Jurnal Varian
Vol. 9, No. 1, February 2026, pp 1–18
e-ISSN: 2581-2017
DOI: https://doi.org/10.30812/varian.v9i1.5767

Heart Disease Classification Using ROSE and I-CHAID with Cramér's V Bias Correction

Annurial Fitrayah Taufiq[1], Siswanto Siswanto[1], Hadijah Hadijah[2], Lilis Dwi Sapta Aprilyani[3]
[1]Universitas Hasanuddin, Makassar, Indonesia
[2]Universitas Negeri Medan, Medan, Indonesia
[3]Universitas Sam Ratulangi, Manado, Indonesia

Article Info

Article history:

Received : 09-29-2025
Revised : 02-27-2026
Accepted : 02-28-2026

Keywords:

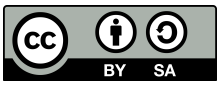
Bias Correction;
CHAID;
Classification Tree;
Cramér's V;
I-CHAID.

ABSTRACT

Machine learning applications in healthcare are increasingly important for disease classification using categorical data. The Chi-square Automatic Interaction Detection (CHAID) method is widely used, but it often produces biased results, especially with small or imbalanced datasets. To overcome this limitation, the Improved CHAID (I-CHAID) was developed by integrating bias correction on Cramér's V. Further performance gains on imbalanced data can be achieved by combining I-CHAID with the Random Oversampling Examples (ROSE) technique. This study aims to determine significant factors influencing heart disease and to evaluate the classification accuracy of the I-CHAID method with bias correction on Cramér's V. The research was conducted in two stages: (1) balancing the dataset with ROSE and (2) constructing a classification tree of heart disease occurrences using I-CHAID with bias correction. The proposed I-CHAID model correctly classified 98 individuals with heart disease and 110 without heart disease out of 253 test cases. However, 30 cases were undetected (false negatives), and 15 were misclassified (false positives). Overall, the model achieved an accuracy of 84.60%, outperforming the standard CHAID method without bias correction, which reached only 71.15%. The I-CHAID method with Cramér's V bias correction proved effective in identifying key factors associated with heart disease in Yogyakarta, including generational differences, smoking habits, and dietary patterns rich in fatty and savory foods. These findings highlight the potential of the proposed framework to support more reliable early risk identification and data-driven public health decision-making, particularly when dealing with imbalanced categorical health data.

Corresponding Author:

Siswanto Siswanto,
Department of Statistics, Universitas Hasanuddin,
Email: siswanto@unhas.ac.id



How to Cite:
Taufiq, A. F., Siswanto, S., Hadijah, H., & Aprilyani, L. D. S. (2026). Heart Disease Classification Using ROSE and I-CHAID with Cramér's V Bias Correction. *Jurnal Varian*, *9*(1), 1–18.

A. INTRODUCTION

The development of artificial intelligence (AI) technology has advanced rapidly and significantly impacted various sectors, including healthcare (Lee & Yoon, 2021). Among the diverse branches of AI, one of the most prominent is machine learning, a concept first introduced by Samuel in 1959 (Dramsch, 2020). Machine learning enables computers to learn patterns from historical data and generate predictions based on complex relationships among variables (Shu & Ye, 2023). It has been widely implemented across domains such as education, industry, government, and healthcare.

In the medical field, machine learning applications include diagnosis, health risk prediction, and data-driven treatment recommendations (Adeniran et al., 2024). Various algorithms have been developed to enhance classification performance in health data, including Neural Networks, Support Vector Machines (SVM), and Chi-Square Automatic Interaction Detection (CHAID) (Blecker

et al., 2019; Syahputri & Hasibuan, 2024). Among these methods, CHAID has been recognized as a decision tree algorithm capable of analyzing categorical data by exploring statistical associations between target and predictor variables.

The CHAID method was first introduced by Kass in 1975 (Díaz-Pérez & Bethencourt-Cejas, 2016). According to Milanović and Stamenković (2016) and Selim et al. (2024), CHAID is an exploratory technique that analyzes relationships between target and predictor variables to generate decision trees, thereby facilitating the understanding of structured associations. This method groups data according to a target variable with two or more categories and combinations of predictor variables, where the significance of Chi-Square tests determines the number of categories. However, as data complexity increases, Chi-Square-based node splitting becomes less accurate due to its susceptibility to sample-size bias (Mohammadpour et al., 2023). To address this issue, the Improved CHAID (I-CHAID) method was developed.

I-CHAID improves upon CHAID in constructing decision trees, particularly for categorical data with relatively small sample sizes and multiple categories. Its advantage lies in more precise association measures, such as Tschuprow's T for nominal variables and Cramér's V for ordinal variables (Ben-Shachar et al., 2023; Berry & Johnston, 2023). Cramér's V is especially valuable for measuring the strength of association, accounting for the number of categories, and is widely used in categorical data analysis. Nevertheless, estimation bias may arise, particularly in small samples, necessitating bias correction to improve accuracy. This correction adjusts the estimator to minimize bias, thus optimizing node splitting in I-CHAID and yielding more reliable interpretations (Khatun & Siddiqui, 2021).

Despite its advantages, I-CHAID with bias-corrected Cramér's V still faces challenges when handling imbalanced datasets. Class imbalance, particularly when the minority class accounts for less than 20% of the total data, is a significant obstacle for classification algorithms (Ahsan & Siddique, 2022). In medical datasets, imbalances often occur when the number of diseased individuals is substantially smaller than non-diseased cases. This imbalance results in models that perform well on the majority class but poorly on the minority class (Fujiwara et al., 2020). Consequently, sensitivity in detecting minority cases declines. Common techniques for handling imbalanced data include Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and Random Over-Sampling Examples (ROSE).

SMOTE and ADASYN generate synthetic samples for minority classes. However, both methods have limitations when applied to qualitative data, such as ordinal and nominal variables. SMOTE interpolates linearly between minority instances in numeric feature space, which is unsuitable for ordinal data, as it may ignore meaningful category order and generate conceptually invalid values (Fadillah et al., 2025). Similarly, ADASYN adjusts synthesis levels based on classification difficulty but relies on distance metrics that are not well-defined for categorical data (Zhang et al., 2024). Hence, these approaches may yield invalid synthetic samples for qualitative data, necessitating alternative methods aligned with categorical characteristics.

Resampling methods such as random oversampling and random undersampling can be applied to categorical data since they do not rely on numeric distance representations. However, random oversampling simply duplicates minority samples, increasing the risk of overfitting, while random undersampling reduces majority samples, potentially discarding valuable information (Aprihartha et al., 2024). This limitation highlights the need for a resampling approach that balances class distribution while preserving data diversity, without merely duplicating or deleting samples. The ROSE method addresses this by generating synthetic samples to balance class proportions (Demir & Sahin, 2022). ROSE synthesizes numerical data using kernel density estimation, producing new samples near original data points with slight random variation (Koldasbayeva et al., 2023). For categorical data, ROSE generates synthetic cases by recombining existing categorical values, creating distributions that reflect the empirical distribution of the original dataset. This enhances representativeness and reduces majority-class bias in decision tree construction (Boudegzdame et al., 2024).

Previous studies have investigated ROSE. For instance, Khushi et al. (2021) compared resampling techniques, including undersampling, oversampling, and hybrid methods, in medical datasets, and found that ROSE combined with random forest achieved superior predictive performance. Another study by Al Anshory et al. (2023) applied I-CHAID with bias correction, demonstrating improved classification accuracy of up to 73.33% on testing data. Combining ROSE with I-CHAID and bias-corrected Cramér's V is thus a promising approach to managing imbalanced health data, particularly for disease classification.

