**Journal of Information System Research (JOSH)**
Volume 6, No. 4, Juli 2025, pp 1938−1947
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7216

# Optimizing Insurance Customer Segmentation with C4.5 Decision Tree Algorithm

**Sigit Candra Setya[1], Moch. Iswan Perangin-angin[2], Marsono[3], Asyahri Hadi Nasyuha[4,*], Lucia Nugraheni Harnaningrum[5]**

[1]Teknik Informatika, Institut Teknologi Pagar Alam, Pagar Alam
Jl. Masik Siagim No.75, Karang Dalo, Kec. Dempo Tengah, Kota Pagar Alam, Sumatera Selatan, Indonesia
[2]Fakultas Ilmu Komputer Dan Teknologi Informasi, Teknik Informatika, Universitas Budi Darma, Medan
Jl. Sisingamangaraja No.338, Siti Rejo I, Kec. Medan Kota, Kota Medan, Sumatera Utara, Indonesia
[3]Sistem Informasi, STMIK Triguna Dharma, Medan
Jl. Jenderal Besar A.H. Nasution No.73, Kwala Bekala, Kec. Medan Johor, Kota Medan, Sumatera Utara, Indonesia
[4]Fakultas Teknologi Informasi, Sistem Informasi, Universitas Teknologi Digital Indonesia, Yogyakarta
Jl. Majapahit No.143, Jaranan, Banguntapan, Kec. Banguntapan, Kabupaten Bantul, Daerah Istimewa Yogyakarta, Indonesia
[5]Fakultas Teknologi Informasi, Informatika, Universitas Teknologi Digital Indonesia, Yogyakarta
Jl. Majapahit No.143, Jaranan, Banguntapan, Kec. Banguntapan, Kabupaten Bantul, Daerah Istimewa Yogyakarta, Indonesia
Email: [1]sigitcs1969@gmail.com, [2]mochammadiswan@gmail.com, marsonotgdsi.@gmail.com
[4,*]asyahrihadi@gmail.com, [5]ningrum@utdi.ac.id
Correspondence Author Email: asyahrihadi@gmail.com
Submitted: **23/04/2025;** Accepted: **19/07/2025;** Published: **20/07/2025**

**Abstract**−Insurance companies rely on premium payments as their primary source of revenue. However, economic instability often causes delays in premium payments, impacting revenue recording. This study applies the C4.5 Decision Tree algorithm to classify insurance customers based on premium amount, age, income, and claim history, thereby improving product recommendations. The research utilizes data mining techniques to analyze customer attributes and generate decision rules for optimal insurance product selection. The findings indicate that customers with a premium of IDR 500,000 are best suited for PRUMed Cover (PMC), while those with IDR 1,000,000 are recommended PRUCritical Benefit 88 (PCB88). For customers with IDR 750,000, additional factors such as age and income level influence the recommended insurance type. The entropy and information gain calculations identify premium amount as the most significant attribute for decision-making, followed by age, income, and claim history. By implementing this method, insurance companies can enhance customer segmentation, streamline product selection, and optimize marketing strategies. The transparent and interpretable decision tree structure ensures regulatory compliance while improving customer satisfaction. Future research should explore additional variables, such as behavioral data and regional trends, and compare C4.5 with other classification algorithms like Random Forest or Support Vector Machines (SVM) to enhance accuracy and scalability.

**Keywords**: Data Mining; Decision Tree; C4.5 Algorithm; Insurance Risk Classification; Customer Segmentation

## 1. INTRODUCTION

The insurance industry is characterized by intense competition, requiring companies to constantly develop and refine strategies to ensure the sustainability and growth of their business[1]. In this context, one of the most critical aspects of maintaining a competitive edge is the effective use of data. Insurance companies accumulate vast amounts of data over the years, including customer demographics, purchase histories, transaction records, and policy details. However, these datasets are often underutilized, remaining in storage without being leveraged to their full potential. Traditionally, these data have been used primarily for documentation purposes or basic statistical analysis. As the industry evolves, there is a pressing need to extract greater value from this data, transforming it into actionable insights that can support strategic decision-making. This has led to the rapid adoption of data mining techniques in the insurance sector. Customer classification is a crucial aspect of the insurance business, impacting risk assessment, pricing strategies, and fraud detection. Additionally, regulatory compliance issues often arise due to inaccurate classification models, leading to hefty penalties and reputational damage. Currently, most insurance companies rely on rule-based systems and basic statistical models for customer classification. These methods are often rigid and fail to adapt to complex, non-linear relationships in customer data. Furthermore, manual intervention in the classification process introduces biases and inefficiencies, making it challenging to personalize insurance offerings effectively.

Compared to these studies, our research implements the C4.5 Decision Tree algorithm, which offers better interpretability and handles both categorical and continuous attributes efficiently. Unlike Naïve Bayes, C4.5 does not assume feature independence and provides a structured decision-making process that can be easily visualized and interpreted. Additionally, in contrast to KNN, which requires extensive computational resources for large datasets, C4.5 efficiently constructs a decision tree that reduces classification complexity. The role of data mining in insurance extends beyond customer classification. Studies have explored its applications in fraud detection, claim prediction, and underwriting risk assessment. Gupta et al. (2021) utilized a Random Forest algorithm to detect fraudulent claims and achieved an 84% fraud detection rate[2], significantly reducing financial losses for insurance providers. Y.Han et al. (2025) leveraged Support Vector Machines (SVM) for health insurance premium optimization, demonstrating a 10% improvement in premium adjustment accuracy compared to traditional actuarial methods[3]. Despite these advancements, there is still a need for a more transparent and adaptable classification framework that can support

**Journal of Information System Research (JOSH)**
Volume 6, No. 4, Juli 2025, pp 1938−1947
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7216

real-time decision-making. This study aims to address this gap by exploring the practical implementation of the C4.5 Decision Tree algorithm for insurance product classification.

