  0.232424  0.492025  0.528182  
Stomachache  0.497296   0.541140  0.148201  0.406612  0.513964  
Cough        1.000000   0.618653  0.348060  0.473338  0.539967  
Back Pain    0.618653   1.000000  0.448957  0.539808  0.650022  
Stress       0.348060   0.448957  1.000000  0.545039  0.347440  
Anxiety      0.473338   0.539808  0.545039  1.000000  0.483324  
Ulcer        0.539967   0.650022  0.347440  0.483324  1.000000  
```

Figure 11 Variance Inflation Factor and Correlation Matrix

With the basis of the values, common cold, flu, sore throat, headache, cough, and back pain were highly correlated to each other, but we remained the headache as the representative of the mentioned variables that were highly correlated to headache since in the Exploratory Data Analysis it showed a slight impact on the CGPA. Also, stomachache and ulcer were highly correlated to each other, and we remained the stomach in the refined model with the same grounds as to the headache. Stress and anxiety had smaller values and did not show strong correlation to other predictors, then they were included in the refined model.

Hence, these perceived illnesses such as common cold, flu, sore throat, cough, back pain, and ulcer were removed to refine the model for better prediction. The selected features after removing those highly correlated were headache, stomachache, anxiety (which were moderately correlated to other predictors), and most importantly, stress which showed a lower correlation across the other predictors which made it less affected by the multicollinearity.

a. Refined Model