One chronic disease where positive cases are often underrepresented compared to negative cases is heart disease. Heart disease remains a leading cause of mortality in many countries, including the United Kingdom, the United States, Australia, Canada, and Indonesia (Rashid & Hossain, 2022). Its prevalence continues to rise alongside lifestyle changes (Roth et al., 2020). According to Indonesia's Basic Health Research (Riskesdas) in 2013 and 2018, the prevalence of heart disease increased from 0.5% to 1.5%. Data from the 2023 Indonesia Health Survey (SKI) showed that Yogyakarta was among the cities with the highest heart disease incidence. Risk factors include unhealthy diets, smoking habits, age, and genetics (Mensah et al., 2023). Identifying and understanding these factors is a crucial step in disease prevention and control (Agraini et al., 2025). Accordingly, classification methods are needed to

analyze factors influencing heart disease.

The gap between this research and previous studies lies in the fact that existing works have mainly examined ROSE and I-CHAID separately, without integrating them, and have used bias-corrected Cramér's V to classify imbalanced categorical health data, particularly in heart disease cases. The difference between this study and prior research is the proposed integration of the ROSE resampling method with I-CHAID, incorporating bias-corrected Cramér's V, which enables more accurate association measurement and node splitting when handling imbalanced categorical variables. The aim of this study is to contribute to the development of a more robust classification framework for imbalanced categorical health data. Accordingly, this research contributes by improving the identification of significant risk factors and enhancing classification accuracy for heart disease cases in Yogyakarta using real-world health survey data.

B. RESEARCH METHOD

1. Imbalanced Data

Imbalanced data refers to a condition in which the class distribution within a dataset is disproportionate, with one class containing significantly more samples than the other (Qadrini et al., 2022). Data imbalance becomes problematic when the proportion of the minority class falls below 20% (Thölke et al., 2023). A common example of imbalanced data can be found in disease diagnosis, where the number of detected patients is far fewer than the number of healthy individuals. Such an imbalance can introduce bias into classification models, as machine learning algorithms tend to prioritize the majority class (Ghosh et al., 2024). Although the overall model accuracy may appear high, performance in identifying minority-class instances is suboptimal (Leevy et al., 2018).

Several methods have been developed to address data imbalance, including oversampling and undersampling techniques (Wongvorachan et al., 2023). Oversampling increases the number of minority class samples by generating synthetic data, whereas undersampling reduces the number of majority class samples to achieve balance. One commonly used oversampling technique is random oversampling, which duplicates minority class samples at random until their proportion approaches that of the majority class. While this method helps balance class distribution, it carries the risk of overfitting, as the model may overly rely on duplicated data without gaining additional informative variation. To overcome this limitation, an extension of random oversampling, known as the Random Over-Sampling Examples (ROSE) method, has been introduced.

2. Random Over-Sampling Examples

The ROSE method is a technique for addressing class imbalance in datasets by synthesizing new data. This technique generates new samples through resampling based on Kernel Density Estimation (KDE). Unlike simple duplication of existing samples, ROSE creates new examples that resemble the minority class distribution, thereby reducing the risk of overfitting. The steps of the ROSE method for synthesizing data are as follows (Menardi & Torelli, 2014):

1) Identifying class imbalance: This process involves calculating the number of samples in each class and determining whether there is a significant difference between the majority and minority classes. If the minority class proportion is below 20%, ROSE can be applied to generate additional samples from the minority class to balance the data distribution. Frequent majority-class samples are removed to achieve a more balanced distribution without reducing diversity. Synthetic minority samples are generated by leveraging the empirical distribution of the original data, using the probability of each category. The minority class proportion is then increased until it reaches at least 30% of the total data to achieve balance.
2) Probability-based approach: ROSE employs the empirical distribution of categorical data to approximate the distribution of the minority class. The estimation of the conditional probability distribution is conducted based on the occurrence proportion of each category in the minority class, as expressed in Equation (1):

$$\hat{P}\left(X_q = x | Y_j\right) = \frac{\sum_{i:y_i=Y_j} 1\left(X_{iq} = x\right)}{n_j} \tag{1}$$

$\hat{P}\left(X_q = x|Y_j\right)$ is the estimated probability that the variable X_q takes the value x within class Y_j. Here, X_q denotes the q-th variable. The term $1\left(X_{iq} = x\right)$ is an indicator function, which equals 1 if $X_{iq} = x$ and 0 otherwise. Furthermore, n_j represents the number of samples in class Y_j, reflecting the estimated distribution of the variable x for that class. Finally, $\sum_{i:y_i=Y_j} 1\left(X_{iq} = x\right)$ indicates the number of observations in class Y_j that have value x for variable X_q.

DOI: https://doi.org/10.30812/varian.v9i1.5767

3) Generating synthetic data through empirical distribution sampling: Synthetic data are generated by sampling from the empirical distribution according to the previously estimated probabilities.
   a) The target class Y^* is selected randomly with probability given in Equation (2):

$$P(Y^* = Y_j) = \pi_j \tag{2}$$

   where π_j is the proportion of the class Y_j in the minority data.
   b) A Sample X^* within class Y^* is then selected based on the empirical distribution using Equation (3):

$$P(X^* \,|Y^*) = \hat{P}(X_q = x \,|Y^*) \tag{3}$$

   This means that the value of X^* is drawn randomly according to the previously estimated probability distribution of categories within the class Y^*.

3. Chi-Square Automatic Interaction Detection

The CHAID method is a type of decision tree (Gorgan-Mohammadi et al., 2023). CHAID employs the Chi-square test to identify the most significant relationship between predictor variables and the target variable (Lin & Fan, 2019). The algorithm works by splitting the data into groups based on the predictor variable categories that have the strongest association with the target variable. If certain categories of a predictor variable do not show a significant difference, those categories are merged to simplify the model (Gunduz & Al-Ajji, 2022). The CHAID method is particularly useful for exploratory analysis because it can handle predictor variables with many categories without converting them to binary form and generates a decision tree (Strzelecka & Zawadzka, 2023). CHAID determines whether a predictor variable influences the target variable using the Chi-square test (Hani & Ahmad, 2024). The hypotheses are defined as follows:

H_0 : There is no association between the predictor variable and the target variable.
H_0 : There is a significant association between the predictor variable and the target variable.

If the Chi-square test result indicates that H_0 can be rejected at the 5% ($p < 0.05$) significance level, the predictor variable is selected to split the data into groups. Otherwise, the predictor variable is not used for partitioning. The main formula in CHAID is the Chi-square test, which measures the relationship between the predictor and target variables. The Chi-square statistic is calculated using Equation (4):

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{4}$$

with

$$e_{ij} = \frac{r_i c_j}{n} \tag{5}$$

Here χ^2 follows a Chi-square distribution with degrees of freedom $(r-1)(c-1)$. o_{ij} represents the observed frequency in the row i and column j. e_{ij} is the expected frequency for row i and column i and column j. r_i denotes the total of the row i ($i = 1, 2, \ldots, r$); c_j denotes the total of the column j ($j = 1, 2, \ldots, c$); and n is the total number of observations. The larger the Chi-square value, the stronger the relationship between the predictor and the target variable. If the resulting p-value is less than the chosen threshold, the predictor variable is considered significant. Significant predictor variables are then used to form nodes in the decision tree.

4. Chi-Square Automatic Interaction Detection Algorithm

The CHAID algorithm constructs decision trees from categorical predictor variables, whether nominal or ordinal. If a continuous predictor variable is present, it must first be transformed into an ordinal variable before further processing. The algorithm consists of three main stages: category merging, node splitting, and stopping tree growth (Yang et al., 2023).