Moreover, the application of data mining techniques aligns with the growing trend of data-driven decision-making in the insurance industry. As digital transformation continues to reshape the industry, insurance companies are increasingly investing in technologies that enable them to harness the power of their data. The ability to classify and predict client behavior not only enhances operational efficiency but also supports the development of innovative products and services tailored to changing consumer needs[4].

This research contributes to the broader body of knowledge on data mining applications in the insurance sector by demonstrating the practical implementation of the C4.5 Decision Tree algorithm for insurance product classification[5]. The study provides a comprehensive approach for transforming large datasets into actionable insights, ultimately supporting more accurate and efficient decision-making processes. By adopting such data mining techniques, insurance companies can improve their competitive positioning in the market, optimize their business strategies, and provide a higher level of service to their clients. In conclusion, the application of the C4.5 Decision Tree algorithm in insurance product classification represents a significant advancement in the use of data mining techniques for business strategy optimization. This study illustrates the potential for such methodologies to revolutionize the insurance sector, enabling companies to make data-driven decisions that align closely with client needs and preferences.

## 2. RESEARCH METHODOLOGY

Knowledge Discovery in Databases (KDD) is a comprehensive process aimed at discovering and identifying patterns within data, where these patterns must be valid, novel, useful, and understandable[6],[7]. KDD encompasses the integration of scientific techniques for the discovery, interpretation, and visualization of patterns in large datasets. It also involves the application of methods to integrate various scientific disciplines to extract valuable insights from the available data[8]. Data mining is a fundamental aspect of KDD that focuses on extracting hidden information from large volumes of data. It is not an entirely new field; rather, it combines and inherits techniques from established disciplines such as statistics, artificial intelligence, and database systems. Data mining can be applied across various domains, including marketing and financial management, telecommunications, astronomy, and other fields, to address complex challenges. The significance of data mining lies in its ability to manage vast amounts of data, facilitate transaction recording, and support data warehousing processes, ultimately providing users with accurate and meaningful information. The primary reason data mining has gained significant attention in the information industry in recent years is the availability of large-scale data and the increasing need to transform this data into meaningful information and knowledge. The core focus of data mining is extracting or mining knowledge from extensive datasets; the information derived from these processes is crucial for development and strategic decision-making

### 2.1 Data Mining

Data mining is a series of processes designed to extract additional value from a dataset by uncovering knowledge that was previously undetectable through manual methods[9], [10] The information generated is obtained by extracting and identifying significant or interesting patterns within the data stored in databases. The emergence of data mining is driven by the data explosion problem encountered in recent years, as many organizations have been collecting vast amounts of data over time (such as purchase data, sales data, customer data, transaction data, etc). Most of this data is entered using computer applications designed to manage daily transactions, which are predominantly OLTP (Online Transaction Processing) systems. The need for data mining has arisen due to the exponential growth in data volume, which has made it increasingly challenging for organizations to manually analyze and extract meaningful insights.

By leveraging sophisticated algorithms and analytical techniques, data mining identifies patterns, trends, and correlations that may not be immediately apparent through traditional methods. This approach enables organizations to transform raw data into actionable insights, supporting decision-making processes and strategic planning. As a result, data mining has become a vital tool for businesses seeking to gain a competitive advantage and optimize their operations by making informed, data-driven decisions. The primary reason for utilizing data mining is to assist in the analysis of behavioral observation collections. Such data are often susceptible to collinearity due to known interrelationships. A critical and unavoidable fact about data mining is that the subset or dataset being analyzed may not fully represent the entire domain, and as a result, it might not contain examples of certain critical relationships and behaviors present in other parts of the domain. To address this issue, the analysis can be enhanced through experimental-based approaches and other methodologies, such as Choice Modeling, which is particularly effective for human-generated data. In these situations, inherent correlations can be either controlled or eliminated during the construction of the experimental design. This ensures that the results obtained are robust and unbiased, providing a more accurate representation of the underlying data patterns and relationships[11]. By integrating these advanced techniques, data mining becomes a more reliable tool for

**Journal of Information System Research (JOSH)**
Volume 6, No. 4, Juli 2025, pp 1938−1947
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7216

extracting meaningful insights, even when faced with challenges related to dataset representativeness and data interdependencies[12],[13].

### 2.2 Clasification Decision Three

A Decision Tree is a structure used to divide a large dataset into smaller subsets of records by applying a series of decision rules[14], [15]. It is a powerful method for classification and prediction that simplifies complex decision-making processes by breaking them down into a hierarchy of simpler decisions[16],[17]. Decision Tree follows a top-down search strategy to find solutions. In the process of classifying unknown data, the value of an attribute is tested by tracing the path from the root node to a leaf node[18][19]. The leaf node then provides a prediction for the class that the new data instance belongs to. This step-by-step traversal ensures that each instance is evaluated based on its attribute values until it reaches the appropriate class prediction. Decision Trees are advantageous because they are intuitive, easy to interpret, and can handle both categorical and numerical data[20][21][22]. Additionally, they provide a visual representation that facilitates understanding of the decision-making process, making them a popular tool in data mining and machine learning applications. By utilizing a systematic approach to partition data based on attribute values, decision trees enable organizations to make data-driven decisions, predict outcomes, and classify data efficiently, even with large and complex datasets.

