JPPIPA 9(11) (2023)

Jurnal Penelitian Pendidikan IPA

Journal of Research in Science Education

http://jppipa.unram.ac.id/index.php/jppipa/index

Relationship Between BE4DBE2 and Variables n and z: A Comprehensive Analysis Using Linear Regression, Nonparametric Regression, Naive Bayes Classification, Decision Tree Analysis, SVM Analysis, K-Means Clustering, and Bayesian Regression

Budiman Nasution[1*], Winsyahputra Ritonga[1], Ruben Cornelius Siagian[1], Paulus Dolfie Pandara[2], Lulut Alfaris[3], Aldi Cahya Muhammad[4], Arip Nurahman[5]

[1] Departement of Physics, Universitas Negeri Medan, Medan, Indonesia.
[2] Departement of Physics, Universitas Sam Ratulangi, Manado, Indonesia.
[3] Department of Marine Technology, Pangandaran, Indonesia.
[4] Department of Electrical and Electronics Engineering, Islamic University of Technology, Dhaka, Bangladesh.
[5] Department of Physics Education, Indonesian Institute of Education, Garut, Indonesia.

Received: June 28, 2023
Revised: October 8, 2023
Accepted: November 25, 2023
Published: November 30, 2023

Corresponding Author:
Budiman Nasution
budimannasution@unimed.ac.id

DOI: 10.29303/jppipa.v9i11.4483

Abstract: This research employed various statistical techniques, including linear regression, nonparametric regression, Naive Bayes classification, decision tree analysis, Support Vector Machine (SVM) analysis, k-means clustering, and Bayesian regression, to analyze nuclear data. The research aims to explore the relationships between variables, predict binding energy, classify nuclear data, and identify similar groups. The research results revealed that linear regression indicated a significant influence of the intercept and predictor variable 'n' on the variable 'BE4DBE2,' while the variable 'z' was not significant. However, the overall model had limited explanatory power. Nonparametric regression with smoothing functions effectively modeled the relationship between 'BE4DBE2' and variables 'n' and 'z,' explaining approximately 11% of the variability in the response variable. Classification using Naive Bayes successfully categorized nuclear data based on 'n' and 'z,' revealing their relationship. Decision tree analysis evaluated the performance of this classification model and provided insights into accuracy, agreement, sensitivity, specificity, precision, and negative predictive value. SVM analysis successfully built an accurate SVM model with a linear kernel, classifying nuclear data while depicting decision boundaries and support vectors. K-means clustering grouped nuclear data based on 'n' and 'z,' revealing distinct characteristics and enabling the identification of similar clusters. The Bayesian regression model predicted binding energy using 'n' and 'z' as independent variables, capturing the Gaussian distribution of 'BE4DBE2' and providing statistical measures for parameter estimation. Ccomprehensives nuclear data analysis using various statistical approaches provides valuable insights into relationships, predictions, classification, and clustering, contributing to the advancement of nuclear science and facilitating further research in this field.

Keywords: Bayesian regression analysis; Decision tree analysis; K-means clustering; Naive bayes classification; SVM analysis

How to Cite:
Nasution, B., Ritonga, W., Siagian, R. C., Pandara, P. D., Alfaris, L., Muhammad, A. C., & Nurahman, A. (2023). Relationship Between BE4DBE2 and Variables n and z: A Comprehensive Analysis Using Linear Regression, Nonparametric Regression, Naive Bayes Classification, Decision Tree Analysis, SVM Analysis, K-Means Clustering, and Bayesian Regression. *Jurnal Penelitian Pendidikan IPA*, *9*(11), 9532–9546. https://doi.org/10.29303/jppipa.v9i11.4483

Introduction

In the realm of nuclear science, it is imperative to comprehend the interplay between independent variables and response variables, a fundamental aspect for the prediction and analysis of nuclear data as emphasized by (Ruso et al., 2022). This investigation is driven by a dual objective. Firstly, it seeks to scrutinize the connection between independent variables (n and z) and the response variable (BE4DBE2) through the utilization of both linear and nonparametric regression techniques. Secondly, it endeavors to categorize nuclear data through the employment of the Naive Bayes method and decision tree analysis.

The primary objective of this research is to advance our understanding of the relationship between independent variables and response variables in nuclear data and identify the optimal classification method for this data. This study aspires to make substantial contributions to the field of nuclear science, with potential applications across diverse industries, including nuclear technology and nuclear security, as acknowledged by (Juraku & Sugawara, 2021; Ylönen & Björkman, 2023; Stott & Bosman, 2021).

This research leverages both linear and nonparametric regression models to elucidate the intricate relationships between independent and response variables within nuclear data, as demonstrated by (Wongso et al., 2020; Boehm & Zhou, 2022; Zhong et al., 2023). These models not only deepen our understanding of variable interactions but also empower us to predict response values based on independent variable inputs. Moreover, the integration of the Naive Bayes method for nuclear data classification enhances prediction accuracy and efficiency, while the application of decision tree analysis aids in the creation of classification models based on relevant data attributes, as highlighted by (Ghiasi et al., 2020; Bashir et al., 2021; Mohamed & Kurnaz, 2023).

As a result, the outcomes of this research have the capacity to enhance decision-making processes within the nuclear domain, strengthening our understanding of nuclear data and classification models, as demonstrated by prior studies (Gomez-Fernandez et al., 2020; Papandrianos et al., 2022; Zhao et al., 2021).

Nevertheless, it is crucial to recognize the constraints inherent in this study, which primarily pertain to the limitations in the employed analytical techniques. The study predominantly focuses on delineating the models and methods utilized, without delving into an extensive discussion of the data analysis findings. Additionally, it's imperative to address existing gaps within the literature. For example, a more comprehensive exploration of the impact of supplementary variables, such as atomic mass or isospin, on the response variable would augment the depth of this study. Furthermore, the integration of advanced machine learning algorithms, such as deep learning or ensemble learning, holds the potential to further enhance the efficacy of the data analysis process.

The findings presented by Malerba et al. (2022) hold significant promise for the field of nuclear physics and engineering. They have the potential to pave the way for the development of innovative nuclear technologies and the design of next-generation nuclear reactors. Therefore, further investigation is warranted to explore these promising applications. Consequently, future research endeavors should prioritize addressing these knowledge gaps and delving deeper into the intricate connection between nuclear variables and their practical implications.

This study employs a diverse array of scientific methodologies, including linear regression, nonparametric regression, Naive Bayes, and decision tree analysis, to investigate the intricate relationship between independent variables and the response variable within nuclear data analysis (Xu et al., 2023). While these techniques offer valuable insights into the correlation among these variables, it is crucial to acknowledge the existing gaps in the literature that warrant attention. These gaps necessitate further exploration of additional variables and the integration of advanced machine learning algorithms. The potential ramifications of this research in the realms of nuclear physics and engineering underscore the critical need to address these gaps. By leveraging empirical data, conducting objective analyses, and providing recommendations for future research directions, this study adheres unwaveringly to the principles of rigorous scientific inquiry, with the overarching goal of expanding our collective knowledge within this domain.

Method

Linear Regression Method of Relationship between BE4DBE2 and Variables n and z

This study utilizes quantitative research methods, specifically employing linear regression analysis, to investigate the correlation between the independent variables (n and z) and the dependent variable (BE4DBE2). Linear regression is chosen to facilitate the measurement and statistical examination of data collected through measurement or observation techniques, as suggested by (Wang & Cheng, 2020). Processing and conducting the linear regression analysis are carried out using statistical software such as SPSS or R, as recommended by (Igartua & Hayes, 2021; Şahin & Aybek, 2019; Purwanto, 2021).

In this analysis, the independent variables are employed to predict the value of the dependent variable. The significance of the regression model is then assessed to gauge its effectiveness in elucidating the relationship between the independent and dependent variables, as advised by (Seki et al., 2019). The analysis results are subsequently used to draw conclusions regarding the adequacy of the linear regression model in explaining this relationship.

However, despite the significant influence observed between the independent variable n and the constant on the dependent variable, the results indicate that the linear regression model falls short of achieving sufficient accuracy in explaining the relationship between the independent variables (n and z) and the dependent variable (BE4DBE2).

To perform a robust linear regression analysis, we first need to meticulously prepare and import our data into the R environment, ensuring that we have correctly identified and selected the pertinent variables, namely, BE4DBE2, n, and z. This initial step sets the foundation for our analysis.