1) Merging
   At this stage, categories within a predictor variable that do not exhibit significant differences are merged to produce a simpler and more optimal model. The steps are as follows:
   a) If a predictor variable contains only one category, the process is terminated, and the adjusted p-value is set to 1.
   b) If a predictor variable has two categories, the process proceeds directly to the node-splitting stage.
   c) When there are more than two categories, the search for the most similar pair of categories is conducted. For ordinal variables, allowable pairs consist of two adjacent categories, whereas for nominal variables, merging can involve non-adjacent categories. Similarity is determined based on the pair of categories with the highest p-value with respect to the target variable.
   d) Merging is performed if the highest p-value exceeds the predetermined significance threshold. Otherwise, the process continues directly to Step 7.
   e) If the merged categories consist of three or more original categories, the best binary split within the merged categories is determined. This split is performed only if the resulting p-value is below the significance threshold.
   f) The process is repeated from Step 2 until no further merging is required.
   g) After the merging process is complete, adjusted p-values are calculated using the Bonferroni method. These values are then used in the subsequent splitting stage.
2) Splitting
   After merging the categories within the predictor variables, the next step is to select the most appropriate predictor variable to split the node. The node-splitting process consists of the following steps:
   a) The predictor variable is selected based on the smallest adjusted p-value, as it indicates the strongest association with the target variable.
   b) Node splitting is performed if the adjusted p-value is less than or equal to the significance threshold for splitting. If this condition is not met, the node is not split and becomes a terminal node (end node).
3) Stopping
   The stopping stage aims to terminate subgroup formation. This decision is made when no predictor variables show a significant effect on the target variable or when the subgroups no longer meet the requirements for the Chi-square test. If significant predictors remain, the number of observations within the resulting subgroups must be re-evaluated.

5. Improved Chi-Square Automatic Interaction Detection

The I-CHAID algorithm consists of two main stages. The first stage involves splitting the dataset into training and test sets, while the second stage involves performing a CHAID analysis using Cramér's V. In the first stage, the dataset is divided into two parts: the training data, used to construct the classification rules in the decision tree, and the test data, used to evaluate the performance of the constructed tree. In the second stage, the CHAID analysis with Cramér In the second stage, the CHAID analysis with Cramér's V is carried out through four main steps:

1) Each predictor variable is cross-tabulated with the target variable to form a two-way contingency table, and the Chi-square statistic for each variable is calculated.
2) The Cramér's V statistic is computed for each pair of categories that could potentially be merged, in order to test independence within the contingency table.
3) The best predictor variable is selected based on the lowest p-value and the highest Cramér's V value. The optimally merged categories within this predictor are then used to define new subgroups. If no predictor variable has a significant p-value, further splitting is not performed.
4) Step 1 is repeated for the subsequent subgroups. This process continues until all subgroups have been evaluated or the number of observations within a subgroup becomes too small for further analysis.

The classification process using I-CHAID follows several rules (Safitri et al., 2022):

1) Group splitting considers both the p-value obtained from the Chi-square test and the value of Cramér's V. If the p-value is greater than the Cramér's V value, the decision tree will not be generated. Lowering the significance level of the p-value results in a shorter decision tree.
2) Category merging is conducted by comparing pairs of categories within a predictor variable that exhibit similarity. Two categories are merged if their differences are not statistically significant. Statistical significance is determined by comparing

the Chi-square test p-value against the significance level defined by the researcher. Decreasing the significance level produces a larger decision tree, whereas increasing the significance level produces.

6. Cramér's V Test

The selection of the best variable in the CHAID algorithm requires a standardized measure of association to allow objective comparison of results. One of the main challenges in measuring association is that the value of χ^2 does not have a fixed upper bound, which can lead to bias in interpreting the strength of relationships between variables. Normalization of χ^2 values are therefore necessary to place them on a more controlled scale, producing a more stable measure of association and enabling fairer comparisons among variables with different numbers of categories. Cramér's V thus plays an important role in the I-CHAID algorithm as a measure of the strength of association between predictor variables and the target variable, and it helps select the best variable when splitting nodes in the decision tree. The solution to this issue is to normalize the value of ϕ^2 into the interval [0,1], which yields Equation (6) (Khatun & Siddiqui, 2021):

$$\phi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(\pi_{ij} - \pi_{i+}\pi_{+j})^2}{\pi_{i+}\pi_{+j}} \tag{6}$$

Here, π_{ij} denotes the population proportion at row i and column j, In terms of sample proportions, this can be expressed as in Equations (7) and (8):

$$\hat{\phi}^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(\frac{p_{ij}}{n} - \frac{p_{i+}}{n}\frac{p_{+j}}{n}\right)^2}{\frac{p_{i+}p_{+j}}{n}} \tag{7}$$

$$\hat{\phi}^2 = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(p_{ij} - \frac{p_{i+}p_{+j}}{n}\right)^2}{\frac{p_{i+}p_{+j}}{n}} \tag{8}$$

where p_{i+} is the marginal frequency for row i and p_{+j} is the marginal frequency for column j. Thus, $\frac{p_{i+}p_{+j}}{n}$ represents the expected frequency e_{ij}. While p_{ij} is the observed frequency at row i and column j. Consequently, Equations (9) and (10) are obtained:

$$\hat{\phi}^2 = \frac{1}{n} \times \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{9}$$

$$\hat{\phi}^2 = \frac{\chi^2}{n} \tag{10}$$

Finally, the value of $\hat{\phi}^2$ is normalized by considering the dimensions of the contingency table to obtain Cramér's V, as expressed in Equation (11):

$$V = \sqrt{\frac{\hat{\phi}^2}{\min(r-1,\ c-1)}} \tag{11}$$

where r is the number of rows and ccc is the number of columns in the contingency table. This normalization ensures that the value of Cramér's V always lies within the range of 0 to 1.

7. Cramér's V Test with Bias Correction

Tschuprow (1925) estimated the existance of bias in $\hat{\phi}^2$, which was later demonstrated by Shrivastava and Chajewska (2024) and expressed in Equation (12).

$$E\left[\chi^2\right] = \frac{n}{n-1}(r-1)(c-1) \tag{12}$$

DOI: https://doi.org/10.30812/varian.v9i1.5767

Therefore, the expectation of $\hat{\phi}^2$ can be calculated using Equation (13):

$$E\left[\hat{\phi}^2\right] = E\left[\frac{\chi^2}{n}\right] = \frac{1}{n-1}(r-1)(c-1) \tag{13}$$