#### 2.2.1 Structure and Working of a Decision Tree

The classification process begins with the root node, which contains the entire dataset. The algorithm evaluates each attribute to determine the best attribute to split the data, based on a criterion that maximizes the separation between the different classes[23]. This process continues recursively, forming branches and nodes until the tree reaches the leaf nodes, which indicate the predicted class labels. The process can be visualized as follows:

a. Root Node: The starting point containing the complete dataset.
b. Internal Nodes: Nodes where an attribute is tested to split the data into subsets.
c. Branches: Outcomes of the test applied at the internal nodes.
d. Leaf Nodes: Terminal points that provide the final class label for instances.

#### 2.2.2 Splitting Criteria

A critical aspect of building a Decision Tree is selecting the attribute that best separates the data. Common criteria for splitting include Information Gain, Gini Index, and Gain Ratio.

a. Gini Impurity (CART Algorithm)
   The Gini impurity measures the likelihood of an incorrect classification of a randomly chosen element if it were labeled according to the class distribution of a node[20],[24]. It is calculated as follows:

   $$\mathrm{Gini(D)} = 1 - \sum_{i=1}^{C} p_i^2 \tag{1}$$

   Gini(D) represents the impurity of the dataset D, where C is the number of classes, and $p_i$ is the proportion of instances belonging to class i in the dataset D.
b. Entropy and Information Gain (ID3 Algorithm)
   Entropy is another measure used to determine the impurity or disorder of a dataset. It quantifies the uncertainty in the dataset by calculating the average amount of information needed to classify an instance correctly. The entropy of a dataset DDD is given by:

   $$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2(x_i) \tag{2}$$

   H(X) represents the entropy of the variable X, where $p(x_i)$ is the probability of occurrence of each category $x_i$ in X, and n is the number of distinct categories within the variable. Information Gain measures the reduction in entropy (uncertainty) after splitting the data on a particular attribute. The formula for Information Gain is:

   $$\mathrm{IG(T, A)} = \mathrm{Entropy(T)} - \sum_{v \in \mathrm{Values(A)}} \frac{|T_v|}{|T|} \times \mathrm{Entropy}(T_v) \tag{3}$$

   IG(T, A) represents the Information Gain of attribute A with respect to dataset T, indicating the reduction in entropy after splitting the dataset based on attribute A. Entropy(T) is the entropy of the entire dataset T before the split, measuring the level of disorder or impurity. Values(A) denotes the set of all possible values that attribute A can take, while $T_v$ is the subset of T where attribute A has the value v. $|T_v|$ is the number of instances in the subset $T_v$, |T| is the total number of instances in the dataset T, and Entropy($T_v$) is the entropy of the subset $T_v$.
c. Gain Ratio (C4.5 Algorithm)
   C4.5 Algorithm is an extension of the ID3 algorithm used for generating a decision tree[25][26]. Developed by Ross Quinlan, C4.5 addresses several limitations of ID3, making it one of the most popular algorithms for decision tree construction. The algorithm is primarily used for classification tasks and can handle both categorical and continuous data, which gives it a significant advantage in various applications. The C4.5 algorithm builds a decision tree by recursively partitioning the dataset based on the attribute that provides the

**Journal of Information System Research (JOSH)**
Volume 6, No. 4, Juli 2025, pp 1938−1947
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7216

most information gain relative to the target variable[27]. It extends the basic principles of the ID3 algorithm by introducing several improvements:

1. Handles both continuous and discrete attributes.
2. Deals with missing values by estimating them using the distribution of known values.
3. Applies pruning to reduce the tree size and avoid overfitting.
4. Computes a Gain Ratio instead of pure information gain to prevent the bias towards attributes with many distinct values.
   The Gain Ratio is an extension of Information Gain, addressing its bias toward attributes with many distinct values. It normalizes the information gain by considering the intrinsic information of the split:

$$\text{Gain Ratio}(T, A) = \frac{IG(T,A)}{\text{Split Information}(T,A)} \tag{4}$$

Gain Ratio(D, A) is a metric used to evaluate the quality of an attribute A for splitting the dataset D, balancing the Information Gain with the distribution of instances after the split. Information Gain(D, A) refers to the reduction in entropy achieved by splitting the dataset based on attribute A, measuring how effectively the attribute separates the data with respect to the target variable. Split Information(D, A), on the other hand, represents the amount of information produced by dividing the dataset according to attribute A, assessing how uniformly the instances are distributed among the resulting subsets.

$$\text{Split Information}(T, A) = \sum_{v \in \text{Values}} \frac{|T_v|}{T} \log_2\left(\frac{|T_v|}{T}\right) \tag{5}$$

Split Information(T, A) quantifies the information generated by splitting the dataset T using attribute A, helping to evaluate how uniformly the data is divided. Values(A) refers to the set of all possible values that attribute A can take, while $T_v$ represents the subset of dataset T where attribute A has the value v. $|T_v|$ is the number of instances in the subset $T_v$, and $|T|$ is the total number of instances in the original dataset T.

### 2.2.3 Type of Insurance Product

Insurance is a financial mechanism designed to mitigate risks and uncertainties faced by individuals or organizations. By paying premiums, policyholders receive assurance that financial losses resulting from specific events, such as accidents, natural disasters, or theft, will be compensated according to the terms of the policy. Insurance plays a critical role in maintaining economic stability by distributing risk and providing financial security. In the global economic context, insurance not only protects individuals and businesses but also supports infrastructure development, investment, and innovation. The insurance industry is rapidly adapting to technological changes, such as the use of big data and artificial intelligence, to streamline claims processes, assess risks more accurately, and develop more personalized products. These advancements enable insurers to respond more effectively to customer needs and improve operational efficiency. The study focuses on two specific types of insurance products, each designed to cater to different needs and stages of life:

1. PRUCritical Benefit 88 (PCB)
   Life insurance that provides comprehensive protection against the risk of critical illness up to the age of 88. This product provides benefits in the form of payment claims if the insured is diagnosed with one of the 60 critical illnesses covered, as well as protection against the risk of death worldwide.
2. PRUPrime Healthcare (PHC)
   A health insurance product that provides health protection with unlimited annual hospitalization benefits and begins by selecting the coverage area (Indonesia, Asia, or global). This insurance covers medical care costs, including surgery and intensive care, and provides protection up to the age of 99.
3. PRUMed Cover (PMC)
   Health insurance that offers protection against the costs of hospitalization, surgery, and intensive care. This product is designed to meet health protection needs and includes annual benefits that can be renewed in accordance with the policy terms.