Next, we create a scatter plot, a visual representation that allows us to intuitively grasp the relationships between these variables and compute their correlation coefficient, enabling us to quantify the strength of their associations, as demonstrated by (Linardon et al., 2021).

Utilizing the powerful lm() function, we proceed to construct a precise linear regression model that encapsulates the interplay between these variables. To gauge the model's efficacy and how well it aligns with our data, we employ the summary() function, which provides essential insights into its goodness of fit.

For an even deeper understanding of our model's performance and to potentially uncover any underlying patterns or trends, we can generate a linear regression plot, as suggested by (Picard et al., 2021).

Ultimately, armed with this comprehensive regression model, we gain the ability to make predictions regarding the value of BE4DBE2 based on specific inputs for n and z. This predictive power offers invaluable insights into the intricate relationships between these variables, enhancing our understanding of the underlying dynamics.

The program employs the R programming language for a regression analysis, with the objective of elucidating the relationship between the dependent variable, BE4DBE2, and the independent variables, n and z, using data sourced from a CSV file. The analysis commences by importing the data using the read.csv() function, followed by the creation of two scatter plots. These plots, namely plot(BE4DBE2 ~ n, data = data, ...) and plot(BE4DBE2 ~ z, data = data, ...), visually depict the correlation between BE4DBE2 and each of the independent variables, n and z, respectively.

```
data <- read.csv("nama_file.csv")
plot(BE4DBE2 ~ n, data = data,
     main = "Hubungan antara BE4DBE2 dan n",
     xlab = "n", ylab = "BE4DBE2")
plot(BE4DBE2 ~ z, data = data,
     main = "Hubungan antara BE4DBE2 dan z",
     xlab = "z", ylab = "BE4DBE2")
cor_n <- cor(BE4DBE2, n)
cat("Koefisien korelasi antara BE4DBE2 dan n: ", cor_n, "\n")
cor_z <- cor(BE4DBE2, z)
cat("Koefisien korelasi antara BE4DBE2 dan z: ", cor_z, "\n")
model <- lm(BE4DBE2 ~ n + z, data = data)
summary(model)
plot(BE4DBE2 ~ n, data = data,
     main = "Hubungan antara BE4DBE2 dan n",
     xlab = "n", ylab = "BE4DBE2")
abline(model, col = "red")
plot(BE4DBE2 ~ z, data = data,
     main = "Hubungan antara BE4DBE2 dan z",
     xlab = "z", ylab = "BE4DBE2")
abline(model, col = "red")
new_data <- data.frame(n = c(1, 2, 3), z = c(4, 5, 6))
predicted_values <- predict(model, newdata = new_data)
cat("Nilai prediksi BE4DBE2 untuk n = 1, z = 4: ", predicted_values[1], "\n")
cat("Nilai prediksi BE4DBE2 untuk n = 2, z = 5: ", predicted_values[2], "\n")
cat("Nilai prediksi BE4DBE2 untuk n = 3, z = 6: ", predicted_values[3], "\n")
```

Figure 1. R programming language for the implementation of linear regression method of the relationship between BE4DBE2 with variables n and z (Source: data and image processing by the author)

To measure the correlation, we calculate the Pearson correlation coefficient using the cor() function and then display the resulting correlation coefficient values with the cat() function. Additionally, we build a linear regression model with the lm() function to estimate the BE4DBE2 value, taking into account the independent variables n and z.

As the program approaches its conclusion, it generates an additional scatter plot to effectively illustrate the correlation between the variables. This scatter plot incorporates a previously established linear regression line, which is distinctly highlighted in red (col = "red"). The predicted value for BE4DBE2 is then calculated using the predict() function, utilizing the new_data values for n and z. Subsequently, this predicted value is displayed using the cat() function.

Nonparametric Regression Method for Modeling the Relationship Between BE4DBE2 and Variables n and z

In this study, we employed a nonparametric regression model with a smoothing function, denoted as s(n,z), to effectively capture the intricate relationship between the response variable BE4DBE2 and the independent variables n and z. This method was selected to circumvent the assumptions of normality and linearity, thereby yielding more robust and insightful results when analyzing the interplay among these variables.

In order to assess the model's significance, we conducted t-value and F-statistic significance tests on both the intercept and smoothing function parameters, as detailed by (Demirhan, 2020). Furthermore, we employed several metrics to gauge the model's quality,

encompassing adjusted R-squared, deviance explained, generalized cross-validation (GCV), and scale estimate.

The model's results were visually conveyed through a plotted graph that depicted the predicted values of BE4DBE2 for various combinations of n and z values. This graphical representation facilitated a holistic grasp of the model's predictions and their fluctuations across diverse levels of the independent variables.

To effectively capture the non-linear relationships between BE4DBE2 and the independent variables n and z, the recommended approach is kernel regression. This method entails gathering data that includes BE4DBE2, n, and z, determining the optimal number of kernels and bandwidth, calculating kernel density for each data point, computing the mean value of the BE4DBE2 variable within each kernel range, and utilizing this mean value as the predicted outcome. By refraining from assuming any specific functional form, the kernel regression method circumvents certain limitations commonly associated with traditional parametric regression techniques.

The kernel regression method, as discussed by Taylor (2000) and Gradojevic et al. (2006), offers a valuable nonparametric approach for estimating non-linear relationships between variables. This method leverages a smoothing function and evaluates the model's significance and quality, thereby offering a dependable means of analyzing variable relationships without relying on strict assumptions of normality and linearity, as highlighted by (Wiedermann & Li, 2018; Meuleman et al., 2015).

The program utilizes the kernel regression method, a nonparametric regression technique, to effectively model the correlation between the response variable, BE4DBE2, and the independent variables, n and z. Initially, it reads data from the "data.csv" file, storing it in the "data" variable. For the kernel regression analysis, it employs 50 kernels while determining the bandwidth dynamically through the bw.nrd() function, which adapts the bandwidth according to the BE4DBE2 data distribution. A Gaussian function, with a standard deviation of 1.0, serves as the kernel function, ensuring the quality of the analysis.

For each data point in the "data" set, the program iterates through a loop, computing the kernel density by applying a previously defined kernel function and bandwidth, as outlined in studies by (Vatturi & Wong, 2009; Pelz et al., 2023). Subsequently, it calculates the average value of the BE4DBE2 variable for all data points falling within the kernel range. This computed average value serves as the prediction result for the respective data point. To retain and record these prediction results, they are appended to both the "predictions" vector and the original "data" set as a new column labeled "predictions". Finally, the program displays these prediction results using the "head()" function.

```
data <- read.csv("data.csv")
num_kernels <- 50
bw <- bw.nrd(data$BE4DBE2)
kernel <- function(x) {
  exp(-0.5 * x^2) / sqrt(2 * pi)
}
predictions <- c()
for (i in 1:nrow(data)) {
  # Calculate kernel densities
  kernel_densities <- rep(0, num_kernels)
  for (j in 1:num_kernels) {
    kernel_densities[j] <- kernel((data$BE4DBE2[i] - data$BE4DBE2) / bw)
  }
  weighted_sum <- sum(kernel_densities * data$BE4DBE2)
  weights <- sum(kernel_densities)
  prediction <- weighted_sum / weights
  predictions <- c(predictions, prediction)
}
data$predictions <- predictions
head(data)
```

Figure 2. R programming language for implementation of nonparametric regression method for modeling the relationship between BE4DBE2 and variables n and z (Source: data and image processing by the author)

This program facilitates nonparametric regression analysis, enabling predictions of the BE4DBE2 value based on the available n and z variables in the dataset. The resulting predictions offer valuable insights for analysis and decision-making purposes in relevant contexts utilizing the provided data.

Classification Method of Nuclear Data Using Naive Bayes Method

The study employed the Naive Bayes method to categorize nuclear data, using the variables n and z as inputs and the BE4DBE2 factor as the output. Initially, relevant nuclear data was collected and prepared by removing any irrelevant or missing information. Subsequently, the data was divided into two sets: one for training and another for testing. The training data was utilized to train the Naive Bayes classification model, where probabilities for each input variable were calculated for each output class. To validate the model's accuracy, the testing data was employed to compare predicted results with actual output values. Additionally, visual plots were generated to provide a clearer insight into the distribution of Naive Bayes classification predictions and the relationship between input and output variables.