Here, $E\left[\hat{\phi}^2\right]$ represents the expected bias of $\hat{\phi}^2$, n is the sample size, r is the number of rows, and c is the number of the columns in the contingency table. Tschuprow (1925) assumed that this approximation was sufficiently accurate, thereby producing a bias correction for $\hat{\phi}^2$ as given in Equation (14):

$$\widetilde{\phi}^2 = \hat{\phi}^2 - E\left[\hat{\phi}^2\right] \tag{14}$$

The value $\hat{\phi}^2$ equals zero when the tested variables are independent, since the observed frequencies in the contingency table approximate the expected frequencies. However, in small samples, $\hat{\phi}^2$ tends to be overestimated even when no association exists. Consequently, bias correction is achieved by subtracting this expectation bias, yielding Equation (14). Although this correction improves estimation accuracy, in some cases $\widetilde{\phi}^2$ may become negative, particularly when the true value is very small, or the sample size is limited. This issue can be addressed by employing a non-negative estimator, as defined in Equation (15), ensuring that results remain within the interval $[0-1]$ and relevant as a measure of association strength between categorical variables:

$$\widetilde{\phi}_+^2 = max(\hat{\phi}^2) \tag{15}$$

The value of $\widetilde{\phi}_+^2$ is always positive and not necessarily zero, meaning that it still carries some bias. A better bias correction can be achieved by substituting $\widetilde{\phi}^2$ with $\widetilde{\phi}_+^2$ in the computation of Cramér's V. For this purpose, both the number of rows and columns are adjusted according to Equations (16) and (17):

$$\widetilde{r} = \frac{1}{n-1}(r-1)^2 \tag{16}$$

$$\widetilde{c} = \frac{1}{n-1}(c-1)^2 \tag{17}$$

This correction aims to reduce bias in association estimates, particularly when sample sizes are small or the distribution of categories is unbalanced. Accordingly, the bias-corrected version of Cramér's V is expressed in Equation (18):

$$\widetilde{V} = \sqrt{\frac{\widetilde{\phi}_+^2}{\min(\widetilde{r}-1,\ \widetilde{c}-1)}} \tag{18}$$

where

$$\widetilde{\phi}_+^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(p_{ij}-p_{i+}p_{+j})^2}{p_{i+}p_{+j}} \tag{19}$$

The correction value $\widetilde{V}$ provides a more stable and accurate estimate of the association between categorical variables, especially when the number of categories is unbalanced or the sample size is small.

8. Confusion Matrix

The confusion matrix is a tabular representation used to evaluate the performance of a classification model by comparing the model's predictions with the actual values (Purwanto & Nugroho, 2023). The confusion matrix consists of four main components: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) (Amin, 2022). The TP component represents the number of cases in which the model correctly predicts the positive class. The FP component refers to the number

of cases in which the model incorrectly classifies the negative class as positive (also known as a Type I error). The TN component denotes the number of cases in which the model correctly classifies the negative class, whereas FN denotes cases in which the model misclassifies the positive class as negative (Type II error) (Vujovic, 2021). This matrix serves as the basis for calculating various evaluation metrics, including accuracy, precision, recall, and the F1-score.

Figure 1 outlines the sequence of the research methodology.

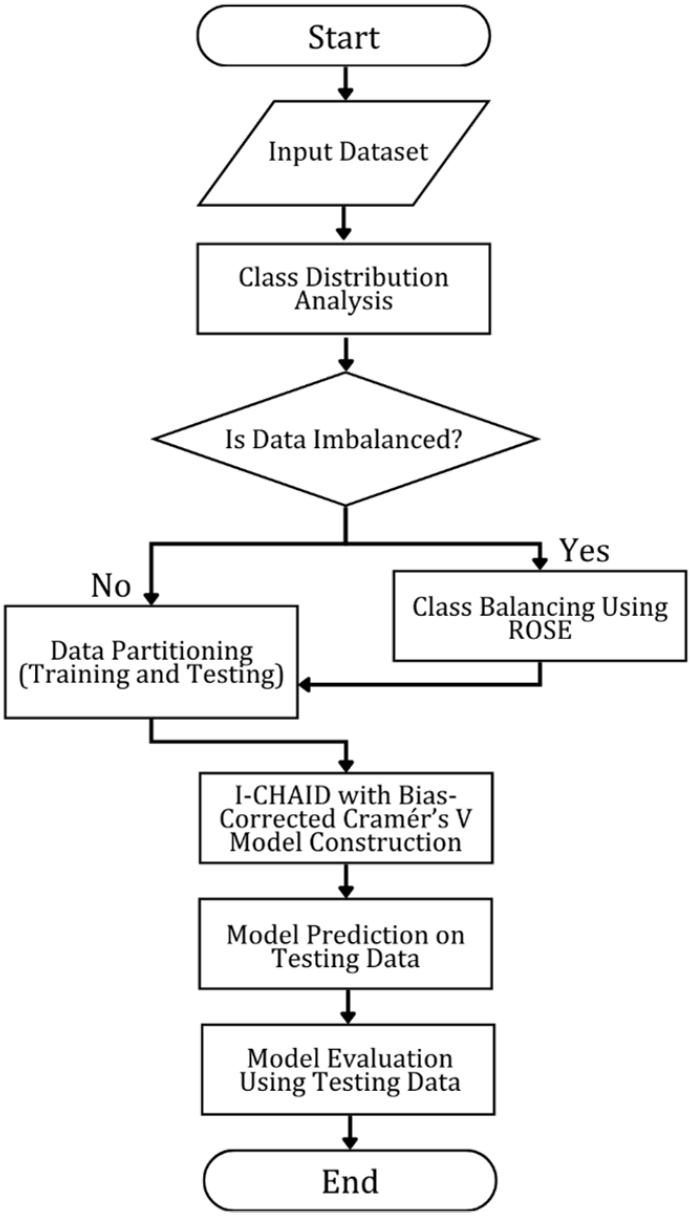


Figure 1. Research Flowchart

C. RESULT AND DISCUSSION

1. Class Distribution Analysis

The data used in this study were obtained from the SKI 2023, a large-scale nationally representative survey integrating the Riskesdas, the Toddler Nutritional Status Survey (SSGI), and biomedical examinations. The survey employs a multi-stage sampling design with explicit stratification at the census block level and implicit stratification at the household level to produce representative estimates at the regency level. In the SKI 2023 dataset for Yogyakarta City, this study applied inclusion criteria based on age, focusing on respondents aged 42 or older, corresponding to the Generation X and Baby Boomer cohorts. This restriction was applied because heart disease prevalence is substantially higher in older populations, making this group more relevant for analyzing risk factors and classification patterns. After applying these criteria, the final dataset comprised 842 respondents, including 59 with heart disease and 783 without. This corresponds to 1 case of heart disease per 14 residents. As shown in Figure 2, only 7% of respondents had a history of heart disease, while 93% reported none. This reflects an imbalanced dataset, as the minority class proportion falls below the 20% threshold. Such an imbalance may bias predictive models toward the majority class, reducing their ability to detect minority patterns. Nevertheless, the analysis was conducted, as identifying minority class characteristics is essential for improving early detection of heart disease, a leading cause of mortality.

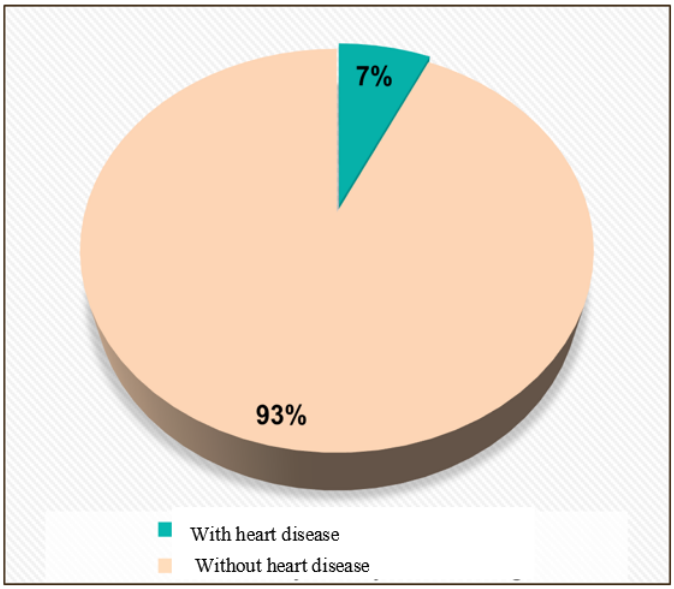


Figure 2. Proportion of Heart Disease Based on the 2023 SKI Data in Yogyakarta City

The dependent variable in this study was the presence of heart disease (Y), defined based on medical diagnosis reported in the SKI dataset and categorized dichotomously (1 = diagnosed with heart disease, 2 = not diagnosed). Independent variables included sex (X_1), classified biologically as male or female; smoking frequency (X_2), measured by self-reported habitual smoking behavior; education level (X_3), representing the highest completed level of formal schooling; employment status (X_4), indicating whether the respondent was engaged in work; generational cohort (X_5), determined by birth year and limited to Generation X and Baby Boomer groups; frequency of consuming fatty foods (X_6), and frequency of consuming foods containing flavor enhancers (X_7), both measured using ordinal categories of intake frequency.