## 3. RESULT AND DISCUSSION

### 3.1. Data Selection

Data analysis involves breaking down complete data into a description with the aim of analyzing insurance data. To analyze the data, several methods or stages can be applied to obtain the desired results. The analysis focuses on insurance financing transaction data using Data Mining techniques. To minimize errors, the data selected is limited to identifying the appropriate type of insurance product based on the customer's monthly premium, starting with premium amounts of IDR 200,000, IDR 350,000, and IDR 500,000. Additionally, it considers customers' income levels, distinguishing between those in the middle-to-upper category (earning more than IDR 3,000,000 per month) and those in the low-income category (earning less than IDR 3,000,000 per month), as well as the age category of prospective customers. The personal data of potential customers is presented in Table 1. below. This

**Journal of Information System Research (JOSH)**
Volume 6, No. 4, Juli 2025, pp 1938−1947
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7216

table provides detailed information about the characteristics of each customer. This data provides information to understand other relevant details of the prospective customer.

**Table 1.** Description of Personal Data of Prospective Customers

| No | Name | Age | Policy Type | Claim History | Income (IDR) | Occupation | Riches |
|---|---|---|---|---|---|---|---|
| 1. | Budi Santoso | 25 | PCB | 1 | 5,000,000 | Engineer | Sedan |
| 2. | Rini Wulandari | 45 | PMC | 2 | 10,000,000 | CEO | Motorbike |
| 3. | Agus Pratama | 35 | PMC | 0 | 7,500,000 | Driver | SUV |
| 4. | Siti Nuraini | 50 | PCB | 3 | 15,000,000 | Business Owner | Apartment |
| 5. | Joni Lesmana | 30 | PHC | 1 | 3,000,000 | Nurse | SUV |
| 6. | Indah Sari | 55 | PCB | 3 | 12,000,000 | Doctor | Sedan |
| 7. | Ahmad Fauzan | 40 | PMC | 0 | 9,000,000 | Accountant | Motorbike |
| 8. | Dewi Hanum | 60 | PCB | 2 | 13,000,000 | Police Officer | House |
| 9. | Andi Saputra | 20 | PMC | 1 | 8,000,000 | Chef | Motorbike |
| 10. | Lina Marlina | 65 | PCB | 3 | 14,000,000 | Lawyer | Sedan |
| … | …….. | …….. | …….. | …….. | …….. | …….. | …….. |
| 30. | Yuliawati | 26 | PMC | 3 | 13,000,000 | Lawyer | Sedan |

The prospective customer data will later become a case in the operational data mining process. From the existing data, the column taken as the attribute/decision variable is the Type of Insurance Product. While the column taken as the determining variable in the formation of the decision tree is:
1. Age
2. Policy Type
3. Claim History
4. Premium Amount
5. Riches

### 3.2 Data Processing

The classification problem ends with the production of knowledge presented in the form of a diagram commonly called a decision tree. To determine the type of insurance product that is right for a prospective customer, the criteria considered are income and age. One of the attributes is an attribute that states the solution data per data item called the target attribute. The following is customer data from individuals who already hold policies at PT. Prudential Life Assurance. The following is data that can be used for training data. Complete data can be seen in Table 2 below:

**Table 2.** List of Customer

| No | Name | Age | Claim History | Premium Amount (IDR) | Riches | Income | Policy Type |
|---|---|---|---|---|---|---|---|
| 1. | Budi Santoso | Young | 1 | 500,000 | Not Rich | Low | PCB |
| 2. | Rini Wulandari | Old | 2 | 750,000 | Not Rich | Low | PMC |
| 3. | Agus Pratama | Young | 0 | 750,000 | Not Rich | Low | PMC |
| 4. | Siti Nuraini | Old | 3 | 1,000,000 | Rich | High | PCB |
| 5. | Joni Lesmana | Young | 1 | 500,000 | Not Rich | Low | PHC |
| 6. | Indah Sari | Old | 3 | 750,000 | Not Rich | Low | PCB |
| 7. | Ahmad Fauzan | Young | 0 | 1,000,000 | Not Rich | Low | PMC |
| 8. | Dewi Sartika | Old | 2 | 1,000,000 | Rich | High | PCB |
| 9. | Andi Saputra | Young | 1 | 750,000 | Not Rich | Low | PMC |
| 10. | Lina Marlina | Old | 3 | 1,000,000 | Not Rich | High | PCB |
| …. | …… | …… | …… | …… | …… | …… | …….. |
| 30. | Yuliawati | Young | 3 | 1,000,000 | Not Rich | High | PMC |

Description:
Age: 20 Years to 40 Years is Young, 41 Years to 65 is Old
Claim History: 0 to 1 is Good, 2 to 3 is Not Good
Income: 5,000,000 to 9,000,000 is Low, above 10,000,000 is High
Property Type: Motorbike, Sedan, SUV is Not Rich, House, Apartement is Rich

### 3.3 Determine attribute values

From the sample data, first determine the selected node, namely by calculating the information gain value of each attribute. Total Insurance Premium, Age, Income, Type of Insurance Product. To determine the selected node, use