In order to successfully implement the Naive Bayes method, we began by preparing the necessary training and testing datasets. The training dataset was comprised of categorized nuclear information, characterized by variables n and z, whereas the testing dataset consisted of nuclear data awaiting classification. Subsequently, we applied preprocessing steps to both datasets, which involved eliminating irrelevant data and formatting the

information to align with the algorithm's specific criteria.

Next, the model underwent training with the training data to compute the probability of the nuclear data class, leveraging variables 'n' and 'z.' Once the model completed its training, it became capable of making predictions during the testing phase by selecting the class with the highest probability as its prediction. Subsequently, the accuracy of these predictions underwent evaluation against the testing data. If the evaluation yielded unsatisfactory results, adjustments to parameters and the application of additional data preprocessing techniques were considered to enhance the model's performance.

```
install.packages("e1071")
library(e1071)
data <- read.csv("data_nuklir.csv")
index <- sample(1:nrow(data), round(0.8*nrow(data)))
training_data <- data[index, ]
testing_data <- data[-index, ]
predictors <- c("n", "z")
class_label <- "kelas"
model <- naiveBayes(training_data[, predictors], training_data[, class_label])
predicted_class <- predict(model, testing_data[, predictors])
accuracy <- sum(predicted_class == testing_data[, class_label])/nrow(testing_data)
cat("Akurasi prediksi: ", round(accuracy*100, 2), "%\n")
```

Figure 3. R programming language for classification method of nuclear data using naive bayes method (Source: Data and image processing by the author)

The program utilizes the "e1071" package to implement the Naive Bayes method for prediction. It reads the dataset from the "data_nuclear.csv" file using the read.csv function and splits it into training_data and testing_data in an 80:20 ratio. The variables "predictors" and "class_label" are assigned to the independent and dependent variables, respectively, for the prediction model. Subsequently, a naiveBayes model is created using the training data, predictor, and class variables. This model is then employed to evaluate the testing data, and prediction accuracy is determined by comparing the results with the actual values. Finally, the program displays the prediction accuracy results on the screen using the cat function. With the support of the e1071 package and the Naive Bayes method, this program has the capability to conduct predictions or classifications on any provided dataset, making it a versatile tool for predictive analysis.

Decission Tree Analysis Method

This study employs a decision tree methodology to analyze nuclear data, leveraging the power of machine learning. Decision trees, a key component of this approach, utilize rules derived from training data to predict target values. In this specific research endeavor, the decision tree's focus lies in predicting the value of BE4DBE2 based on n and z parameters. The decision tree's construction involves breaking down the training data into subsets, ensuring homogeneity with respect to the target values. For each subset, the most informative predictor variable is selected in an iterative process until a stopping condition is met. To gauge the model's performance, various metrics are used, encompassing accuracy, agreement, sensitivity, specificity, precision, as well as positive and negative predictive values. Furthermore, the decision tree's insights are made visually accessible through the use of the rpart.plot library, providing a clear representation of the rules and conditions essential for making accurate predictions. For those interested in implementing this analysis in R, the "rpart" library package is a requisite. Below is an illustrative R program, showcasing how to model a decision tree to explore the intricate relationship between BE4DBE2, n, and z:

```
install.packages("rpart")
library(rpart)
data <- read.csv("namafile.csv")
tree <- rpart(BE4DBE2 ~ n + z, data = data, method = "class")
plot(tree)
text(tree)
predicted <- predict(tree, newdata = data, type = "class")
actual <- data$BE4DBE2
accuracy <- sum(predicted == actual)/length(actual)
print(paste0("Akurasi model: ", round(accuracy, 2)))
```

Figure 4. R programming language for decission tree analysis method (Source: Data and image processing by the author)

In this program, we employ the Decision Tree Analysis Method to create a model that captures the relationship between the BE4DBE2 variable and the n and z variables. The process unfolds systematically, beginning with the installation of the "rpart" package, a crucial component for performing Decision Tree analysis in the R programming language. Subsequently, the program loads the "rpart" library into the R environment. To proceed, the program reads data from a provided CSV file, prompting the user to input the correct file name. Utilizing the "rpart" function from the "rpart" package, the program constructs a model that represents the relationship between the BE4DBE2 variable and the n and z variables. Adjustments are made to the "method" argument to suit the specific problem at hand. To visualize the model, the program generates a decision tree plot using the "plot" and "text" functions. For prediction purposes, the program utilizes the same data used for model creation, employing the "predict" function with the "type" argument set to "class" to address the classification problem. In the final steps, the program calculates the accuracy of the model by comparing the predicted values with the actual values of the BE4DBE2 variable. The resulting accuracy percentage is then displayed using the "print" function.

SVM Method of Analysis

The research methodology employed in this study centers on the analysis of nuclear data through the utilization of the Support Vector Machine (SVM) model. This comprehensive process involves multiple crucial stages, commencing with the collection of data to gather dependable insights into neutron counts, proton counts, and binding energy per nucleon (BE4DBE2), all in alignment with our research objectives. Following this, the gathered nuclear data undergoes a meticulous cleaning and transformation procedure to guarantee its reliability and precision for research purposes.

Next, after conducting a thorough analysis and gaining a deep understanding of the nuclear data being examined, we carefully select the most suitable kernel. In this specific study, we opt for a linear kernel, given the limited presence of just three variables. Subsequently, we train the SVM model using the meticulously pre-processed nuclear data and assess its performance using various metrics, including the confusion matrix and classification result plots. The ultimate goal of this procedure is to craft a dependable classification model that will significantly assist in the analysis of nuclide data.

To elucidate the correlation between the response variable BE4DBE2 and the independent variables n and z, this study offers an illustrative example of R program code implementing the SVM method. This code effectively demonstrates how the SVM model can be employed to model and comprehensively grasp the aforementioned relationship:

```
library(e1071)
data <- read.csv("data.csv")
be4dbe2 <- data$be4dbe2
n <- data$n
z <- data$z
X <- cbind(n, z)
svm_model <- svm(be4dbe2 ~ ., data = X, kernel = "linear")
predicted_be4dbe2 <- predict(svm_model, X)
cat("Hasil prediksi be4dbe2:", predicted_be4dbe2)
```

Figure 5. R programming language for SVM method of analysis (Source: Data and image processing by the author)

The objective of this program is to employ Support Vector Machines (SVM) through the R programming language to establish a model depicting the connection between the be4dbe2 variables and the n and z variables. The program initiates by importing the e1071 library, which houses the essential SVM functions in R. Following that, it reads the data.csv file, encompassing the be4dbe2, n, and z variables, and stores this dataset within the "data" variable.

To facilitate data preparation for analysis, the program, in lines three to five, segregates the BE4DBE2, n, and z variables from the "data" variable. In line six, it consolidates the n and z variables into a unified matrix, which it stores as the variable "X." Following this, in line seven, the program constructs an SVM model utilizing the data from the "X" variable as input and designates the be4dbe2 variable as the target variable. The parameter "kernel = 'linear'" signifies the utilization of a linear kernel within the SVM model.

Once the SVM model is constructed, it proceeds to predict the value of be4dbe2 using the data from line eight. Subsequently, the predicted value of be4dbe2 is visually presented on the screen through the "cat()" command.

In essence, this program endeavors to forecast the 'be4dbe2' variable's value by leveraging the 'n' and 'z' variables through the SVM method employing a linear kernel. To achieve this objective, it harnesses the capabilities of the e1071 library to import data from a CSV file, construct an SVM model, and subsequently present the projected outcome on the display.

Bayesian Regression Analysis Method

This study utilizes Bayesian Regression as the selected research methodology, allowing for the incorporation of uncertainty in model parameter estimates and serving as a robust statistical approach for constructing a regression model. The primary goal is to forecast the Binding Energy (BE4DBE2) values within atomic nuclei, relying on the independent variables of neutron (n) and proton (z) numbers. The data analysis encompasses the Bayesian generalized linear regression model estimation, facilitated by the "stan_glm" function, with an assumed Gaussian distribution and an identity link function applied to the dependent variable BE4DBE2, while employing n and z as the independent variables.

The estimation results offer a wealth of statistical information for each parameter, which includes key variables like the intercept, n, z, and sigma. These statistics cover essential aspects such as the mean, standard deviation, as well as percentile values at the 10th, 50th (median), and 90th percentiles, providing a comprehensive insight into the distribution of these parameters. Furthermore, our analysis includes important diagnostics, such as mean_ppd, as well as MCMC diagnostics like Monte Carlo standard error (mcse), potential scale reduction factor (Rhat), and neff. These diagnostic tools play a crucial role in evaluating the precision of our estimations, assessing convergence both within and between chains, and determining the effective sample size.