2. Class Balancing Using Random Oversampling Examples Method

The empirical distribution is computed for each predictor variable using data from the minority class, namely, respondents with heart disease. The calculation is performed by determining the relative frequency of each categorical value for each predictor variable. The results of this calculation serve as a reference for sampling to ensure that the proportions of each category remain consistent with the original dataset.

The synthesized data from the minority class is used to randomize the majority class, which shares high similarity in characteristics. Samples from the dominant, majority class are randomly removed to reduce bias in the model. Subsequently, synthetic data is added to the minority class using the ROSE method, resulting in a more balanced class distribution without eliminating the inherent patterns in the original data, as shown in Figure 3. The application of the ROSE method resulted in a more balanced class distribution, thereby reducing model bias toward the majority class. By preserving the empirical distribution

of categorical variables, this approach allows the decision tree to better represent minority-class characteristics without distorting the original data structure.

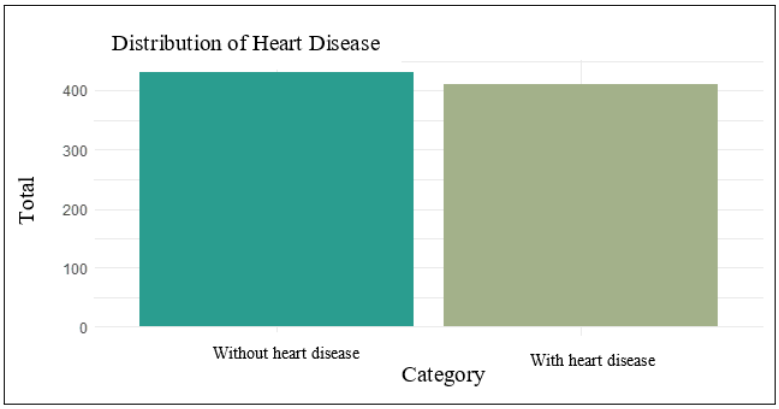


Figure 3. Distribution of Heart Disease After Applying the ROSE Method

3. Improved Chi-Square Automatic Interaction Detection Method with Bias-Corrected Cramér's V Model Construction

The Chi-Square test was applied in the merging stage. This stage involves examining the significance of each predictor variable category against the response variable prior to category merging. The merging step is only performed on variables with more than two categories; therefore, it was not applied to variables X_1, X_4, and X_5. For ordinal variables, categories that do not meet the significance threshold can be merged with adjacent categories to preserve the data hierarchy and ensure the distribution remains representative. The first step in the merging process is to construct a contingency table for each predictor variable and the response variable.

Based on the Chi-Square test results presented in Table 1, all categories of smoking frequency exhibited a significant relationship with heart disease, as the calculated χ^2 values were greater than the critical χ^2 values. Moreover, the p-values of the three categories were all below 0.05, indicating that the categories of the variable X_2 were significant at the 5% significance level. The Chi-Square calculations for each category of the other predictor variables against variable Y are also presented in Table 1.

Table 1. Results of Chi-Square Calculation between Variable Y and All Categories of Each Variable X

Variable	$\chi^2_{calculated}$	χ^2_{table}	P-value	Description
$Y \sim X_3$ Category 1	0.271	3.841	0.603	Not Significant
$Y \sim X_3$ Category 2	0.106	3.841	0.745	Not Significant
$Y \sim X_3$ Category 3	0.899	3.841	0.343	Not Significant
$Y \sim X_6$ Category 1	7.722	3.841	0.005	Significant
$Y \sim X_6$ Category 2	26.513	3.841	0.000	Significant
$Y \sim X_6$ Category 3	6.830	3.841	0.009	Significant
$Y \sim X_6$ Category 4	1.327	3.841	0.249	Not Significant
$Y \sim X_6$ Category 5	9.828	3.841	0.002	Significant
$Y \sim X_6$ Category 6	2.342	3.841	0.126	Not Significant
$Y \sim X_7$ Category 1	14.495	3.841	0.000	Significant
$Y \sim X_7$ Category 2	2.801	3.841	0.094	Not Significant
$Y \sim X_7$ Category 3	0.220	3.841	0.639	Not Significant
$Y \sim X_7$ Category 4	11.455	3.841	0.000	Significant
$Y \sim X_7$ Category 5	0.145	3.841	0.703	Not Significant
$Y \sim X_7$ Category 6	30.151	3.841	0.000	Significant

Table 1 further shows that no category within the variable X_3 was significant; therefore, this variable cannot be used in the splitting stage. Following the Chi-Square test for each predictor variable category, the non-significant categories were merged with the nearest categories within the same predictor variable. After merging, the Chi-Square test was recalculated. The results of the merging process and the corresponding Chi-Square calculations are presented in Table 2. The merged results indicated

significance across all predictor variable categories. Consequently, the splitting stage can proceed with adjustments made to the newly merged categories.

Table 2. Category Merging Based on the Chi-Square Test

Variable	Merged Category	$\chi^2_{calculated}$	p-value	Description
$Y \sim X_6$	3 and 4	11.463	0.000	Significant
$Y \sim X_6$	5 and 6	13.413	0.000	Significant
$Y \sim X_7$	1 and 2	33.743	0.000	Significant
$Y \sim X_7$	3 and 4	4.916	0.027	Significant
$Y \sim X_7$	5 and 6	24.914	0.000	Significant

The initial stage of decision tree construction involved splitting the dataset into 70% for training and 30% for testing to evaluate the model's performance. Consequently, 589 observations were allocated to the training data, while 253 were assigned to the testing data. Following this split, the decision tree was constructed using the I-CHAID method with bias correction on the training data, which involved the Chi-Square test, Cramér's V test with bias correction, and the formation of the decision tree at each node. The training data were used to build the decision tree, and the resulting model was then applied to the testing data to evaluate classification accuracy.

