Sinkron: Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13444
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Implementation of Classification Decision Tree and C4.5 Algorithm in selecting Insurance Products.

Sri Redjeki[1*], Ariesta Damayanti[2], Erna Hudianti[3], Asyahri Hadi Nasyuha[4]
[1,2,3)] Informatika, Fakultas Teknologi Informasi, Universitas Teknologi Digital Indonesia, Yogyakarta, Indonesia
[4)] Sistem Informasi, Fakultas Teknologi Informasi, Universitas Teknologi Digital Indonesia, Yogyakarta, Indonesia
[1)*]dzeky@utdi.ac.id, [2)]ariesta@utdi.ac.id, [3)]ernahudi@utdi.ac.id, [4)]asyahrihadi@gmail.com

Submitted :Jan 24, 2024 | **Accepted** : Jan 27, 2024 | **Published** : Jan 28, 2024

Abstract: Every insurance customer will receive a policy card, as a sign that the person is included in the insurance and is obliged to pay the insurance premium, the amount of which has been determined by the company in accordance with the agreement. Premium payments are Insurance's biggest source of income. Unfavorable economic conditions often cause customers not to pay their premiums by the specified time limit, resulting in a delay in completing the recording of premium income. This research aims to find out the right type of insurance product for prospective customers. The research method used is Classification Decision Tree. Classification Decision Tree is a research method used to examine existing facts systematically based on research objects, existing facts to be collected and processed into data, then explained based on theory so that in the end it produces a conclusion. This research is for selecting the right type of insurance product for prospective customers based on the age and income categories of prospective customers. Insurers must be more careful, especially in selecting prospective customers, and in determining the right type of insurance product for prospective customers so that the power in selecting the right type of insurance product for prospective customers is right at the intended target.

Keywords: *Data Mining, Classification Decision Tree, Algoritma C4.5*, Insurance.

INTRODUCTION

The competitive world of the insurance business means that players must always think about strategies that can ensure the continuity of the insurance company's business. The needs of the business world who want to obtain added value from the data they have collected have encouraged the application of data analysis techniques from various fields such as statistics, artificial intelligence, databases and so on to large-scale data, which ultimately gave rise to a new methodology called data mining. One of the main assets owned by insurance companies is an extraordinary amount of business data. This gives rise to the need for technology that can be used to generate new knowledge, which can help in setting insurance business strategies. Prediction of consumer interest is very important for an insurance company, whereby predicting consumer interest insurance companies can make decisions or strategies that are correct and appropriate for their consumers.

Insurance comes from the terms: Verzekering or assurantie (Dutch) Assurance or insurance (English) Insurance which means coverage or protection for an object from the threat of danger that causes loss. Insurance is a term used to refer to an action, system or business where financial protection for life, property, health and so on is compensated for unexpected events that can occur such as death, loss, damage, or illness, where involves regular premium payments over a specified period of time in exchange for a policy that guarantees such protection. The presence of data mining is motivated by the data problems experienced recently by many companies or banks. or the organization has been collecting data for several years (purchase data, sales data, customer data, transaction data, etc.). Premiums are one of the important elements in insurance because they are the main obligation that must be fulfilled by the insured to the insurer, because insurance can run or can be transferred from the insured to the insurer if the insured has paid the premium to the insurer or insurance company. Meanwhile, the policy is a written deed which is used as proof that an insurance agreement has been entered into. As is the case with companies operating in the insurance sector, the data stored so far is only for documentation and is only used for analysis needs. This data mining application utilizes input data to select the right type of insurance product for prospective customers.