**Journal of Information System Research (JOSH)**
Volume 6, No. 4, Juli 2025, pp 1938−1947
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7216

the information gain value of each criterion with the specified sample data. The selected node is the criterion with the largest information gain. By using the equation no 2:

p(PRUMed Cover)=$\frac{16}{30}$ = 0.533 = − 0.533·log2 (0.533) = 0.487

p(PRUCritical Benefit 88)= $\frac{9}{30}$ = 0.3 = − 0.3·log2 (0.3) = 0.521

p(PRUPrime Healthcare)= $\frac{5}{30}$ = 0.167 = − 0.167·log2 (0.167) = 0.427

Total Entropy=0.487+0.521+0.427=1.435

After obtaining the total Entropy information, it is continued by calculating the information value of each attribute. From the data presented, the information value of the Insurance Premium attribute is known as in table 3. as follows:

**Table 3.** Value of Premium Amount Attribute Information

| Insurance Premium (IDR) | Type of Insurance Product | Total |
|---|---|---|
| 500,000 | PRUPrime Healthcare | 2 |
| 500,000 | PRUMed Cover | 2 |
| 500,000 | PRUCritical Benefit 88 | 0 |
| 750,000 | PRUPrime Healthcare | 9 |
| 750,000 | PRUMed Cover | 3 |
| 750,000 | PRUCritical Benefit 88 | 2 |
| 1,000,000 | PRUPrime Healthcare | 7 |
| 1,000,000 | PRUMed Cover | 5 |
| 1,000,000 | PRUCritical Benefit 88 | 0 |

For Premium Amount 500,000:

p(PRUMed Cover)=$\frac{2}{4}$ = 0.5;

p(PRUCritical Benefit 88)= $\frac{2}{4}$ = 0.5;

p(PRUPrime Healthcare)= $\frac{0}{4}$ =0

Entropy= − [0.5·log2(0.5) + 0.5·log2 (0.5) + 0·log2(0)] = − [0.5·(−1)+0.5·(−1)] = 1

For Premium Amount 750,000:

p(PRUMed Cover)=$\frac{9}{14}$ = 0.64;

p(PRUCritical Benefit 88)= $\frac{3}{14}$ = 0.214;

p(PRUPrime Healthcare)= $\frac{2}{14}$ =0.143

Entropy = − [0.643·log2(0.643) + 0.214·log2(0.214) + 0.143·log2 (0.143)]
= − [0.643·(−0.641) + 0.214·(−2.222) + 0.143·(−2.807)] = 1.287

For Premium Amount 1,000,000:

p(PRUMed Cover)=$\frac{7}{12}$ = 0.583;

p(PRUCritical Benefit 88)= $\frac{5}{12}$ = 0.417;

p(PRUPrime Healthcare)= $\frac{0}{12}$ =0

Entropy = − [0.583·log2(0.583) + 0.417·log2(0.417) + 0·log2(0)]
= −[0.583·(−0.781)+0.417·(−1.266)]= 0.98

After getting the Insurance Premium attribute information, continue by calculating the value of the age attribute information. From the data presented, the value of the age attribute information is known as in table 4. as follows:

**Table 4.** Information Value of Age Attribute

| Age Category | Type of Insurance Product | Total |
|---|---|---|
| Young | PRUMed Cover | 9 |
| Young | PRUPrime Healthcare | 5 |
| Young | PRUCritical Benefit 88 | 2 |
| Old | PRUMed Cover | 7 |
| Old | PRUPrime Healthcare | 7 |
| Old | PRUCritical Benefit 88 | 0 |

For Young Age Category

p(PRUMed Cover)= $\frac{9}{16}$ =0.562;

p(PRUPrime Healthcare)= $\frac{5}{16}$ =0.312;

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 1938−1947
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7216

p(PRUCritical Benefit 88)= $\frac{2}{16}$ =0.125

Entropy = − [0.562·log2 (0.562)+0.3125·log2 (0.3125) + 0.125·log2 (0.125)]

= − [0.5625·(−0.837) + 0.3125· (−1.678) + 0.125·(−3)] = 1.366

For Old Age Category

p(PRUMed Cover)= $\frac{7}{14}$ =0.5;

p(PRUPrime Healthcare)= $\frac{7}{14}$ =0.5;

p(PRUCritical Benefit 88)= $\frac{0}{14}$ =0

H=−[0.5·log2(0.5)+0.5·log2(0.5)+0·log2(0)] = −[0.5·(−1)+0.5·(−1)] = 1

After getting the age attribute information, continue by calculating the value of the Claim History attribute information. From the data presented, the value of the Claim History attribute information is known as in table 5. as follows:

**Table 5.** Information Value of Claim History Attribute

| Claim History | Type of Insurance Product | Total |
|---|---|---|
| Good | PRUMed Cover | 5 |
| Good | PRUPrime Healthcare | 6 |
| Good | PRUCritical Benefit 88 | 3 |
| Not Good | PRUMed Cover | 1 |
| Not Good | PRUPrime Healthcare | 8 |
| Not Good | PRUCritical Benefit 88 | 7 |

For Claim History Category

p(PRUMed Cover)= $\frac{5}{14}$ =0.357;

p(PRUPrime Healthcare)= $\frac{6}{14}$ =0.429;

p(PRUCritical Benefit 88)= $\frac{3}{14}$ =0.214

H = −(0.357 log2 (0.357) + 0.429log2 (0.429) + 0.214 log2 (0.214)) = 1.531

For Low Income Category

p(PRUMed Cover)= $\frac{1}{16}$ =0.063;

p(PRUPrime Healthcare)= $\frac{8}{16}$ =0.5;

p(PRUCritical Benefit 88)= $\frac{7}{16}$ =0.437

H= − (0.063 log2 (0.063) + 0.500 log2 (0.500) + 0.437 log2 (0.437)) = 1.272

After getting the claim history attribute information, continue by calculating the value of the Riches attribute information. From the data presented, the value of the Riches attribute information is known as in table 6. as follows:

**Table 6.** Information Value of Riches Attribute

| Riches | Type of Insurance Product | Total |
|---|---|---|
| Rich | PRUMed Cover | 2 |
| Rich | PRUPrime Healthcare | 3 |
| Rich | PRUCritical Benefit 88 | 3 |
| Not Rich | PRUMed Cover | 8 |
| Not Rich | PRUPrime Healthcare | 10 |
| Not Rich | PRUCritical Benefit 88 | 4 |

For Riches (Rich) Category

p(PRUMed Cover)= $\frac{2}{8}$ =0.25;

p(PRUPrime Healthcare)= $\frac{3}{8}$ =0.375;

p(PRUCritical Benefit 88)= $\frac{3}{8}$ =0.375

Entropy = − (0.25 log2 (0.25) + 0.375 log2 (0.375) + 0.375 log2 (0.375)) = 1.561

For Low Property Category

p(PRUMed Cover)= $\frac{8}{22}$ =0.364;

p(PRUPrime Healthcare)= $\frac{10}{22}$ =0.455;

p(PRUCritical Benefit 88)= $\frac{4}{22}$ =0.182

Entropy = − (0.364 log2 (0.364) + 0.455 log2 (0.455) + 0.182 log2 (0.182)) = 1.495

**Journal of Information System Research (JOSH)**
Volume 6, No. 4, Juli 2025, pp 1938−1947
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7216

After getting the Riches attribute information, continue by calculating the value of the income attribute information. From the data presented, the value of the Income attribute information is known as in table 7. as follows:

**Table 7.** Information Value of Income Attribute

| Income | Type of Insurance Product | Total |
|---|---|---|
| High | PRUMed Cover | 6 |
| High | PRUPrime Healthcare | 5 |
| High | PRUCritical Benefit 88 | 0 |
| Low | PRUMed Cover | 11 |
| Low | PRUPrime Healthcare | 5 |
| Low | PRUCritical Benefit 88 | 3 |

For High Income Category
p(PRUMed Cover)= $\frac{6}{11}$ =0.545;
p(PRUPrime Healthcare)= $\frac{5}{11}$ =0.455;
p(PRUCritical Benefit 88)= $\frac{0}{11}$ =0
Entropy = − (0.545 log2 (0.545) + 0.455 log2 (0.455)+0) = 0.994
For Low Income Category
p(PRUMed Cover)= $\frac{11}{19}$ =0.579;
p(PRUPrime Healthcare)= $\frac{5}{19}$ =0.263;
p(PRUCritical Benefit 88)= $\frac{3}{19}$ =0.158
Entropy = − (0.579 log2 (0.579) + 0.263 log2 (0.263) + 0.158 log2 (0.158)) = 1.384

### 3.4 Calculating Entropy and Information Gain Values of Attributes

The process of calculating entropy and information gain is essential in decision tree algorithms, particularly in determining which attribute best splits the dataset to improve prediction accuracy. These calculations with formula no. 3 helps evaluate the uncertainty (entropy) within a dataset and the effectiveness of an attribute in reducing that uncertainty (information gain).

a. Calculating the Information Gain Value of Insurance Premium Attributes

$$\text{Gain(Total, Insurance Premium)} = 1.435 - ((\frac{4}{30} \times 1) + (\frac{14}{30} \times 1.287) + (\frac{12}{30} \times 0.98))$$
$$= 1.435 - (0.133) + (0.601) + (0.392) = 1.435 - 1.126 = 0.309$$

b. Calculating the Information Gain Value of Age Attribute.

$$\text{Gain(Total, Age)} = 1.435 - ((\frac{16}{30} \times 1.366) + (\frac{14}{30} \times 0.1))$$
$$= 1.435 - (0.729) + (0.467) = 0.239$$

c. Calculating the Information Gain Value of Claim History Attributes

$$\text{Gain(Total, Claim History)} = 1.435 - ((\frac{14}{30} \times 1.531) + (\frac{16}{30} \times 1.272)$$
$$= 1.435 - (0.715) + (0.679) = 1.435 - 1.394 = 0.041$$

d. Calculating the Information Gain Value of Riches Attribute.

$$\text{Gain(Total, Riches)} = 1.435 - ((\frac{8}{30} \times 1.561) + (\frac{22}{30} \times 1.495))$$
$$= 1.435 - (0.416) + (1.097) = -0.078$$

e. Calculating the Information Gain Value of Income Attribute

$$\text{Gain(Total, Income)} = 1.435 - ((\frac{11}{30} \times 0.76) + (\frac{19}{30} \times 0.99)) = 0.194$$