Table 3. Composition of Respondents Based on the Results of Node Separation Using I-CHAID with Bias Correction on Cramér's V

Nodes	Description	Number		Percentage	
		Heart Disease	Without Heart Disease	Heart Disease	Without Heart Disease
1 – 3 – 11	Generation X with a frequency of consuming fatty and seasoned foods more than once a day	54	31	67.530%	36.470%
1 – 3 – 12	Generation X with a frequency of consuming fatty foods daily and consuming seasoned foods weekly	1	6	14.290%	85.710%
1 – 3 – 13	Generation X with a frequency of consuming fatty foods daily and consuming seasoned foods monthly	2	0	100%	0%
1 – 4	Generation X with a frequency of consuming fatty foods daily	30	0	100%	0%
1 – 5 – 14	Generation X with a frequency of consuming fatty foods weekly and seasoned foods daily	59	20	76.480%	25.320%
1 – 5 – 15	Generation X with a frequency of consuming fatty foods and seasoned foods weekly	8	10	44.440%	55.560%
1 – 5 – 16	Generation X with a frequency of consuming fatty foods weekly and seasoned foods monthly	7	3	70%	30%
1 – 6	Generation X with a frequency of consuming fatty foods monthly	14	0	100%	0%
2 – 7 – 17	Baby Boomer generation with a frequency of consuming fatty and seasoned foods more than once a day	47	20	70.150%	29,.50%
2 – 7 – 18	Baby Boomer generation with a frequency of consuming fatty foods more than once a day and seasoned foods weekly	2	0	100%	0%
2 – 7 – 19	Baby Boomer generation with a frequency of consuming fatty foods more than once a day and seasoned foods monthly	1	18	5.260%	94.740%
2 – 8 – 20	Baby Boomer generation with a frequency of consuming fatty foods and a daily smoking habit	4	0	100%	0%
2 – 8 – 21	Baby Boomer generation with a frequency of consuming fatty foods daily and a non-daily smoking habit	0	5	0%	100%

DOI: https://doi.org/10.30812/varian.v9i1.5767

Nodes	Description	Number		Percentage	
		Heart Disease	Without Heart Disease	Heart Disease	Without Heart Disease
2 – 8 – 22 – 29	Baby Boomer generation with a frequency of consuming fatty foods daily, a non-daily smoking habit, and consuming fatty foods daily	16	7	69.570%	30.430%
2 – 8 – 22 – 30	Baby Boomer generation with a frequency of consuming fatty foods daily, a non-daily smoking habit, and consuming fatty foods weekly	5	0	100%	0%
2 – 8 – 22 – 31	Baby Boomer generation with a frequency of consuming fatty foods daily, a non-daily smoking habit, and consuming fatty foods monthly	2	7	22.220%	77.780%
2 – 9 – 23	Baby Boomer generation with a frequency of consuming fatty foods weekly and a daily smoking habit	14	38	26.920%	73.080%
2 – 9 – 24	Baby Boomer generation with a frequency of consuming fatty foods weekly and a non-daily smoking habit	10	0	100%	0%

Each node underwent merging, splitting, and stopping. These stages resulted in a Decision Tree with 32 nodes. Node Separation Results Using I-CHAID with Bias Correction on Cramér's V are shown in Table 3. The decision tree analysis using the I-CHAID method with bias correction, applied to heart disease data, identified several variables that significantly characterize the presence of heart disease among respondents. These variables include generation, frequency of consumption of fatty foods, frequency of consumption of food additives, and smoking habits. The constructed Decision Tree comprises 32 nodes: one root node, 31 branch nodes, and 22 terminal nodes, which serve as the final classification points.

a. First Layer

The first layer indicates that the variable that most distinguishes groups of respondents based on heart disease status is generation. Respondents belonging to the Baby Boomer generation showed a heart disease proportion of 39.950%, whereas those from Generation X had a much higher proportion of 71.430%. This significant difference demonstrates that age or generational cohort is an important initial indicator of heart disease risk.

b. Second Layer

Branching at the second level is determined by the variable of fatty food consumption frequency, particularly within the Baby Boomer group. Interestingly, individuals who consumed fatty foods frequently, such as more than once per day or 3–6 times per week, had a high proportion of heart disease cases, exceeding 69%. Although the number of cases at this node was relatively small, this finding suggests that among Baby Boomers, the risk of heart disease remains high regardless of fatty food consumption frequency, implying that age and possible long-term risk accumulation should be taken into account. Meanwhile, among Generation X, a clear trend emerges: the less frequently fatty foods are consumed, the lower the proportion of heart disease cases. For example, among respondents who consumed fatty foods fewer than three times per month, the proportion of heart disease cases dropped sharply to 16.420%.

c. Third Layer

The third layer of the decision tree highlights the role of seasoning consumption frequency as an additional node-splitting variable. For instance, respondents who frequently consumed both fatty foods (>1 time per day) and flavor enhancers (>1 time per day) exhibited a heart disease proportion of 63.530%, while those who consumed flavor enhancers 3–6 times per week had a much lower proportion of 14.290%. This finding illustrates that the intensity of consuming additives such as flavor enhancers shows a higher proportion of respondents with heart disease, particularly when combined with a high-fat dietary pattern.

d. Fourth Layer

The fourth and deepest layer of the decision tree structure incorporates smoking habits as an additional node-splitting variable. Interestingly, in one node, it was found that even respondents who had never smoked still exhibited a high heart disease proportion of 62.16%. This suggests that smoking is not the sole major risk factor; rather, there are complex interactions with other factors, such as high-fat dietary patterns and generational cohort (age group). Conversely, respondents

who smoked, whether daily or occasionally, demonstrated varying levels of risk depending on the context of the preceding node characteristics. Overall, the layer-by-layer analysis of the decision tree shows that factors such as generation, smoking habits, and consumption of fatty and seasoned foods play significant roles in determining heart disease risk.

The results of this study are in line with or supported by previous research demonstrating that demographic characteristics and lifestyle behaviors play a central role in heart disease classification models (Das et al., 2023). Prior decision tree–based investigations consistently identified age-related attributes as dominant predictors, driven by cumulative cardiovascular risk exposure over time. In agreement with these findings, the present I-CHAID model selected the generational cohort as the root node, indicating its strongest discriminative power in separating respondents by heart disease status. This reinforces the theoretical expectation that age-linked exposure remains a fundamental stratification mechanism in cardiovascular risk modeling.

Similarly, earlier studies have highlighted dietary patterns and smoking behavior as influential lifestyle predictors within classification frameworks (Das et al., 2023). The current analysis supports this perspective, as fatty food consumption frequency and seasoned food intake emerged as significant branching variables in the second and third layers of the tree. These results suggest that behavioral variables interact with demographic factors to shape subgroup risk patterns. However, a contextual distinction was observed: smoking behavior appeared only at deeper levels of the hierarchy, whereas some clinical-based decision tree models report smoking as an earlier splitting factor. This difference indicates that predictor importance may shift depending on dataset composition and the nature of available features.

A comparison between the study from Das et al. (2023) and prior decision tree research reveals both conceptual alignment and methodological differences. Both approaches demonstrate the multifactorial nature of heart disease classification, where outcomes are influenced by interacting demographic and behavioral predictors rather than single variables. However, differences arise in the predictor hierarchy and data context. Previous studies generally relied on clinical and physiological attributes as primary splitting variables, whereas the current analysis employed population-based lifestyle indicators derived from survey data. Consequently, the resulting model emphasizes behavioral segmentation rather than prioritizing clinical measurement.

Furthermore, the I-CHAID model generated a relatively detailed hierarchical structure comprising 32 nodes and 22 terminal nodes, enabling nuanced subgroup characterization by generational cohort, dietary patterns, and smoking behavior. In contrast, earlier decision tree models typically prioritized biomedical indicators and therefore produced branching structures centered on clinical risk markers. These distinctions highlight how dataset composition and feature availability shape the interpretability and focus of classification outcomes, even when similar analytical techniques are applied.

4. Model Evaluation Using Testing Data

To assess the predictive performance of the proposed model, an evaluation was conducted on the test dataset using a confusion matrix. This approach measures the classification accuracy of heart disease prediction by the bias-corrected I-CHAID model using Cramér's V. The confusion matrix calculated from the test data is presented in Table 4.