*name of corresponding author

Sinkron: Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13444
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The concept of data mining is an effort to dig up information hidden in large amounts of data. Data mining is not a completely new field(Pranata & Utomo, 2020). One of the difficulties in defining data mining is the fact that data mining inherits many aspects and techniques from previously established scientific fields. Some of the solutions that can be solved by data mining are in the fields of markets and financial management, telecommunications, finance, astronomy, and other fields. Knowledge Discovery in Databases (KDD) is the entire process of searching for and identifying patterns in data, where the patterns found are valid, can be useful and can be understood(Nasyuha et al., 2022). KDD relates to integration techniques and scientific discovery, inteIDRretation, and visualization of patterns of data(Ikhwan, 2018). Data mining is a series of processes for extracting added value from a collection of data in the form of knowledge that has not been known manually. This study used two methods because different methods can help to validate the results. If both methods produce the same or similar results, it can increase confidence in those results. Conversely, if there are significant differences between two methods, this may be a sign to examine and re-evaluate the steps or assumptions used in each method. The resulting information is obtained by extracting and recognizing important or interesting patterns from the data contained in the database. Data mining is very necessary, especially in managing very large data to facilitate transaction recording activities and for data warehousing processes to provide accurate information for users. The main reason why data mining has attracted the attention of the information industry in recent years is because of the availability of large amounts of data and the increasing need to convert this data into useful information and knowledge because it fits the focus of this field of science, namely carrying out activities to extract knowledge from data. which are large or quantity, this information will be very useful for development.

LITERATURE REVIEW

Data Mining

Data Mining is not a completely new field. One of the difficulties in defining data mining is the fact that data mining inherits many aspects and techniques from previously established scientific fields (Zulham and Asyahri Hadi Nasyuha, 2018). An inevitable fact of data mining is that the subset or set of data analyzed may not be representative of the entire domain, and therefore may not contain examples of certain critical relationships and behavior that exist in other parts of the domain. To overcome this kind of problem, analysis can be augmented using experiment-based and other approaches, such as Choice Modeling for human-generated data. In this situation, the inherent correlations can be controlled for or removed altogether, during the construction of the experimental design(Hartama et al., 2019). Data mining has long roots in fields of science such as artificial intelligence, machine learning, statistics, databases, and information retrieval.

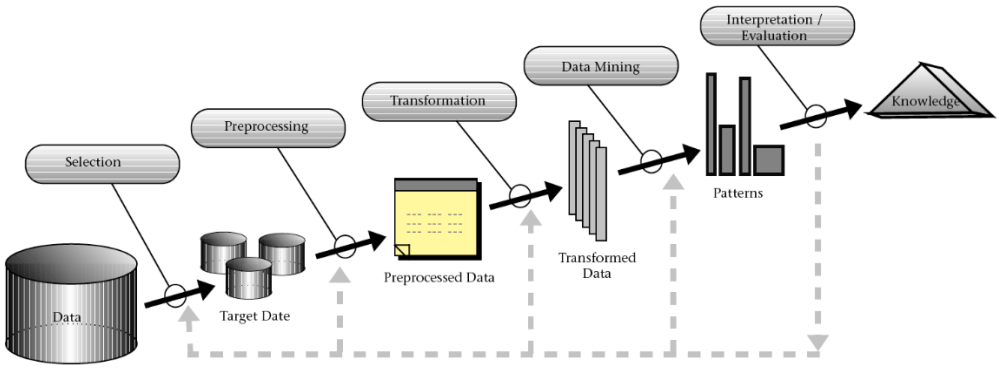

Figure 1. Stages in the Knowledge Discovery in Databases process

The following are the stages in the Knowledge Discovery in Databases (KDD) process, namely:

1. Data Cleaning (to remove inconsistent data noise)
2. Data Integration (where fragmented data sources can be combined)
3. Data Selection (where data relevant to the analysis task is returned to the database)
4. Data Transformation (where data is transformed or combined into the right form for mining with performance summaries or aggressive operations)
5. Data mining (the essential process in which intelligent methods are used to extract data patterns)
6. Pattern Evolution (to identify really interesting patterns that represent knowledge based on some interesting actions)

*name of corresponding author

Sinkron: Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13444
e-ISSN : 2541-2019
p-ISSN : 2541-044X

7. Knowledge Presentation (where an overview of visualization techniques and knowledge is used to provide the knowledge that has been mined to the user).
8. Pattern Evolution (to identify really interesting patterns that represent knowledge based on some interesting actions)
9. Knowledge Presentation (where an overview of visualization techniques and knowledge is used to provide the knowledge that has been mined to the user).