The refined model included predictors such as headache, stomachache, anxiety, and stress as predictors. The results showed still the same conclusion as the previous model. The variable importance plot of the refined model as shown in Figure 12, R^2 values remained high regardless of removing other predictors which implied that the selected variables (stress, headache, stomachache, and anxiety) did not significantly affect the model predictive performance since the explained variance did not change much.

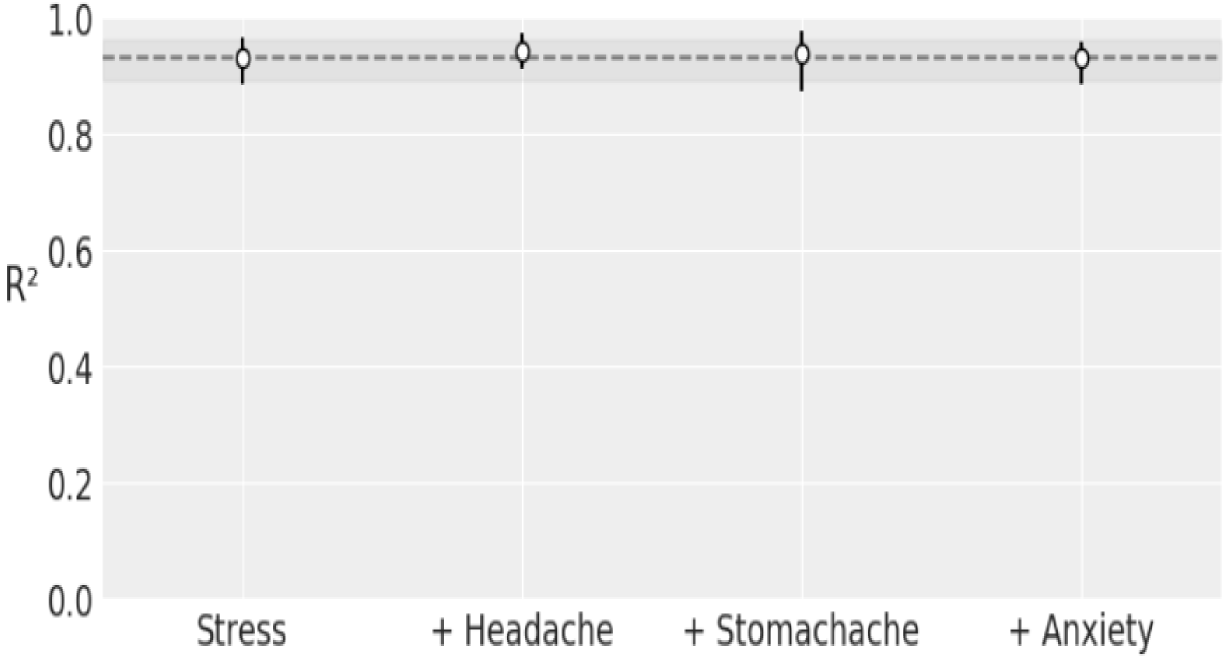
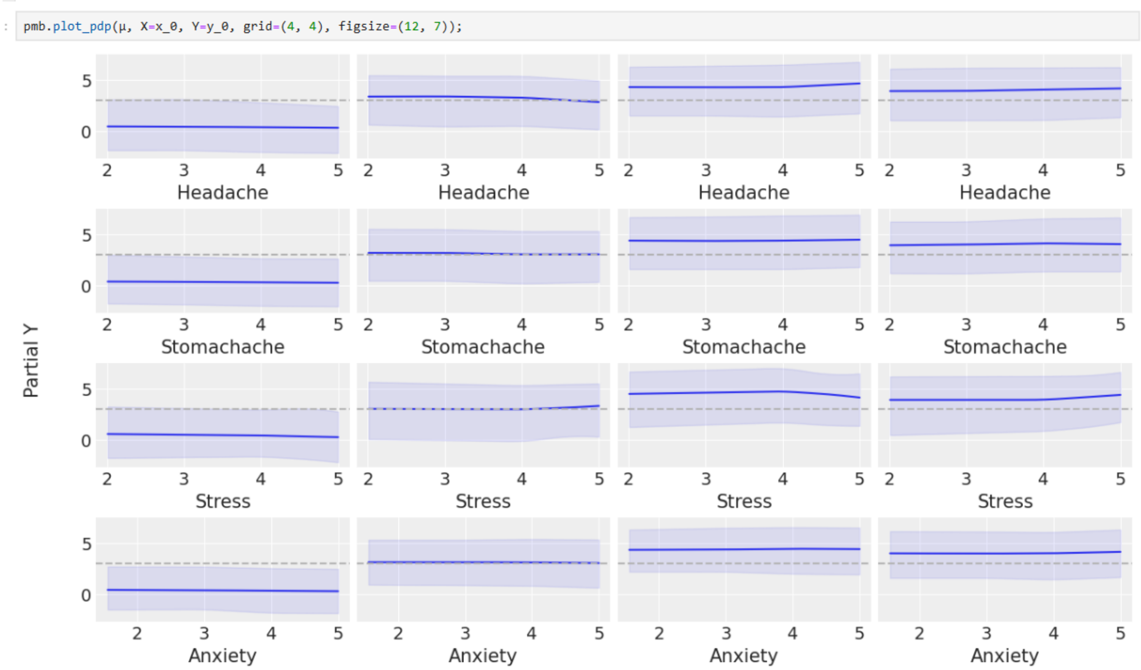


Figure 12 Variable Importance Results of Refined Model



```
pmb.plot_pdp(μ, X=x_0, Y=y_0, grid=(4, 4), figsize=(12, 7));
```

Figure 13 Partial Dependence Plot of Refined Model

As shown in Figure 13, the partial dependence plot of refined model revealed and supported the findings in Figure 12 that since most plots did not show strong increasing or decreasing trends, the predictors did not significantly impact CGPA. On the other hand, there was no meaningful relationship between the perceived illnesses and CGPA

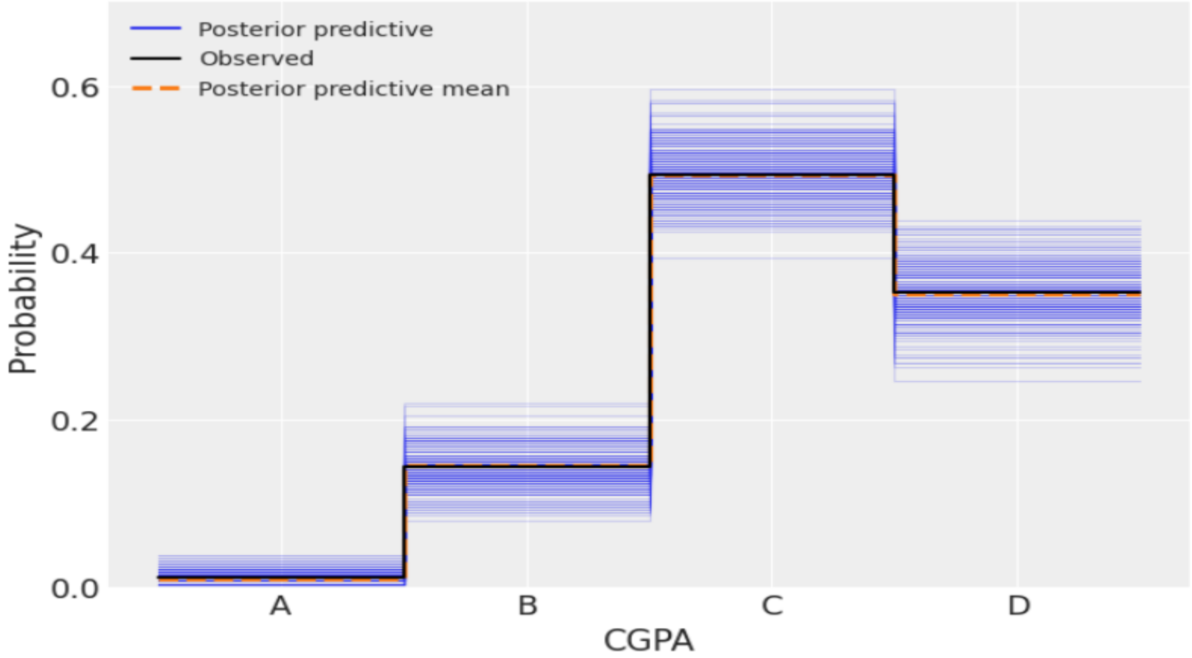
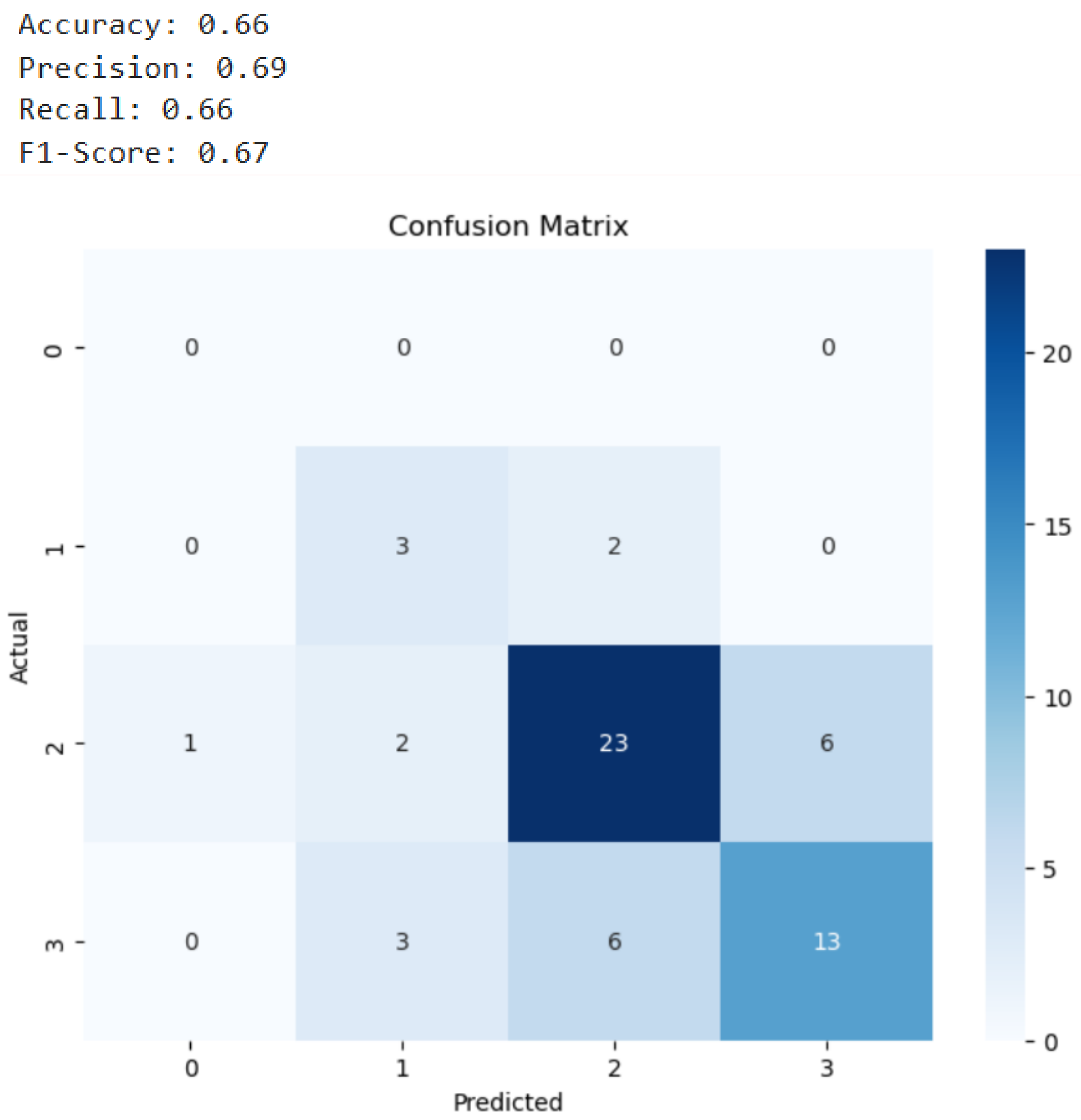


Figure 14 Posterior Predictive Plot of Refined Model

The posterior predictive plot of the refined model as shown in Figure 14, it showed the same plot as the previous models. The results further implied that perceived illnesses alone were not strong predictors of CGPA even after multicollinearity was resolved.

Model Evaluation

To support the results, the refined model was evaluated to examine why the perceived illnesses were not strong predictors to classify CGPA by computing the accuracy, precision, recall and F1-score values of the model and by plotting confusion matrix.



```
Accuracy: 0.66
Precision: 0.69
Recall: 0.66
F1-Score: 0.67
```

Figure 15 Model Evaluation Results

(Note: Class 0 is CGPA-A, Class 1 is CGPA-B, Class 2 is CGPA-C, and Class 3 is CGPA-D)