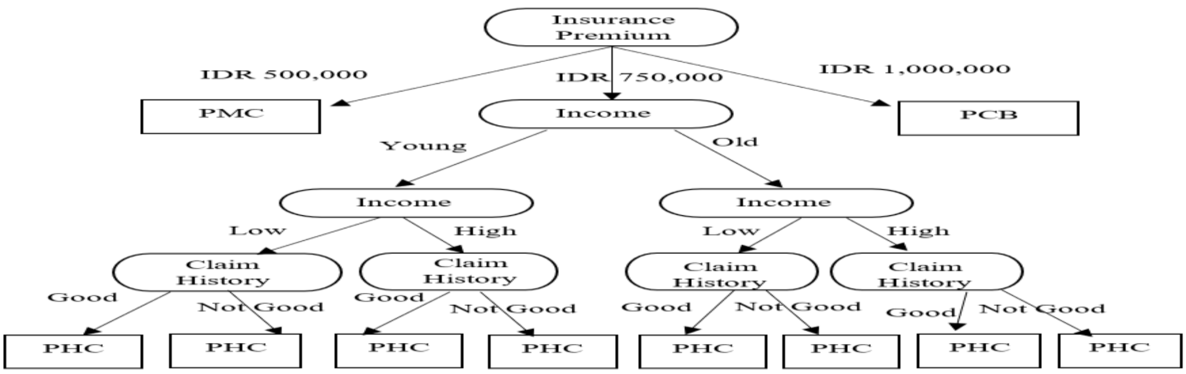


**Figure 1.** Decision Tree Results

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 1938−1947
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7216

Based on the results, we can observe the entropy values and information gain for each attribute. The attribute with the highest information gain is selected as the most informative attribute, which serves as the root node in the decision tree, as shown in Figure 1. Using this attribute, the data set is divided, and the process is repeated recursively to build a complete decision tree, which is shown below:

By observing the decision tree in Figure 1. it is known that the decision tree has been formed. And after the final tree is obtained, it is then changed into a rule. The following is the form of the tree that is changed into a rule:

R1: if Premium = IDR. 500,000 then Insurance Product Type = PRUMed Cover (PMC)
R2: if Premium = IDR. 750,000
If Age = Young:
    If Income = Low:
        If Claim History = Good: PRUPrime Healthcare (PHC)
        If Claim History = Not Good: PRUMed Cover (PMC)
    If Income = High:
        If Claim History = Good: PRUPrime Healthcare (PHC)
        If Claim History = Not Good: PRUMed Cover (PMC)
If Age = Old:
    If Income = Low:
        If Claim History = Good: PRUPrime Healthcare (PHC)
        If Claim History = Not Good: PRUMed Cover (PMC)
    If Income = High:
        If Claim History = Good: PRUPrime Healthcare (PHC)
        If Claim History = Not Good: PRUCritical Benefit 88 (PCB88)
R3: if Premium = IDR. 1,000,000 then Insurance Product Type = PRUCritical Benefit 88 (PCB)

Insurance Premium is the root decision, followed by Age, then Income, and finally Claim History to refine the decision. Depending on the combination of these attributes, the final insurance product decision is either PRUMed Cover (PMC), PRUPrime Healthcare (PHC), or PRUCritical Benefit 88 (PCB88).

## 4. CONCLUSION

The application of the C4.5 Decision Tree algorithm effectively classifies insurance clients based on premium, age, income, and claims history, generating decision rules for product recommendations. This method enhances customer segmentation, streamlines product selection, and improves marketing strategies by offering personalized solutions. The transparent decision tree structure also ensures regulatory compliance and interpretability. For instance, a client with a premium of IDR 500,000 is recommended PRUMed Cover (PMC), while a client with a premium of IDR 1,000,000 should consider PRUCritical Benefit 88 (PCB88). For premiums around IDR 750,000, age and income also play a role in determining the most appropriate recommendation. This approach streamlines the product selection process, ensures alignment with the client's financial status, and enhances marketing strategies by providing personalized solutions. Moreover, the clear and interpretable decision tree structure promotes transparency and ensures regulatory compliance. In practice, insurance companies can leverage this method to improve customer segmentation, refine product offerings, and increase overall satisfaction. Future work should explore alternative classification models, such as Random Forest and Support Vector Machines (SVM), to compare accuracy and robustness. Incorporating behavioral data, regional trends, and a larger dataset will enhance the model's scalability and generalizability, ensuring better adaptability to diverse market conditions. Expanding the study to hybrid models or ensemble learning techniques could further optimize predictive performance and decision-making in insurance risk classification.

## REFERENCES

[1] K. Sandner, S. Sieber, M. Tellermann, and F. Walthes, A Lean Six Sigma framework for the insurance industry: insights and lessons learned from a case study, vol. 90, no. 5–6. 2020.
[2] S. Gupta, P. Lehmann, S. Bonetti, A. Papritz, and D. Or, "Global Prediction of Soil Saturated Hydraulic Conductivity Using Random Forest in a Covariate-Based GeoTransfer Function (CoGTF) Framework," J. Adv. Model. Earth Syst., vol. 13, no. 4, pp. 1–7, Apr. 2021, doi: 10.1029/2020MS002242.
[3] Y. Han et al., "Leverage Classifier: Another Look at Support Vector Machine," Stat. Sin., vol. 23, no. 2, p. 242752, 2025, doi: 10.5705/ss.202023.0124.
[4] K. Dutta, S. Chandra, M. K. Gourisaria, and H. Gm, "A Data Mining based Target Regression-Oriented Approach to Modelling of Health Insurance Claims," Proc. - 5th Int. Conf. Comput. Methodol. Commun. ICCMC 2021, no. Iccmc, pp. 1168–1175, 2021, doi: 10.1109/ICCMC51019.2021.9418038.
[5] S. Chanmee and K. Kesorn, "Semantic decision Trees: A new learning system for the ID3-Based algorithm using a knowledge base," Adv. Eng. Informatics, vol. 58, no. 2, p. 102156, Oct. 2023, doi: 10.1016/j.aei.2023.102156.
[6] A. H. Nasyuha, Z. Zulham, and I. Rusydi, "Implementation of K-means algorithm in data analysis," TELKOMNIKA (Telecommunication Comput. Electron. Control., vol. 20, no. 2, p. 307, Apr. 2022, doi: 10.12928/telkomnika.v20i2.21986.