Table 4. Confusion Matrix of Data Test

Actual/Prediction	Positive	Negative
Positive	93	16
Negative	23	121

$$\text{Accuracy} = \frac{\text{The ratio of correct predictions}}{\text{The total number of predictions}} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Accuracy} = \frac{93 + 121}{93 + 16 + 121 + 23} = \frac{214}{253} = 0.846$$

The I-CHAID classification model with bias correction, using Cramér's V, correctly identified 98 individuals with heart disease and 110 without. However, there were still 30 false negatives and 15 false positives. Overall, the number of correct predictions reached 208 out of 253 observations in the test data. A comparison of the accuracy between the I-CHAID and CHAID methods is presented in Table 5.

Table 5. Comparison of Accuracy between I-CHAID and CHAID

Algorithm	Accuracy
I-CHAID	84.600%
CHAID	71.150%

The accuracy of 84.600% indicates how often the model's predictions match the actual outcomes in the test data. In comparison, the accuracy of the CHAID model without bias correction only reached 71.150%. This indicates that the bias correction in the I-CHAID method successfully improves the classification accuracy for heart disease cases in Yogyakarta City in 2023.

D. CONCLUSION AND SUGGESTION

This study demonstrates that generational differences, smoking frequency, consumption of fatty foods, and intake of foods containing flavor enhancers are significant factors associated with heart disease incidence in Yogyakarta City when analyzed using the I-CHAID method with bias correction on Cramér's V. The proposed model achieved an accuracy of 84.600%, correctly classifying most cases in the testing data and outperforming the standard CHAID approach, which yielded only 71.150% accuracy. This improvement highlights the novelty of integrating ROSE-based class balancing with bias-corrected association measurement in the I-CHAID framework, enabling more reliable node splitting and better identification of relevant predictors in imbalanced categorical health data. The findings imply that methodological refinement in handling imbalance and bias can enhance the interpretability and predictive reliability of classification models used in public health decision support.

This study demonstrates that generational differences, smoking frequency, consumption of fatty foods, and intake of foods containing flavor enhancers are significant factors associated with heart disease incidence in Yogyakarta City when analyzed using the I-CHAID method with bias correction on Cramér's V. The proposed model achieved an accuracy of 84.600%, correctly classifying most cases in the testing data and outperforming the standard CHAID approach, which yielded only 71.150% accuracy. This improvement highlights the novelty of integrating ROSE-based class balancing with bias-corrected association measurement in the I-CHAID framework, enabling more reliable node splitting and better identification of relevant predictors in imbalanced categorical health data. The findings imply that methodological refinement in handling imbalance and bias can enhance the interpretability and predictive reliability of classification models used in public health decision support.

ACKNOWLEDGEMENT

The author sincerely extends gratitude to all individuals and institutions who provided support and assistance throughout this research. The author hopes that the findings of this study will offer meaningful insights and advance scientific knowledge.

DECLARATIONS

AUTHOR CONTRIBUTION

The author sincerely extends gratitude to all individuals and institutions who provided support and assistance throughout this research. The author hopes that the findings of this study will offer meaningful insights and advance scientific knowledge.

FUNDING STATEMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

COMPETING INTEREST

The authors declare that there are no financial or personal conflicts of interest that could have influenced the outcomes of this study.

REFERENCES

Adeniran, I. A., Efunniyi, C. P., Osundare, O. S., & Abhulimen, A. O. (2024). Data-driven decision-making in healthcare: Improving patient outcomes through predictive modeling. *International Journal of Scholarly Research in Multidisciplinary Studies*, *5*(1), 059–067. https://doi.org/10.56781/ijsrms.2024.5.1.0040

Agraini, A., Fitriana, E., Saquro, A., & Karwiti, W. (2025). Pemberdayaan Masyarakat Dalam Pengendalian Risiko Penyakit Jantung di Desa Penyengat Olak Kabupaten Muaro Jambi. *Jurnal Pengabdian Meambo*, *4*(1), 1–7. https://doi.org/10.56742/jpm.v4i1.102

Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, *128*, 102289. https://doi.org/10.1016/j.artmed.2022.102289

Al Anshory, F., Siswanto, S., Thamrin, S. A., & Inayah, I. (2023). Improved Chi Square Automatic Interaction Detection on Students Discontinuation to Secondary School. *Jurnal Varian*, *7*(1), 15–26. https://doi.org/10.30812/varian.v7i1.2627

Amin, M. F. (2022). Confusion Matrix in Binary Classification Problems: A Step-by-Step Tutorial. *Journal of Engineering Research*, *6*(5), 0–0. https://doi.org/10.21608/erjeng.2022.274526

Aprihartha, M. A., Putrawan, Z., Zulhan, D., & Nurfaizal, F. A. (2024). Algoritma Synthetic Minority Oversampling Technique dan C5.0 dalam Mengatasi Ketidakseimbangan Data pada Klasifikasi Kelulusan Siswa. *UPGRADE : Jurnal Pendidikan Teknologi Informasi*, *2*(1), 1–10. https://doi.org/10.30812/upgrade.v2i1.4148

Ben-Shachar, M. S., Patil, I., Thériault, R., Wiernik, B. M., & Lüdecke, D. (2023). Phi, Fei, Fo, Fum: Effect Sizes for Categorical Data That Use the Chi-Squared Statistic. *Mathematics*, *11*(9), 1982. https://doi.org/10.3390/math11091982

Berry, K. J., & Johnston, J. E. (2023). Measures of Nominal Association II. In *Statistical Methods: Connections, Equivalencies, and Relationships* (pp. 559–632). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-41896-9_12

Blecker, S., Austrian, J. S., Horwitz, L. I., Kuperman, G., Shelley, D., Ferrauiola, M., & Katz, S. D. (2019). Interrupting providers with clinical decision support to improve care for heart failure. *International Journal of Medical Informatics*, *131*, 103956. https://doi.org/10.1016/j.ijmedinf.2019.103956

Boudegzdame, N., Sedki, K., Tspora, R., & Lamy, J.-B. (2024). An Approach for Improving Oversampling by Filtering out Unrealistic Synthetic Data: *Proceedings of the 16th International Conference on Agents and Artificial Intelligence*, 291–298. https://doi.org/10.5220/0012325400003636

Das, S., Sultana, M., Bhattacharya, S., Sengupta, D., & De, D. (2023). XAI–reduct: Accuracy preservation despite dimensionality reduction for heart disease classification using explainable AI. *The Journal of Supercomputing*, *79*(16), 18167–18197. https://doi.org/10.1007/s11227-023-05356-3

Demir, S., & Sahin, E. K. (2022). Evaluation of Oversampling Methods (OVER, SMOTE, and ROSE) in Classifying Soil Liquefaction Dataset based on SVM, RF, and Naïve Bayes. *European Journal of Science and Technology*, *34*, 142–147. https://doi.org/10.31590/ejosat.1077867

Díaz-Pérez, F. M., & Bethencourt-Cejas, M. (2016). CHAID algorithm as an appropriate analytical method for tourism market segmentation. *Journal of Destination Marketing & Management*, *5*(3), 275–282. https://doi.org/10.1016/j.jdmm.2016.01.006

Dramsch, J. S. (2020). 70 years of machine learning in geoscience in review. In *Advances in Geophysics* (pp. 1–55, Vol. 61). Elsevier. https://doi.org/10.1016/bs.agph.2020.08.002

Fadillah, D., Haerani, E., Wulandari, F., & Syafria, F. (2025). Klasifikasi Kondisi Janin Menggunakan Algoritma K-Nearest Neighbors dan Teknik SMOTE Berdasarkan Data Kardiotogram. *Bulletin of Computer Science Research*, *5*(4), 482–489. https://doi.org/10.47065/bulletincsr.v5i4.585