As shown in Figure 15, the confusion matrix and performance metrics provided insights into the classification model's effectiveness. The overall accuracy of 66% indicated that the model correctly predicted class labels in about two-thirds of the cases. However, a closer look at the confusion matrix revealed significant issues, particularly with CGPA-A, which was entirely misclassified. This suggested that the model struggled to recognize CGPA-A instances, due to an imbalance in the dataset or insufficient distinguishing features for this category. As a result, the recall for CGPA-A is 0%, which implied that the model failed to identify any true instances of this category.

For CGPA-B, the model correctly classified only three instances, while misclassifying others into CGPA-C and CGPA-D. This moderate misclassification rate indicated that the model has difficulty distinguishing CGPA-B from others, due to feature overlap. On the other hand, CGPA-C has the highest number of correct predictions, with 23 instances correctly classified, but 6 samples were mistakenly categorized as CGPA-D. This pattern suggested a possible similarity between the features of CGPA-C and CGPA-D, which led to confusion in the model.

Similarly, CGPA-D showed significant misclassification, with 6 instances incorrectly predicted as CGPA-C. This highlighted a challenge in differentiating between these two categories. The misclassifications between CGPA-C and CGPA-D indicated that the model did not effectively capture the subtle distinctions between these categories, which could be improved through better feature selection or refined classification techniques. The precision of 69% suggested that when the model predicted a class, it was often correct, but the recall of 66% implied that it missed a considerable number of actual instances.