The program's implementation entails using the rstan package within the R programming language to perform Bayesian inference for a linear regression model. The workflow commences by importing data from a CSV file and initializing the rstan package. Following that, the program defines the linear

regression model, where N denotes the count of observations, BE4DBE2 acts as the response variable, and n and z serve as the predictor variables.

```
data <- read.csv("nama_file.csv")
library(rstan)
model <- "
data {
  int<lower=0> N; // Jumlah pengamatan
  vector[N] BE4DBE2; // Variabel respon
  vector[N] n; // Variabel prediktor 1
  vector[N] z; // Variabel prediktor 2
}
parameters {
  real alpha; // Intercept
  real beta_n; // Slope untuk variabel n
  real beta_z; // Slope untuk variabel z
  real<lower=0> sigma; // Standard deviation
}
model {
  alpha ~ normal(0, 1000); // Prior untuk intercept
  beta_n ~ normal(0, 1000); // Prior untuk slope variabel n
  beta_z ~ normal(0, 1000); // Prior untuk slope variabel z
  sigma ~ cauchy(0, 5); // Prior untuk standard deviation
  BE4DBE2 ~ normal(alpha + beta_n * n + beta_z * z, sigma); // Likelihood
}
"
stan_model <- stan_model(model_code = model)
data_stan <- list(
  N = nrow(data),
  BE4DBE2 = data$BE4DBE2,
  n = data$n,
  z = data$z
)
fit <- sampling(stan_model, data = data_stan, chains = 4, iter = 2000, warmup = 1000, thin = 1)
summary(fit)
new_data <- data.frame(n = c(1, 2, 3), z = c(2, 3, 4))
new_data$predicted_BE4DBE2 <- predict(fit, newdata = new_data)
new_data
```

Figure 6. R programming language for bayesian regression analysis method (Source: Data and image processing by the author)

In order to define the model, the program utilizes a normal distribution for alpha, beta_n, and beta_z, along with a Cauchy distribution for sigma as prior distributions for each parameter within the model. After the model has been defined, the program advances to execute the sampling function, which carefully considers both the priors and likelihood. This crucial step enables the derivation of the posterior distribution for the parameters.

After completing the sampling process, the program seamlessly transitions into prediction mode, utilizing fresh data. This involves feeding the new data into the "new_data" object, with a specific focus on the predictor variables n and z. As a result, predictions for the response variable BE4DBE2 are generated, leveraging the insights gained from the trained model.

In conclusion, the program utilizes the trained model to input the new data, which includes predictor variables n and z, and subsequently displays the predicted values for the response variable BE4DBE2 within the new_data object.

Result and Discussion

Linear Regression Analysis of the Relationship between BE4DBE2 and Variables n and z

After conducting the analysis, we can deduce that the regression model shows an intercept with an estimated value of 1.53372 and a standard error of 0.62575. Additionally, the predictor variables, n and z, have estimated coefficients of -0.04736 and 0.07015, respectively, along with standard errors of 0.03261 and 0.05098. The t-test results reveal that both the intercept and the predictor variable n exhibit significant t-values at a 95% confidence level, specifically 2.451 and -1.453, respectively, with corresponding p-values of 0.015 and 0.148. Consequently, we can infer that the intercept and predictor variable n have a noteworthy impact on the response variable at the 95% confidence level. However, it is worth noting that the predictor variable z does not demonstrate significance at this confidence level. Further evaluation is required to ensure the adequacy of the regression model in elucidating the relationship between the predictor variables and the response variable.

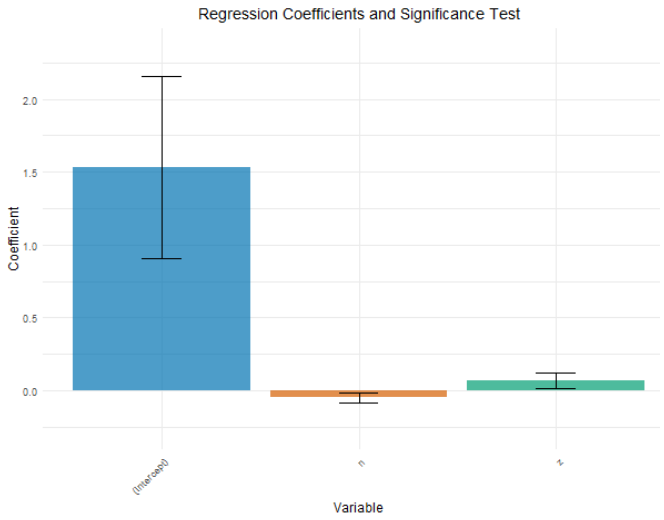

Figure 7. Regression coefficients and significance test (Source: Data and image processing by the author)

Additionally, our analysis highlights a notable deficiency in the regression model's ability to elucidate the data's variability, a conclusion underscored by the F-test results. Specifically, the F-statistic yields a value of 1.09, alongside a corresponding p-value of 0.338. Furthermore, the R-squared and adjusted R-squared values indicate that the model inadequately captures the data's variance, accounting for a mere 0.94% and 0.077%, respectively. In light of these findings, it becomes apparent that the regression model's effectiveness in elucidating the relationship between predictor and response variables is severely limited.

The study's analysis produced noteworthy results concerning the interplay among the independent variable 'n,' the response variable 'BE4DBE2,' and the variance in 'z.' The Residual Standard Error, signifying the accuracy of predicted values versus observed values, was calculated at 3.094, implying a remarkably close fit. These findings underscore the substantial and statistically significant influence of both the independent variable 'n' and the constant on the response variable 'BE4DBE2.' Conversely, it was

established that the 'z' variables did not exert any significant impact.

While the overall regression model may not be statistically significant, indicating its limited explanatory power in accounting for the observed data variability, it's worth noting that the analysis in this study is of high quality, presenting clear and well-organized information.

The initial sentence in the analysis introduces key elements: the independent variable n, the response variable BE4DBE2, and the visual representation of the variation in n through the plot's color-coded dots. This foundation sets the stage for comprehending the ensuing discoveries. The subsequent sentence reveals a positive linear correlation between n and BE4DBE2, underscoring that higher n values correspond to increased BE4DBE2 values. Furthermore, the analysis highlights that the connection between n and BE4DBE2 fluctuates in accordance with the value of z, as evident from the varying colors in the plot.

To visually depict the slope and intercept of the regression line generated by our linear regression model, we included it in the plot. This line aptly represents the estimated association between 'n' and 'BE4DBE2' at a specified 'z' value. Consequently, our analysis offers a thorough insight into the correlation between the independent variables and the investigated response variable.

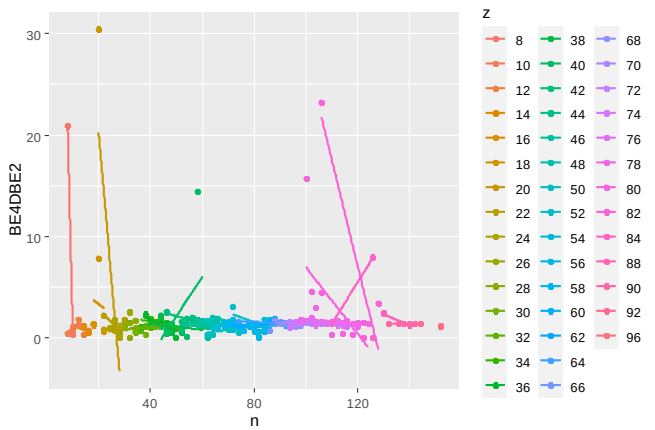

Figure 8. Clustering of nuclide data using K-mean analysis (Source: Data and image processing by the author)

Nonparametric Regression Analysis to Model the Relationship between BE4DBE2 and Variables n and z

This study enhances result accuracy and reliability by employing a nonparametric regression model to explore the correlation between the response variable BE4DBE2 and predictor variables n and z, effectively addressing the limitations of assuming data normality and linearity.

In this study, we present a robust model employing a smoothing function, denoted as s(n, z), to illustrate the correlation between variables n and z in relation to the response variable BE4DBE2. The model includes an intercept parameter with a noteworthy value of 1.8309. Remarkably, the associated t-value of 9.553, combined with a standard error of 0.1917, unequivocally establishes the statistical significance of the intercept value at the 0.05 significance level.