Journal of Information System Research (JOSH)
Volume 6, No. 4, Juli 2025, pp 1938−1947
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i4.7216

[7] M. Rambabu, S. Gupta, and R. S. Singh, “Data Mining in Cloud Computing: Survey,” 2021, pp. 48–56.
[8] A. M. Sarhan, “Data Mining in Internet of Things Systems: A Literature Review,” J. Eng. Res., vol. 6, no. 5, pp. 252–263, 2023.
[9] A. H. Nasyuha et al., “Frequent pattern growth algorithm for maximizing display items,” Telkomnika (Telecommunication Comput. Electron. Control., vol. 19, no. 2, pp. 390–396, 2021, doi: 10.12928/TELKOMNIKA.v19i2.16192.
[10] S. Wang, J. Cao, and P. S. Yu, “Deep Learning for Spatio-Temporal Data Mining: A Survey,” IEEE Trans. Knowl. Data Eng., vol. 34, no. 8, pp. 3681–3700, Aug. 2022, doi: 10.1109/TKDE.2020.3025580.
[11] S. M. Dol and P. M. Jawandhiya, “Classification Technique and its Combination with Clustering and Association Rule Mining in Educational Data Mining—A survey,” Eng. Appl. Artif. Intell., vol. 122, p. 106071, 2023.
[12] M. A. Baherifard, R. Kazemzadeh, A. S. Yazdankhah, and M. Marzband, “Improving the Effect of Electric Vehicle Charging on Imbalance Index in the Unbalanced Distribution Network Using Demand Response Considering Data Mining Techniques,” J. Oper. Autom. Power Eng., vol. 11, no. 3, pp. 182–192, 2023.
[13] Dr. T. Senthil Kumar, “Data Mining Based Marketing Decision Support System Using Hybrid Machine Learning Algorithm,” J. Artif. Intell. Capsul. Networks, vol. 2, no. 3, pp. 185–193, Aug. 2020, doi: 10.36548//jaicn.2020.3.006.
[14] M. M. Ghiasi and S. Zendehboudi, “Application of decision tree-based ensemble learning in the classification of breast cancer,” Comput. Biol. Med., vol. 128, p. 104089, Jan. 2021, doi: 10.1016/j.compbiomed.2020.104089.
[15] E. Alyahyan and D. Dusteaor, “Decision Trees for Very Early Prediction of Student’s Achievement,” in 2020 2nd International Conference on Computer and Information Sciences (ICCIS), 2020, vol. 12, no. 2, pp. 1–7, doi: 10.1109/ICCIS49240.2020.9257646.
[16] F. Nie, W. Zhu, and X. Li, “Decision Tree SVM: An extension of linear SVM for non-linear classification,” Neurocomputing, vol. 401, pp. 153–159, 2020, doi: 10.1016/j.neucom.2019.10.051.
[17] L. Vanfretti and V. S. N. Arava, “Decision tree-based classification of multiple operating conditions for power system voltage stability assessment,” Int. J. Electr. Power Energy Syst., vol. 123, no. October 2019, 2020, doi: 10.1016/j.ijepes.2020.106251.
[18] S. R. Jiao, J. Song, and B. Liu, “A Review of Decision Tree Classification Algorithms for Continuous Variables,” J. Phys. Conf. Ser., vol. 1651, no. 1, 2020, doi: 10.1088/1742-6596/1651/1/012083.
[19] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, “An improved random forest based on the classification accuracy and correlation measurement of decision trees,” Expert Syst. Appl., vol. 237, no. 1, p. 121549, Mar. 2024, doi: 10.1016/j.eswa.2023.121549.
[20] S. Tangirala, “Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm,” Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 2, pp. 612–619, 2020, doi: 10.14569/ijacsa.2020.0110277.
[21] N. Yuvaraj et al., “Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification,” Comput. Electr. Eng., vol. 92, no. April, 2021, doi: 10.1016/j.compeleceng.2021.107186.
[22] A. A. Dehghani, N. Movahedi, K. Ghorbani, and S. Eslamian, “Decision tree algorithms,” in Handbook of Hydroinformatics, vol. 5, no. 2, Elsevier, 2023, pp. 171–187.
[23] K. Maswadi, N. A. Ghani, S. Hamid, and M. B. Rasheed, “Human activity classification using Decision Tree and Naïve Bayes classifiers,” Multimed. Tools Appl., vol. 80, no. 14, pp. 21709–21726, 2021, doi: 10.1007/s11042-020-10447-x.
[24] S. Lee, C. Lee, K. G. Mun, and D. Kim, “Decision Tree Algorithm Considering Distances between Classes,” IEEE Access, vol. 10, no. June, pp. 69750–69756, 2022, doi: 10.1109/ACCESS.2022.3187172.
[25] A. Z. Abdullah, B. Winarno, and D. R. S. Saputro, “The decision tree classification with C4.5 and C5.0 algorithm based on R to detect case fatality rate of dengue hemorrhagic fever in Indonesia,” J. Phys. Conf. Ser., vol. 1776, no. 1, pp. 0–10, 2021, doi: 10.1088/1742-6596/1776/1/012040.
[26] H. Mardiansyah, M. Zarlis, and O. S. Sitompul, “Analysis of C4.5 Algorithm of Water Quality Dataset,” J. Phys. Conf. Ser., vol. 1898, no. 1, 2021, doi: 10.1088/1742-6596/1898/1/012002.
[27] M. Heydari, P. Pahlavani, and B. Bigdeli, “Modern Geomatics Technologies and Applications 1 Comparison of CART and C4 . 5 decision tree algorithms for classification of particulate matter pollution Modern Geomatics Technologies and Applications Comparison of CART and C4 . 5 decision tree algorithm,” no. April 2024, pp. 1–9, 2020.