Moreso, the Bayesian Additive Regression Trees (BART) model demonstrated strong performance in the analysis, as evidenced by its ability to accurately assessed the relationship between illnesses and CGPA. Despite prior concerns about multicollinearity among the illness predictors, BART effectively handled this issue and provided stable estimates. The Partial Dependence Plots (PDPs) revealed that individual perceived illnesses exhibited minimal impact on CGPA, reinforcing the conclusion that these predictors did not significantly contribute to the model's explanatory power. Furthermore, the posterior predictive plot aligned closely with observed probabilities, indicated that BART successfully captured the structure of the data, even when the relationship between the predictors and the outcome was weak.

A key strength of BART was its ability to avoid overfitting by flexibly adapting to nonlinear relationships while incorporating uncertainty. The variable importance plot confirmed that perceived illnesses, even after refining the feature set to remove highly correlated predictors, had negligible influence on CGPA. This outcome did not indicate model failure but rather affirmed that BART was robust in detecting genuine patterns rather than forcing misleading relationships. Additionally, the high R^2 value suggested that while illnesses did not strongly predict CGPA, the model as a whole still explained variance well, due to other covariates not included in the subset of predictors. If BART had performed poorly, we would expect a lower R^2 and erratic predictive behavior, neither of which was observed in the analysis.