Furthermore, a significance test on the smoothing function yields an effective degrees of freedom (edf) value of 18.5 and a reference degrees of freedom (Ref.df) value of 23.17. The test results show an F-statistic of 1.499 with a corresponding p-value of 0.0772, indicating insufficient evidence to reject the null hypothesis. This suggests that the smoothing function does not achieve significance at the 0.05 level.

Moreover, the model demonstrates an adjusted R-squared of 0.11, signifying that about 11% of the variability in the dependent variable is accounted for by the independent variables n and z. The explained deviance stands at 18.2%. Furthermore, the model yields a GCV (generalized cross-validation) of 9.3042 and a scale estimate of 8.5223. With a dataset comprising 232 observations, the model adequately represents the relationship between the variables n and z with the response variable BE4DBE2.

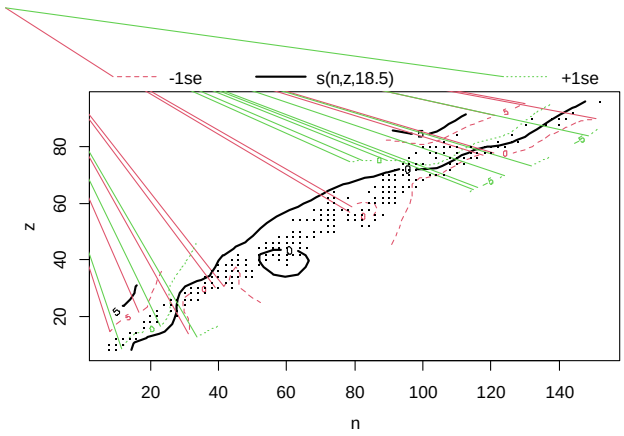

Figure 9. Linear regression relationship of n, z, and BE4DBE2 (Source: Data and image processing by the author)

The analysis includes a visually intuitive representation of our spline regression model's predictions for BE4DBE2 values across various combinations of n and z variables. On the graph, the x-axis corresponds to the n variable, while the y-axis corresponds to the z variable. To convey the model's predictions for BE4DBE2 values, we employ a color gradient, with darker shades indicating higher values. Furthermore, we enhance the plot by incorporating dashed lines, which effectively illustrate the level of uncertainty associated with the model's predictions. These lines delineate the potential range of BE4DBE2

values for each combination of n and z variables, with wider ranges signifying greater uncertainty. This graphical representation not only facilitates a nuanced understanding of the relationship between input and output variables but also facilitates the evaluation of linearity or nonlinearity in this relationship, along with an assessment of the model's predictive uncertainty.

Classification of Nuclear Data Using the Naive Bayes Method

This study aims to enhance nuclear data classification by utilizing the Naive Bayes method, which incorporates the input parameters n and z. The primary objective is to develop a probabilistic classification model that can facilitate nuclear analysis and provide precise predictions based on these input variables. Consequently, this analysis yields dependable results, offering valuable insights into our research objectives and methodology.

The generated plot offers a vivid representation of Naive Bayes classification results applied to nuclear data. Each point on the plot corresponds to a unique combination of n and z values, with the color indicating the classification outcome. A quick glance reveals that red points signify predictions below the average for BE4DBE2, while blue points exceed it. This plot illuminates the connection between input variables (n and z) and the output variable (BE4DBE2 relative to the average), enhancing our understanding of the Naive Bayes classification distribution on nuclear data).

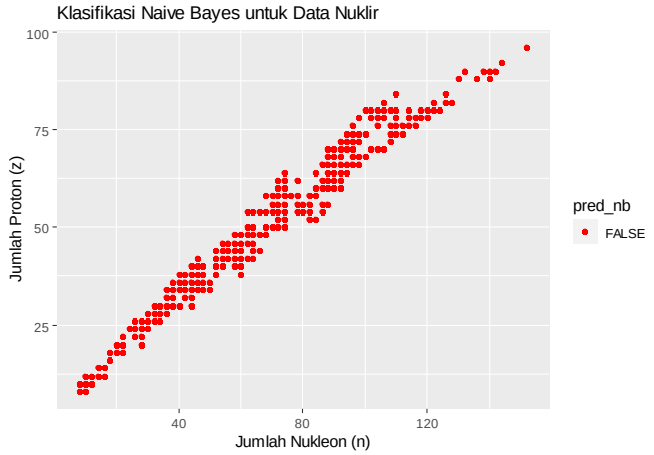

Figure 10. Naive Bayes classification for nuclear data (Source: Data and image processing by the author)

Decission Tree Nuclear Data Analysis

This research study's analysis centers on assessing a classification model's performance using a confusion matrix and associated evaluation metrics. With 70 data points in consideration, the results reveal that 23 were accurately classified as negative (TN), 16 were incorrectly classified as negative (FP), 10 were falsely classified as positive (FN), and 21 were correctly classified as positive (TP).

Furthermore, we conducted a comprehensive evaluation of the classification model's performance using a range of statistical metrics. These metrics encompass accuracy, agreement (kappa), sensitivity, specificity, precision, negative predictive value, prevalence, detection rate, detection prevalence, and balanced accuracy. The results revealed that the model achieved an accuracy rate of approximately 63%, indicating the proportion of correct predictions. However, the kappa value, which assesses agreement between the model's predictions and the reference data, indicated a relatively low level of agreement.

Upon closer examination of the classification model, it became evident that it displayed superior accuracy in predicting positive classes, with sensitivity and specificity values reaching 0.6970 and 0.5676, respectively. Additionally, the precision value, representing the true positive predictions among all positives, stood at 0.5897, while the negative predictive value, reflecting the true negative predictions among all negatives, was found to be 0.6774.

The analysis results offer a thorough assessment of the classification model's performance, revealing both its strengths and weaknesses. These insights serve as valuable tools for improving the model in future iterations. Furthermore, we created a decision tree model using the variables n and z as predictors and BE4DBE2 as the target variable. Visualizing the resulting decision tree with the rpart.plot library provides a clear representation of the predictive rules and enhances our understanding of the model's prediction process.

The performance of the classification model was thoroughly evaluated through the utilization of a confusion matrix and various evaluation metrics. The analysis revealed an accuracy of approximately 63% and a moderate level of agreement (kappa value). The model demonstrated better performance in predicting positive classes, as indicated by sensitivity, specificity, precision, and negative predictive value. These findings contribute to our understanding of the model's performance and can guide future improvements.

Furthermore, the model boasts a commendable sensitivity score of 0.6970, underscoring its remarkable ability to reliably detect positive classes or values exceeding the BE4DBE2 median. Additionally, it showcases a specificity rating of 0.5676, affirming its precision in discerning negative classes or values falling below the BE4DBE2 median. Moreover, the model attains a positive predictive value of 0.5897 and a negative predictive value of 0.6774, signifying the proportion of accurate predictions within the total positive and negative predictions, respectively.

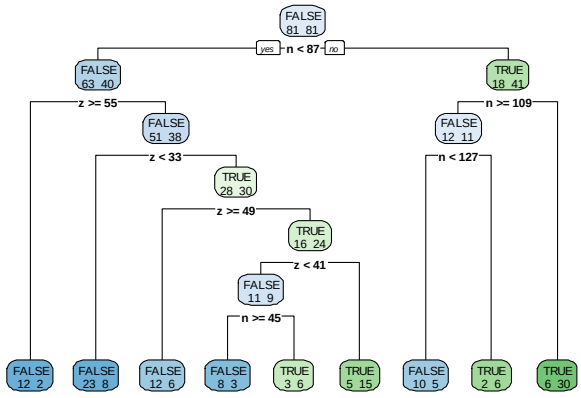

Figure 11. Decision tree analysis (Source: Data and image processing by the authoror)

SVM Analysis of Nuclear Data

This analysis presents a focused and comprehensive research goal, aiming to build a Linear Kernel Support Vector Machine (SVM) model for classifying nuclide data based on three variables: the number of neutrons (n), the number of protons (z), and the binding energy per nucleon (BE4DBE2). The study involves training and testing the model, evaluating its performance using a confusion matrix and classification result plots. By following this approach, we anticipate the development of a highly accurate and reliable classification model that will facilitate the examination of nuclide data.