Association Rule

Association rules are a technique in data mining for determining the relationship between items in a predetermined dataset(Dol & Jawandhiya, 2023). This concept itself is derived from the terminology of market basket analysis, namely searching for relationships between several products in purchasing transactions. This technique looks for possible combinations that frequently appear in an itemset. There are several algorithms that have been developed regarding association rules, but there is one classic algorithm that is often used, namely the a priori algorithm. The basic idea of this algorithm is to develop frequent item sets(Hartama et al., 2019). By using one item and recursively developing frequent itemsets with two items, three items and so on until frequent item sets of all sizes. To develop frequent itemsets the reason is that if one item set does not exceed the minimum support, then any larger item set size will not exceed that minimum support. In general, developing sets with fc-items uses the frequent sets with k – 1 item developed in the previous step. Each step requires a single check of the entire contents of the database. Of the large number of rules that may be developed, it is necessary to have a fairly strong level of dependency between items in the antecedent and consequent. To measure the strength of this association rule, support and confidence measures are used. Support is the ratio between the number of transactions containing antecedents and consequents and the number of transactions. Confidence is the ratio between the number of transactions covering all items in the antecedent and consequent to the number of transactions covering all items in the antecedent.

$$Support = p(A \cap B) = \frac{\text{number of transactions containing items in } A \cap B}{\text{total number of transactions on D}}$$

$$Confidence = p(A/B) = \frac{\text{number of transactions containing items in } A \cap B}{\text{number of transactions containing the items in A}}$$

Clustering

Basically, clustering of data is a process for grouping a set of data without a previously defined class attribute, based on the conceptual principle of clustering, namely maximizing, and minimizing intra-class similarity. For example, a set of commodity objects can first be clustered into a set of classes and then into a set of rules that can be derived based on a certain classification(Syakur et al., 2018). The process of physically or abstractly grouping objects into classes or similar objects is called clustering or unsupervised classification(Kapil et al., 2016). Carrying out analysis using clustering, will be very helpful to form useful partitions for a large set of objects based on the "drive and conquer" principle which decomposes a large-scale system into smaller components, to simplify the design and implementation process. The main difference between Clustering Analysis and classification is that Clustering Analysis is used to predict classes whose real number format is in categorical or Boolean format.

Clasification

Classification is a process of finding a model that explains or differentiates concepts or data classes, with the aim of being able to estimate the class of an object whose class is unknown(Dol & Jawandhiya, 2023). In classification, a few records are given which is called a training set, which consists of several attributes, the attributes can be continuous or categorical, one of the attributes indicates the class of the record. Classification is the process of learning an objective function (target) f that maps each set of attributes x to one of the previously defined class labels y(Heaton, 2016). The target function is also called classification.

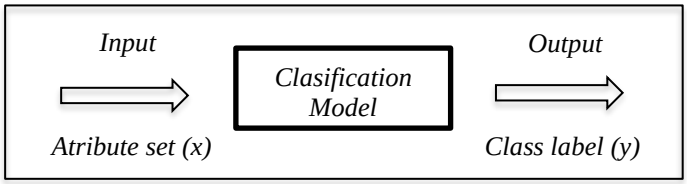

Figure 2. Classification as a task maps Attribute x into class label y

*name of corresponding author

Sinkron: Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13444
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Clasification *Decision Tree*

Decision Tree is a structure that can be used to divide large data sets into smaller sets of records by applying a series of decision rules. In general, Decision Trees carry out a top-down search strategy for solutions. In the process of classifying unknown data, attribute values will be tested by tracing the path from the root node to the final node (leaf) and then the class belonging to a particular new data will be predicted(Indini et al., 2022). Decision trees are one of the most popular classification methods because they are easy to inteIDRret by humans. A decision tree is a prediction model using a tree structure or hierarchical structure. The concept of a decision tree is to transform data into a decision tree and decision rules. The main benefit of using decision trees is their ability to break down complex decision-making processes into simpler ones so that decision-making will better inteIDRret solutions to problems. Decision trees are also useful for exploring data, finding hidden relationships between several candidate input variables and a target variable(Arnawisuda Ningsi, 2023). Decision trees combine data exploration and modeling, so they are great as a first step in the modeling process even when used as the final model for some other techniques. There is often a trade-off between model accuracy and model transparency. In some applications, the accuracy of a classification or prediction is the only thing that is highlighted or displayed, for example a Direct Mail Company creates an accurate model to predict which members are likely to respond to a request, without regard to how or why the model works.