Ultimately, BART's effectiveness lied in its ability to uncover true association or the lack thereof—without imposing assumptions about linearity or interactions. The results indicated that perceived illnesses did not strongly determine CGPA, and rather than fabricating a misleading influence, BART correctly assigned low importance to these variables which ranged from 0.8 to 1.00 which supported the findings from the study of Zhang et al. (2020) entitled "Application of Bayesian Additive Regression Trees for Estimating Daily Concentrations of PM2.5 Components" which concluded that concluded that BART demonstrated strong predictive accuracy (R^2 values of 0.62 – 0.73).

This outcome underscored that BART was functioning as expected, providing reliable insights rather than artificially inflating the role of weak predictors. If perceived illnesses were indeed significant factors that impact CGPA, BART would have identified and leveraged those patterns. Therefore, the model's performance should be regarded as strong, as it successfully differentiated between influential and non-influential predictors while maintaining high predictive accuracy between the Cumulative Grade Point Average of the students and the perceived illnesses they experienced.

SUMMARY

This study aimed to predict and build a model for the Cumulative Grade Point Average (CGPA) of the College of Education students based on the perceived illnesses they experienced using the Bayesian Additive Regression Trees (BART). Python specifically Jupyter Notebook was used to execute the codes prior to performing BART and evaluating the obtained BART model.

The following was the summary of the results from the findings:

1. The multiprocess sampling (using the Markov Chain Monte Claro) had 0 divergence indicating that model was well-behaved and implied that the sampler has accurately explored the posterior distribution. Also, the sample has been adjusted for accuracy and efficiency since it has tuned 1000 iterations, and the 1000 draws implies collecting samples after tuning to estimate the posterior distribution.

2. The first variable importance plot, as shown in Figure 4.3, has revealed that all the predictor variables contribute to CGPA since all has higher R^2 values. Implied that there was a case of multicollinearity.

3. The Exploratory Data Analysis (EDA) has shown that while other perceived illnesses moderately correlated to other perceived illnesses such as stress, anxiety, stomachache and headache, the CGPA categories were intertwined across all the plots, suggested that these perceived illnesses alone may not significantly predict CGPA.

4. The first partial dependence plot, as shown in Figure 4.5, has shown that all predictors had flat or near zero patterns implied that they had little to no variation in their dependent values, which

suggested that none of the variables had a strong sole impact on the outcome when considered controlled.

5. The first posterior predictive plot, as shown in Table 4.6, had the same implication as Figure 4 and Figure 6 which has shown uncertainty. However, the Bayesian model effectively captured the CGPA probabilities distribution since the observed data was closely aligned with the posterior predictive mean.

6. For fitting independent trees, it did not show a difference to the previous results that the Bayesian model we have was good and efficient, however, the perceived illnesses were not strong predictors enough to classify the CGPA of the students as the exploratory data analysis explained.

7. Performing Variance Inflation Factor (VIF) to see if there were dependencies in the predictors (to check for multicollinearity), it has shown in Table 4.10 that there were highly correlated predictors such as cold and sore throat, flu and sore throat, back pain and ulcer, and cough and flu. After removing these predictors, refined model was obtained where stress, anxiety, stomachache and headache were the predictors. These perceived illnesses were the features that the EDA had detected to have moderate correlation to each of them.

8. Reperforming the process of applying Bayesian Additive Regression Trees (BART) to the refined model, it has shown similar implication to the previous indicated that perceived illnesses alone did not classify CGPA regardless of how BART model was found efficient and good, and even after multicollinearity has resolved.

9. The model evaluation has shown an overall accuracy of 66%, indicated that the model correctly predicted class labels in about two-thirds of the cases, precision of 69% suggesting that when the model predicted a class, it was often correct, however, the recall of 66% implied that it missed a considerable number of actual instances and an F1 score of 67%.