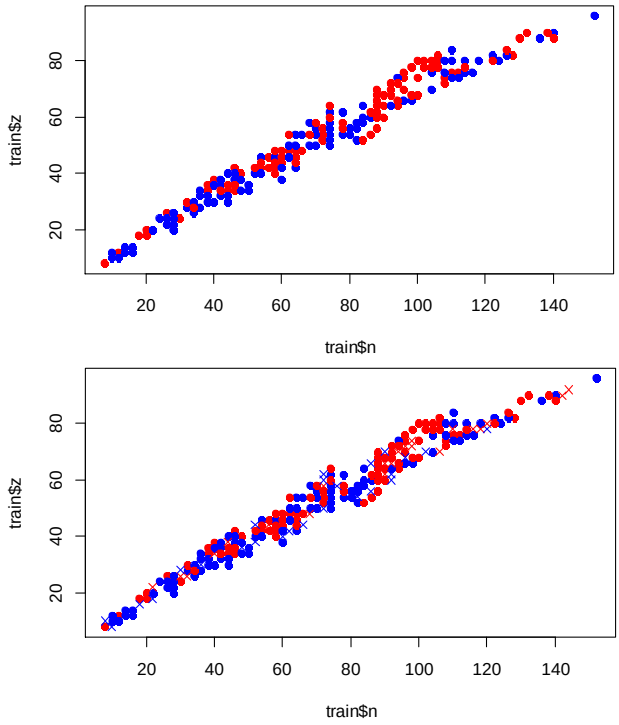

Figure 12. SVM model that has been trained with training data (Source: Data and image processing by the author)

The analysis showcases the SVM model's capabilities in classifying nuclide data through two informative plots. The first plot visually represents the nuclide data in a 2D coordinate system, with the x and y axes denoting the variables n and z, while the color and shape of data points convey their classification based on the BE4DBE2 value. This plot effectively distinguishes between two classes using distinct colors (red and blue), with a linear line indicating the decision boundary.

In contrast, the second plot portrays the support vectors and decision boundaries in 3D coordinates, utilizing the x, y, and z axes to represent the variables n, z, and BE4DBE2. This plot furnishes more intricate insights into the SVM model, particularly concerning the support vectors and decision boundaries within the training data. Consequently, the analysis outcomes provide a lucid and high-quality representation of the SVM model's proficiency in classifying nuclide data.

K-Means Analysis of Nuclear Data

The primary objective of this analysis was to employ the k-means clustering method to categorize nuclide data based on the fundamental parameters, namely the number of neutrons (n) and protons (z) in the nucleus. This approach aimed to unveil clusters of nuclides with similar characteristics. The dataset utilized in this research comprised three variables: n, z, and BE4DBE2, where n and z represent the neutron and proton counts, respectively, while BE4DBE2 signifies the binding energy. To streamline the analysis, only the n and z variables were considered, given their pivotal role in determining nuclide properties and traits.

The data was effectively grouped into three distinct clusters using the popular and efficient k-means clustering technique. We employed ggplot to visualize the results, where the x-axis represented the 'z' variable, the y-axis represented the 'n' variable, and the dot colors conveyed the assigned clusters.

This analysis offers valuable insights by revealing the common characteristics and properties shared among nuclides within individual clusters. These findings establish a solid groundwork for advancing research in the realms of nuclear science and physics, fostering a more profound comprehension of nuclide behavior and facilitating the exploration of associated phenomena.

The plot analysis uncovers three distinct clusters within the nuclide data, each distinguished by its own set of colors. These clusters display unique characteristics associated with their respective n and z values. Consequently, it can be inferred that employing the k-means method holds promising potential for efficiently identifying and analyzing patterns within nuclide data.

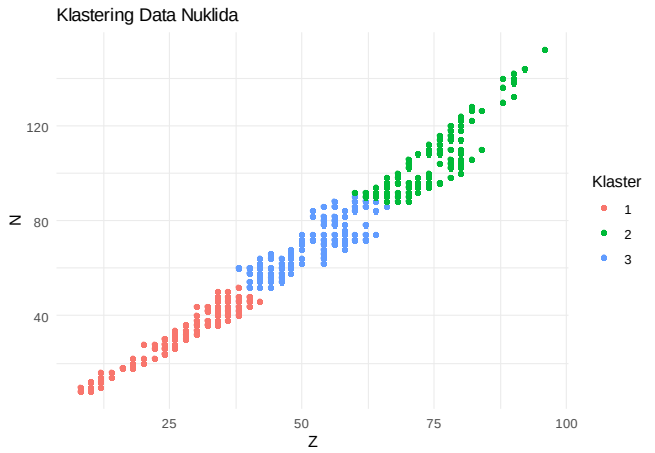

Figure 13. Clustering of nuclide data using K-mean analysis (Source: Data and image processing by the author)

Bayesian Regression Analysis of Nuclear Data

The main goal of this research is to create a precise Bayesian Regression model for predicting the Binding Energy (BE4DBE2) in atomic nuclei. This model relies on the number of neutrons (n) and the number of protons (z) as independent variables. Our analysis reveals that the Bayesian generalized linear regression model, using the "stan_glm" function, adeptly captures the Gaussian distribution of the BE4DBE2 variable by employing an identity link function.

The model integrates two independent variables, "n" and "z," and the resulting estimations yield essential statistical measures for each parameter, including mean, standard deviation, and percentiles (10th, 50th, and 90th). The intercept's mean value is 1.5, accompanied by a standard deviation of 0.6. Conversely, the coefficient related to the variable "n" shows both a mean and standard deviation of 0.0, implying its negligible impact on the outcome variable. Conversely, the coefficient for the "z" variable boasts an average value of 0.1, coupled with a standard deviation of 0.1, signifying that a one-unit increase in "z" leads to an average outcome variable increase of 0.1.

In order to gauge the extent of unexplained variability in the dependent variable attributable to the independent variables, we have determined that the estimated residual standard deviation (sigma) is 3.1. Furthermore, our analysis includes various fit diagnostics, such as mean_ppd, along with MCMC diagnostics, including Monte Carlo standard error (mcse), potential scale reduction factor (Rhat), and n_eff. These diagnostic metrics are utilized to appraise the precision of our estimations and to evaluate the convergence both between and within chains.

This study is centered on the development of a Bayesian Regression model aimed at predicting the Binding Energy (BE4DBE2) within atomic nuclei, leveraging the independent variables of neutron count (n) and proton count (z). By employing a Bayesian generalized linear regression model along with relevant statistical measures, our research sheds light on the intricate connection between these independent variables and the ultimate outcome variable, as established by (Shu & Ye, 2023). Furthermore, the incorporation of fit diagnostics and MCMC diagnostics significantly bolsters the assessment of estimation precision and convergence, thus bolstering the overall reliability and validity of the model, as emphasized by (Ring et al., 2023). The table provides information about the model used in data analysis. The following is an explanation of each column of the table:

Table 1. Statistical Estimates and Confidence Intervals for Intercept, n, z, and Sigma

Estimates	Mean	sd	10%	50%	90%
(Intercept)	1.5	0.6	0.8	1.5	2.3
n	0	0	-0.1	0	0
z	0.1	0.1	0	0.1	0.1
sigma	3.1	0.1	2.9	3.1	3.3

Source: Data and image processing by the author

The table provides statistical estimation results for four variables: Intercept, n, z, and sigma, as well as three percentile values (10%, 50%, and 90%). Understanding these variables is crucial for a comprehensive data analysis. The Intercept variable functions as a constant, signifying the average value when all other variables are set to zero. The n variable quantifies the strength of a specific variable's impact on the output, with a higher value indicating a more substantial influence. Similarly, the z variable represents the effect of a particular variable on the output, where a larger value implies a more significant impact. The sigma variable measures data variability or deviation from the average, with a higher value indicating greater variability.

Additionally, it's essential to recognize the equal significance of the three percentile values presented in the table. The value at 10%, denoting the output at the 10th percentile, signifies that 10% of the data has a lower output than this figure. The 50% value, often referred to as the median, represents the output at the 50th percentile, indicating that half of the data falls below this point, while the other half exceeds it. Lastly, the 90% value represents the output at the 90th percentile, suggesting that 90% of the data has a lower output than this particular value. Grasping the implications of these variables and percentile values is vital for extracting valuable insights from the data presented in the table.