A decision tree is a simple representation of a classification technique for a finite number of classes, where the root node is marked with a name attributes, the edges are labeled with possible attribute values and the leaf nodes are marked with different classes. An example of a decision tree can be seen in Figure 3 below:

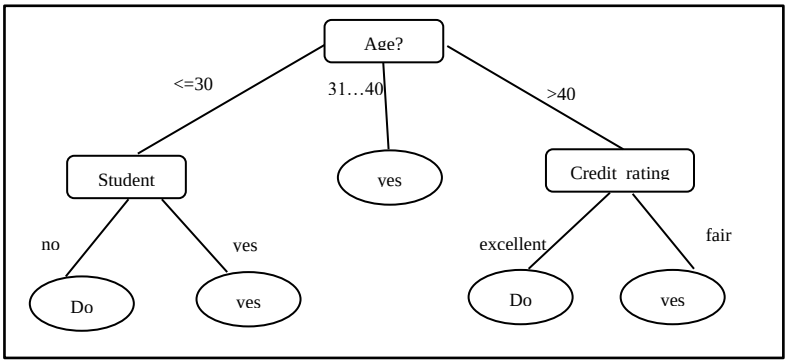

Figure 3. Decision Tree Model

Once a decision tree is built, it can be used to classify records that do not yet have a class. Starting from the root node, using tests on the attributes of records that do not yet have a class, then following the branch according to the results of the test, which will lead to the internal node (a node that has one incoming branch and two or more outgoing branches), with the method must do another test on the attributes or leaf nodes. Records whose class is unknown are then given a class that corresponds to the class at the leaf node. In a decision tree each leaf node marks a class label. The process in a decision tree is changing the form of data (table) into a tree model and then changing the tree model into rules.

Algorithm C4.5

The C4.5 algorithm is a group of decision tree algorithms. This algorithm has input in the form of training samples and samples(Andarista & Jananto, n.d.; Ubaedi & Djaksana, 2022). Training samples are sample data that will be used to build a tree that has been tested for correctness(Arnawisuda Ningsi, 2023; Girsang et al., 2022; Ucha Putri et al., 2021) Meanwhile, samples are data fields that we will later use as parameters in classifying data. The C4.5 algorithm is a development of the ID3 algorithm. The C4.5 and ID3 algorithms were created by a researcher in the field of artificial intelligence named J. Rose quinlan in the late 1970s. Algorithm C4.5 creates a decision tree from top to bottom, where the top attribute is the root, and the bottom attribute is called the leaf(Marlina & Bakri, 2021). In general, the C4.5 algorithm for building a decision tree is as follows:

1. Calculate the amount of data, the amount of data based on the result attribute members with certain conditions. For the first process, the conditions are still empty.
2. Select the attribute as Node.
3. Create a branch for each member of the Node.
4. Check whether the Entrophy value of any Node member is zero. If there are, determine which leaves are formed. If all the Entrophy values of Node members
5. is zero, then the process stops.

*name of corresponding author

Sinkron: Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13444
e-ISSN : 2541-2019
p-ISSN : 2541-044X

6. If any member of the Node has an Entropy value greater than zero, repeat the process from the beginning with the Node as a condition until all members of the Node have a value of zero.
7. Node is the attribute that has the highest gain value from the existing attributes.