Fujiwara, K., Huang, Y., Hori, K., Nishioji, K., Kobayashi, M., Kamaguchi, M., & Kano, M. (2020). Over- and Under-sampling Approach for Extremely Imbalanced and Small Minority Data Problem in Health Record Analysis. *Frontiers in Public Health*, *8*, 178. https://doi.org/10.3389/fpubh.2020.00178

Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B., & Japkowicz, N. (2024). The class imbalance problem in deep learning. *Machine Learning*, *113*(7), 4845–4901. https://doi.org/10.1007/s10994-022-06268-8

Gorgan-Mohammadi, F., Rajaee, T., & Zounemat-Kermani, M. (2023). Decision tree models in predicting water quality parameters of dissolved oxygen and phosphorus in lake water. *Sustainable Water Resources Management*, *9*(1), 1. https://doi.org/10.1007/s40899-022-00776-0

Gunduz, M., & Al-Ajji, I. (2022). Employment of CHAID and CRT decision tree algorithms to develop bid/no-bid decision-making models for contractors. *Engineering, Construction and Architectural Management*, *29*(9), 3712–3736. https://doi.org/10.1108/ECAM-01-2021-0042

Hani, S. B., & Ahmad, M. (2024). Predicting mortality amongst Jordanian men with heart attacks using the chi-square automatic interaction detection model. *Health Informatics Journal*, *30*(3), 14604582241270830. https://doi.org/10.1177/14604582241270830

Khatun, M., & Siddiqui, S. (2021). Testing pairs of continuous random variables for independence: A simple heuristic. *Journal of Computational Mathematics and Data Science*, *1*, 100012. https://doi.org/10.1016/j.jcmds.2021.100012

Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., Yang, X., & Reyes, M. C. (2021). A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. *IEEE Access*, *9*, 109960–109975. https://doi.org/10.1109/ACCESS.2021.3102399

Koldasbayeva, D., Tregubova, P., Gasanov, M., Zaytsev, A., Petrovskaia, A., & Burnaev, E. (2023). *Challenges in data-based geospatial modeling for environmental research and practice* (1). https://doi.org/10.48550/ARXIV.2311.11057

Lee, D., & Yoon, S. N. (2021). Application of Artificial Intelligence-Based Technologies in the Healthcare Industry: Opportunities and Challenges. *International Journal of Environmental Research and Public Health*, *18*(1), 271. https://doi.org/10.3390/ijerph18010271

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*, *5*(1), 42. https://doi.org/10.1186/s40537-018-0151-6

Lin, C.-L., & Fan, C.-L. (2019). Evaluation of CART, CHAID, and QUEST algorithms: A case study of construction defects in Taiwan. *Journal of Asian Architecture and Building Engineering*, *18*(6), 539–553. https://doi.org/10.1080/13467581.2019.1696203

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, *28*(1), 92–122. https://doi.org/10.1007/s10618-012-0295-5

Mensah, G. A., Fuster, V., Murray, C. J., & Roth, G. A. (2023). Global Burden of Cardiovascular Diseases and Risks, 1990-2022. *JACC*, *82*(25), 2350–2473. https://doi.org/10.1016/j.jacc.2023.11.007

Milanović, M., & Stamenković, M. (2016). CHAID Decision Tree: Methodological Frame and Application. *Economic Themes*, *54*(4), 563–586. https://doi.org/10.1515/ethemes-2016-0029

Mohammadpour, S. I., Khedmati, M., & Zada, M. J. H. (2023). Classification of truck-involved crash severity: Dealing with missing, imbalanced, and high dimensional safety data (G. Li, Ed.). *PLOS ONE*, *18*(3), e0281901. https://doi.org/10.1371/journal.pone.0281901

Purwanto, A., & Nugroho, H. W. (2023). Analisa Perbandingan Kinerja Algoritma C4.5 dan Algoritma K-Nearest Neighbors untuk Klasifikasi Penerima Beasiswa. *Jurnal Teknoinfo*, *17*(1), 236. https://doi.org/10.33365/jti.v17i1.2370

Qadrini, L., Hikmah, H., & Megasari, M. (2022). Oversampling, Undersampling, Smote SVM dan Random Forest pada Klasifikasi Penerima Bidikmisi Sejawa Timur Tahun 2017. *Journal of Computer System and Informatics (JoSYC)*, *3*(4), 386–391. https://doi.org/10.47065/josyc.v3i4.2154

Rashid, S. M. A., & Hossain, S. M. (2022). Stroke and Coronary Heart Diseases, Global and Asian Trend and Risk Factors -A Perspective. *Medicine Today*, *34*(1), 27–35. https://doi.org/10.3329/medtoday.v34i1.58671

Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., Barengo, N. C., Beaton, A. Z., Benjamin, E. J., Benziger, C. P., Bonny, A., Brauer, M., Brodmann, M., Cahill, T. J., Carapetis, J., Catapano, A. L., Chugh, S. S., Cooper, L. T., Coresh, J., … Fuster, V. (2020). Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019. *Journal of the American College of Cardiology*, *76*(25), 2982–3021. https://doi.org/10.1016/j.jacc.2020.11.010

Safitri, S. N., Haryono Setiadi, & Suryani, E. (2022). Educational Data Mining Using Cluster Analysis Methods and Decision Trees based on Log Mining. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *6*(3), 448–456. https://doi.org/10.29207/resti.v6i3.3935

Selim, A., Ali, I., Saracevic, M., & Ristevski, B. (2024). Application of the digital twin model in higher education. *Multimedia Tools and Applications*, *84*(21), 24255–24272. https://doi.org/10.1007/s11042-024-20014-3

Shrivastava, H., & Chajewska, U. (2024). Methods for Recovering Conditional Independence Graphs: A Survey. *Journal of Artificial Intelligence Research*, *80*, 593–612. https://doi.org/10.1613/jair.1.14676

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

Strzelecka, A., & Zawadzka, D. (2023). The use of Chi-squared Automatic Interaction Detector (CHAID) analysis to identify characteristics of agricultural households at risk of financial self-exclusions. *Procedia Computer Science*, *225*, 4443–4452. https://doi.org/10.1016/j.procs.2023.10.442

Syahputri, C. N., & Hasibuan, M. S. (2024). Optimasi Klasifikasi Decision Tree dengan Teknik Pruning untuk Mengurangi Overfitting. *JSiI (Jurnal Sistem Informasi)*, *11*(2), 87–96. https://doi.org/10.30656/jsii.v11i2.9161

Thölke, P., Mantilla-Ramos, Y.-J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Mekki Berrada, L., Sahraoui, M., Young, T., Bellemare Pépin, A., El Khantour, C., Landry, M., Pascarella, A., Hadid, V., Combrisson, E., O'Byrne, J., & Jerbi, K. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage*, *277*, 120253. https://doi.org/10.1016/j.neuroimage.2023.120253

Vujovic, Ž. Đ. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *12*(6). https://doi.org/10.14569/IJACSA.2021.0120670

Wongvorachan, T., He, S., & Bulut, O. (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information*, *14*(1), 54. https://doi.org/10.3390/info14010054

Yang, Y., Yi, F., Deng, C., & Sun, G. (2023). Performance Analysis of the CHAID Algorithm for Accuracy. *Mathematics*, *11*(11), 2558. https://doi.org/10.3390/math11112558

Zhang, H., Wang, J., & Zhu, W. (2024). Modeling consistency and consensus in social network group decision making: The role of limited dual tolerance and compromise behaviors. *Applied Soft Computing*, *166*, 112130. https://doi.org/10.1016/j.asoc.2024.112130

[This page is intentionally left blank.]

Vol. 9, No. 1, February 2026, pp 1–18
DOI: https://doi.org/10.30812/varian.v9i1.5767