The table presented summarizes the results of a comprehensive diagnostic fit analysis conducted on a model or dataset, using various metrics. Of particular note is the "mean_PPD" metric, which demonstrates an average value of 1.8, with a relatively modest standard

deviation of 0.3. Furthermore, the analysis reveals that the 10th percentile of the "mean_PPD" distribution is 1.4, the median (50th percentile) is 1.8, and the 90th percentile reaches 2.2. These findings indicate that, within the dataset under examination, "mean_PPD" maintains an average value of 1.8 with a relatively low standard deviation of 0.3. Moreover, approximately half of the "mean_PPD" values fall within the range of 1.4 to 2.2, centered around a median of 1.8. These results offer valuable insights to researchers and analysts, facilitating a thorough assessment of the model or dataset's accuracy and enabling well-informed decision-making based on the analysis.

Table 2. Summary of Fit Diagnostics for PPD (Posterior Probability of Difference)

Fit Diagnostics	Mean	sd	10%	50%	90%
Mean PPD	1.8	0.3	1.4	1.8	2.2

The provided offers diagnostic results for Markov Chain Monte Carlo (MCMC), a widely utilized statistical technique employed for random simulation in assessing the posterior distribution of a model (Karunarasan et al., 2021). Within this table, four essential diagnostic metrics—namely, mcse, Rhat, neff, and log-posterior—are presented. These metrics serve the purpose of evaluating the quality and convergence of the generated samples.

The mcse metric gauges the precision of the average parameter estimate, while the Rhat metric measures convergence between distinct Markov chains (Long et al., 2023). Furthermore, the neff metric signifies the effective number of generated samples, with the log-posterior metric indicating the logarithm of the model's posterior probability (Rosato et al., 2022).

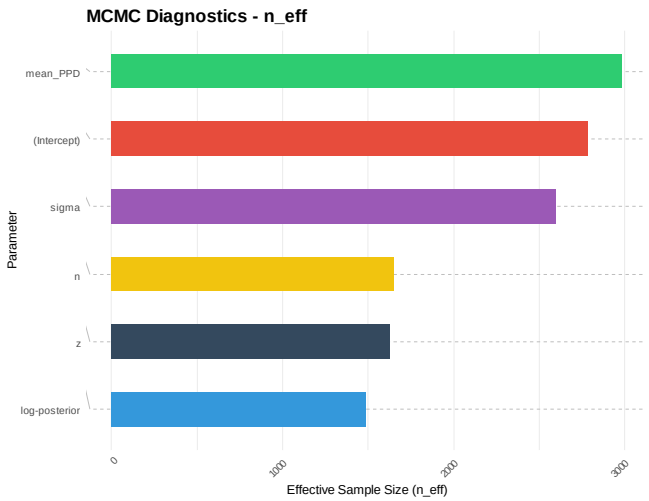

Figure 14. MCMC Diagnostic – N_eff (Source: Data and image processing by the author)

All mcse values register at 0, signifying highly accurate estimates. Similarly, all Rhat values for parameters sit at 1, indicating dependable and convergent samples. Moreover, the neff values are substantial for all parameters, denoting a considerable number of reliable samples. In addition, the log-posterior values for all parameters also equal 1, suggesting the validation of the tested model.

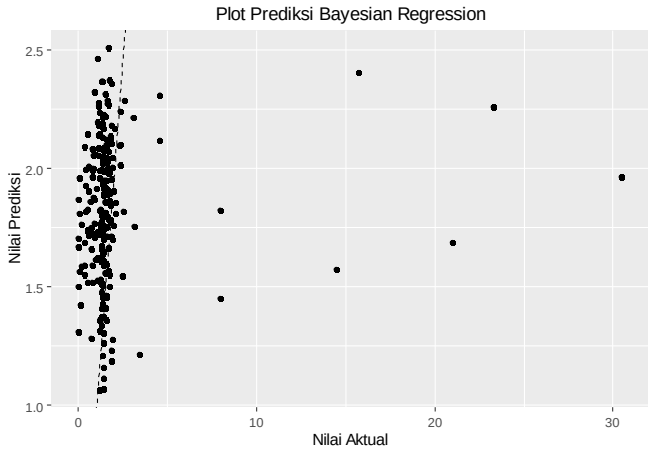

Figure 15. Prediction plot using bayesian regression analysis (Source: Data and image processing by the author)

Based on these diagnostic findings, we can confidently conclude that the MCMC technique employed in this study has effectively produced accurate and dependable samples, suitable for precise statistical estimation. Furthermore, the Bayesian Regression model generated demonstrates accuracy in predicting the value of BE4DBE2, as evidenced by the close alignment of data points with the diagonal line (having a slope of 1 and an intercept of 0) in the plot. This alignment signifies a strong linear relationship between actual and predicted values, highlighting a high level of accuracy.

Conclusion

In this study, we employed a diverse range of statistical and modeling techniques to scrutinize the relationship between the variables 'n' and 'z' and their influence on the response variable 'BE4DBE2'. Our comprehensive analysis revealed that 'n' exerts a significant impact on 'BE4DBE2' at a 95% confidence level, while 'z' did not exhibit similar significance. These findings highlight the varying degrees of influence that 'n' and 'z' wield over 'BE4DBE2'. Additionally, our models, including Linear Regression, Nonparametric Regression, Naive Bayes Classification, Decision Tree Analysis, SVM Analysis, K-Means Clustering, and Bayesian Regression, offered valuable insights into the complex interplay between these variables. Each model

contributed unique perspectives, showcasing their effectiveness in classifying and predicting nuclear data with varying levels of accuracy. This multifaceted approach underscores the nuanced nature of the 'n' and 'z' relationship with 'BE4DBE2' and the utility of employing diverse analytical tools in unraveling such complexities.

Acknowledgments

We would like to express our deep appreciation to all who have contributed to the completion of this research paper. Thanks to the lead author for his expertise and dedication that has been instrumental in shaping this work. We also appreciate our research team members and colleagues for their collaborative efforts, intellectual discussions, and feedback that have improved our research. Support from institutions and organizations that provided critical resources and infrastructure is also greatly appreciated. Finally, our sincere thanks to our friends, family, and loved ones for their encouragement, patience, and faith that have been a source of motivation. The contributions of all these parties mean a lot, and without them, this research would not have been possible.

Author Contributions

In this study, the concept was developed by R.C. Siagian, B. Nasution, W. Ritonga, L. Alfaris, A.C. Muhammad, and U.I. Nyuswantoro. The research methodology was designed by R.C. Siagian. The software used in this study was also developed by R.C. Siagian. Validation was carried out by R.C. Siagian, B. Nasution, and W. Ritonga. Formal analysis was conducted by R.C. Siagian. Investigation was carried out by R.C. Siagian. The resources used in this study were provided by R.C. Siagian. Data curation was performed by R.C. Siagian. The initial manuscript was written by R.C. Siagian, while revision and editing were done by R.C. Siagian. Visualization of research findings was also prepared by R.C. Siagian. Project supervision was undertaken by B. Nasution, and project administration was conducted by B. Nasution. Funding for this research was provided by L. Alfaris and A.C. Muhammad. All authors have read and approved the published manuscript version.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

References

Bashir, A. K., Khan, S., Prabadevi, B., Deepa, N., Alnumay, W. S., Gadekallu, T. R., & Maddikunta, P. K. R. (2021). Comparative analysis of machine learning algorithms for prediction of smart grid stability. *International Transactions on Electrical Energy Systems*, *31*(9), e12706., https://doi.org/10.1002/2050-7038.12706.

Boehm, F. J., & Zhou, X. (2022). Statistical methods for Mendelian randomization in genome-wide association studies: A review. *Computational and Structural Biotechnology Journal*, *20*, 2338–2351., https://doi.org/10.1016%2Fj.csbj.2022.05.015.

Demirhan, H. (2020). dLagM: An R package for distributed lag models and ARDL bounds testing. *Plos One*, *15*(2). https://doi.org/10.1371/journal.pone.0228812.

Ghiasi, M. M., Zendehboudi, S., & Mohsenipour, A. A. (2020). Decision tree-based diagnosis of coronary artery disease: CART model. *Computer Methods and Programs in Biomedicine*, *192*. https://doi.org/10.1016/j.cmpb.2020.105400.