10. The BART has shown a consistent result of determining that the perceived illnesses alone were not strong predictors enough to classify the CGPA. This indicated that no matter how intense or often a student experienced a certain (perceived) illness, they were not letting that to be a hindrance in achieving higher marks or let alone affecting their CGPA.

CONCLUSIONS

Based on the results and findings, BART detected genuine patterns and relationships between response and predictor variables. It consistently showed that from the beginning of the analysis the perceived illnesses that the students at College of Education experienced throughout their academe journey did not influence their Cumulative Grade Point Average (CGPA). This served an important information to Mindanao State University – Main Campus, Marawi City students that despite experiencing certain perceived illness, they had prioritized their academe achieving a higher mark or CGPA. Also, this was an indication that data which are perceptionally measured was not effective to use since it did not truly affect the actual data, for instance the CGPA. Perceptional datasets have been found in this study to not impact the actual data using BART, that was, classifying CGPA.

Moreover, BART's effectiveness lied in its ability to uncover true associations- or the lack thereof—without imposing assumptions about linearity or interactions. The results consistently indicated that perceived illnesses did not strongly classify CGPA, and rather than fabricating a misleading influence, BART correctly assigned low importance to the perceived illnesses variables. Therefore, the performance of the BART model was strong and efficient to have determined the true relationship between the CGPA, and the perceived illnesses students have experienced.

RECOMMENDATIONS

This study has revealed that BART was good and efficient in handling and detecting genuine relationships between response and predictor variables. To future researchers, since the perceived illnesses of the students experienced did not alone classify the Cumulative Grade Point Average (CGPA) of the College of Education students, broaden the population into the students at Mindanao State University – Main Campus, Marawi City to further validate the findings of this study and draw out meaningful patterns in that complex dataset. Since BART is known to accurately handle complexity, broadening the population of the dataset will not be a problem. Also, researchers may consider other predictor variables that are actual data such as the numbers of days absent, number of study hours, units enrolled and other variables that would likely affect CGPA or the demographic profile of the students since the illnesses were measured through how students perceived it by experiencing them not how it actually impacted their academic performance.

Moreso, since this study used CGPA as categorical variable, other researchers may consider CGPA to be continuous variable (numerical values). BART can perform similarly regardless of the type of variable to predict- may it be categorical or numerical. In lieu, researchers can also perform the same application of BART in classification using different statistical advanced software such as R studio. Additionally, researchers may also opt to do a comparative study of BART and other non-parametric regression methods to see if there are methods that excel in predictive performance other than BART.

REFERENCES

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. CRC Press.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics, 4*(1), 266–298. https://doi.org/10.1214/09-AOAS285

Diebold, F. X. (2012). On the origin(s) and development of the term "big data." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2152421

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography, 36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.

Good, I. J. (1983). *Good thinking: The foundations of probability and its applications*. University of Minnesota Press.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*(1), 389–422. https://doi.org/10.1023/A:1012487302797

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Hassan, R., & Macarambon, S. (2024). *The illnesses experienced, causes, and remedies as perceived by College of Education students*. [Unpublished manuscript].

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hill, J. L., Linero, A. R., & Murray, J. S. (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application, 7*(1), 251–278. https://doi.org/10.1146/annurev-statistics-031219-041110

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge University Press.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), Article 20150202. https://doi.org/10.1098/rsta.2015.0202

Kapelner, A., & Bleich, J. (2016). bartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software, 70*(4), 1–40. https://doi.org/10.18637/jss.v070.i04

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Liu, Y., Luo, Y., & Kapelner, A. (2023). Co-data learning for Bayesian additive regression trees. In *Proceedings of Machine Learning Research*. https://proceedings.mlr.press/v162/luo22a.html

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press.

MathWorks. (2020). *MATLAB documentation*. The MathWorks, Inc.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Murray, J. S. (2021). Log-linear Bayesian additive regression trees for multinomial logistic and count regression models. *Journal of Computational and Graphical Statistics, 30*(2), 408–422. https://doi.org/10.1080/10618600.2020.1833107

Murray, J. S., et al. (2021). Bayesian regression trees in high-dimensional environmental modeling. *Journal of the American Statistical Association*. https://link.springer.com

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

RStudio Team. (2020). *RStudio: Integrated development for R*. RStudio, PBC.

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310. https://doi.org/10.1214/10-STS330

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data*. O'Reilly Media.