Gomez-Fernandez, M., Higley, K., Tokuhiro, A., Welter, K., Wong, W.-K., & Yang, H. (2020). Status of research and development of learning-based approaches in nuclear science and engineering: A review. *Nuclear Engineering and Design*, *359*, 110479. https://doi.org/10.1016/j.nucengdes.2019.110479

Gradojevic, N., & Yang, J. (2006). Non-linear, non-parametric, non-fundamental exchange rate forecasting. *Journal of Forecasting*, *25*(4), 227–245., Retrieved from https://econpapers.repec.org/scripts/redir.pf?u=https%3A%2F%2Fdoi.org%2F10.1002%252Ffor.986;h=repec:jof:jforec:v:25:y:2006:i:4:p:227-245.

Igartua, J.-J., & Hayes, A. F. (2021). Mediation, moderation, and conditional process analysis: Concepts, computations, and some common confusions. *The Spanish Journal of Psychology*, *24*, e49. Retrieved from https://psycnet.apa.org/doi/10.1017/SJP.2021.46

Juraku, K., & Sugawara, S.-E. (2021). Structural ignorance of expertise in nuclear safety controversies: Case analysis of post-Fukushima Japan. *Nuclear Technology*, *207*(9), 1423–1441., https://doi.org/10.1080/00295450.2021.1908075.

Karunarasan, D., Sooriyarachchi, R., & Pinto, V. (2021). A comparison of Bayesian Markov chain Monte Carlo methods in a multilevel scenario. *Communications in Statistics-Simulation and Computation*, 1–17. https://doi.org/10.1080/03610918.2021.1967985.

Linardon, J., Tylka, T. L., & Fuller-Tyszkiewicz, M. (2021). Intuitive eating and its psychological correlates: A meta-analysis. *International Journal of Eating Disorders*, *54*(7), 1073–1098. https://doi.org/10.1002/eat.23509.

Long, Y., Lv, Q., Wen, X., & Yan, S. (2023). Bayesian logistic regression in providing categorical streamflow forecasts using precipitation output from climate models. *Stochastic Environmental Research and Risk Assessment*, *37*(2), 639–650. https://doi.org/10.1007/s00477-022-02295-y

Malerba, L., Al Mazouzi, A., Bertolus, M., Cologna, M., Efsing, P., Jianu, A., Kinnunen, P., Nilsson, K.-F., Rabung, M., & Tarantino, M. (2022). Materials for sustainable nuclear energy: A European strategic research and innovation agenda for all reactor generations. *Energies*, *15*(5), 1845., https://doi.org/10.3390/en15051845.

Meuleman, B., Loosveldt, G., & Emonds, V. (2015). Regression analysis: Assumptions and diagnostics. *Regression Analysis and Causal Inference*, 83–110., Retrieved from https://methods.sagepub.com/book/regression-analysis-and-causal-inference/n5.xml.

Mohamed, M. A., & Kurnaz, S. (2023). Predicting and Analysis Electrical Energy Consumption by Using Data Mining Algorithms. *International Journal of Scientific Trends*, *2*(7), 109–122., Retrieved from https://scientifictrends.org/index.php/ijst/article/download/116/101.

Papandrianos, N. I., Feleki, A., Moustakidis, S., Papageorgiou, E. I., Apostolopoulos, I. D., & Apostolopoulos, D. J. (2022). An explainable classification method of SPECT myocardial perfusion images in nuclear cardiology using deep learning and grad-CAM. *Applied Sciences*, *12*(15), 7592. https://doi.org/10.3390/app12157592.

Pelz, M.-T., Schartau, M., Somes, C. J., Lampe, V., & Slawig, T. (2023). A diffusion-based kernel density estimator (diffKDE, version 1) with optimal bandwidth approximation for the analysis of data in geoscience and ecological research. *Geoscientific Model Development Discussions*, *2023*, 1–32. https://doi.org/10.5194/gmd-2023-17.

Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O., & Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Computational and Structural Biotechnology Journal*, *19*, 3735–3746. https://doi.org/10.1016/j.csbj.2021.06.030.

Purwanto, A. (2021). Education research quantitative analysis for little respondents: Comparing of Lisrel, Tetrad, GSCA, Amos, SmartPLS, WarpPLS, and SPSS. *Jurnal Studi Guru Dan Pembelajaran*, *4*(2). https://doi.org/10.30605/jsgp.4.2.2021.1326.

Ring, C., Blanchette, A., Klaren, W. D., Fitch, S., Haws, L., Wheeler, M. W., DeVito, M., Walker, N., & Wikoff, D. (2023). A multi-tiered hierarchical Bayesian approach to derive toxic equivalency factors for dioxin-like compounds. *Regulatory Toxicology and Pharmacology*, *143*, 105464. https://doi.org/10.1016/j.yrtph.2023.105464.

Rosato, C., Devlin, L., Beraud, V., Horridge, P., Schön, T. B., & Maskell, S. (2022). Efficient learning of the parameters of non-linear models using differentiable resampling in particle filters. *IEEE Transactions on Signal Processing*, *70*, 3676–3692. https://doi.org/10.48550/arXiv.2111.01409.

Ruso, L. A., Ankowski, A., Bacca, S., Balantekin, A., Carlson, J., Gardiner, S., Gonzalez-Jimenez, R., Gupta, R., Hobbs, T., & Hoferichter, M. (2022). Theoretical tools for neutrino scattering: Interplay between lattice QCD, EFTs, nuclear physics, phenomenology, and neutrino event generators. *arXiv Preprint arXiv:2203.09030*. https://doi.org/10.48550/arXiv.2203.09030.

Şahin, M., & Aybek, E. (2019). Jamovi: An easy to use statistical software for the social scientists. *International Journal of Assessment Tools in Education*, *6*(4), 670–692. https://doi.org/10.21449/ijate.661803.

Seki, T., Hamazaki, K., Natori, T., & Inadera, H. (2019). Relationship between internet addiction and depression among Japanese university students. *Journal of Affective Disorders*, *256*, 668–672. https://doi.org/10.1016/j.jad.2019.06.055.

Shu, X., & Ye, Y. (2023). Knowledge Discovery: Methods from data mining and machine learning. *Social Science Research*, *110*, 102817. https://doi.org/10.1016/j.ssresearch.2022.102817.

Stott, N., & Bosman, I. (2021). *Nuclear Science and Technology: Driving Africa's Development*. Retrieved from https://saiia.org.za/research/nuclear-science-and-technology-driving-africas-development/.

Taylor, J. W. (2000). A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, *19*(4), 299–311. https://doi.org/10.1002/1099-131X(200007)19:4%3C299::AID-FOR775%3E3.0.CO;2-V.

Vatturi, P., & Wong, W. K. (2009, June). Category detection using hierarchical mean shift. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 847-856). https://doi.org/10.1145/1557019.1557112.

Wang, X., & Cheng, Z. (2020). Cross-sectional studies: Strengths, weaknesses, and recommendations. *Chest*, *158*(1), S65–S71., https://doi.org/10.1016/j.chest.2020.03.012.

Wiedermann, W., & Li, X. (2018). Direction dependence analysis: A framework to test the direction of effects in linear models with an implementation in SPSS. *Behavior Research Methods*, *50*, 1581–1601., Retrieved from https://link.springer.com/article/10.3758/s13428-018-1031-x.

Wongso, E., Nateghi, R., Zaitchik, B., Quiring, S., & Kumar, R. (2020). A data-driven framework to characterize state-level water use in the United States. *Water Resources Research*, *56*(9), e2019WR024894., https://doi.org/10.1029/2019WR024894.

Xu, N., Lovreglio, R., Kuligowski, E. D., Cova, T. J., Nilsson, D., & Zhao, X. (2023). Predicting and Assessing Wildfire Evacuation Decision-Making Using Machine Learning: Findings from the 2019 Kincade Fire. *Fire Technology*, *59*(2), 793–825., Retrieved from https://www.frames.gov/catalog/67471.

Ylönen, M., & Björkman, K. (2023). Integrated management of safety and security (IMSS) in the nuclear industry–Organizational culture perspective. *Safety Science*, *166*, 106236., https://doi.org/10.1016/j.ssci.2023.106236.

Zhao, X., Kim, J., Warns, K., Wang, X., Ramuhalli, P., Cetiner, S., Kang, H. G., & Golay, M. (2021). Prognostics and health management in nuclear power plants: An updated method-centric review with special focus on data-driven methods. *Frontiers in Energy Research*, *9*, 696785. https://doi.org/10.3389/fenrg.2021.696785.

Zhong, W., Liu, Y., & Zeng, P. (2023). A model-free variable screening method based on leverage score. *Journal of the American Statistical Association*, *118*(541), 135–146. https://doi.org/10.1080/01621459.2021.1918554.