METHOD

Data Mining Analysis

Data analysis can be defined as the decomposition of complete data into descriptions with the aim of analyzing the data. To analyze data, several methods or stages can be processed to obtain the desired results. The analysis taken as Data Mining is financing transaction data at an Assurance company. To avoid errors, limited data is selected for selecting the right type of insurance product according to the customer's premium each month starting with a premium amount of IDR. 200,000, IDR. 350,000, and IDR. 500,000,- and those with income in the middle to upper category. The middle-income category is those with income >=IDR. 3,000,000,- (three million and above) every month and those in the low category are those with income <IDR. 3,000,000,- (under three million) every month. And based on the age category of prospective customers. The following is the personal data of a prospective customer for an insurance company:

Tabel 1. List of Customer Transactions

No.	Name of Prospective Customer	Insurance Premium	Age	Income	Type of Insurance Product
1.	Ranton Sihotang	IDR. 200.000	Old	High	*PRULife Cover*
2.	Yuniarti Siregar	IDR. 200.000	Old	Low	*PRULife Cover*
3.	Frans Mayor Elowe	IDR. 200.000	Old	Low	*PRULife Cover*
4.	Sarjo Haryono	IDR. 200.000	Old	Low	*PRULife Cover*
5.	Obe Tridasuki	IDR. 200.000	Young	High	*PRULife Cover*
6.	Baikuni Wahyunita	IDR. 200.000	Young	Low	*PRULife Cover*
7.	Delyana	IDR. 350.000	Old	Low	*PRULife Cover*
8.	Rosanny, Br Saragih	IDR. 350.000	Old	Low	*PRULife Cover*
9.	Juliana	IDR. 350.000	Young	Low	PAA
10.	Ronesianty Sumbayak	IDR. 350.000	Young	Low	PAA
11.	Anisa Purnama	IDR. 350.000	Young	High	PAA
12.	St. Mangita Sihombing	IDR. 350.000	Old	High	PAA
13.	Hasian Buyung Silalahi	IDR. 350.000	Old	High	PAA
14.	Kartika fithri	IDR. 350.000	Young	Low	PAA
15.	Fery Fernando Saragih	IDR. 500.000	Young	High	PAA
16.	A Sian	IDR. 500.000	Young	Low	PAA
17	Hasnah Samad, Dra	IDR. 500.000	Old	High	PAA
18.	Fenti Isdayati	IDR. 500.000	Old	High	PAA
19.	H.S Saleh HSB, S. AG	IDR. 500.000	Old	High	PAA
20.	Khairiyah	IDR. 500.000	Young	Low	PAA

Information:
- Young Age = 21 Years to 40 Years
- Old Age = 41 Years to 69 Years
- High Income >IDR. 3,000,000
- Low Income <= IDR. 3,000,000

Decision tree models are commonly used in data mining to analyze data and induce trees and rules that will be used to make predictions. By following a decision tree, we can assign value to a case by deciding which branch to take, starting from the root node, and moving down to the leaves. By using this method, an insurance agent who

*name of corresponding author

Sinkron: Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13444
e-ISSN : 2541-2019
p-ISSN : 2541-044X

is responsible for deciding whether to grant an insurance policy to a customer can determine the right type of insurance product for the prospective customer.

Determining Selected Nodes or Determining Attribute Values

From the sample data, first determine the selected node, namely by calculating the information gain value for each attribute. Total Insurance Premium, Age, Income, Type of Insurance Product. To determine the selected node, use the information gain value from each criterion with the specified sample data. The selected node is the criterion with the greatest information gain.

By using the equation –p(+) log2 p(+) -p(-) log2 p(-), the information value (I) of all training data can be calculated:

Entropy (Total) = -(12/20) * LOG((12/20),2) - (8/20)*LOG((8/20),2) = 0.97

After obtaining the total Entropy information, proceed with calculating the information value of each attribute.

As for calculating the information value of the Insurance Premium attribute, it is as follows:

Table 2. Information Value of Insurance Premium attributes

Insurance Premium	Type of Insurance	Amount
200000	*PRULife Cover*	6
200000	PAA	0
350000	*PRULife Cover*	2
350000	PAA	6
500000	*PRULife Cover*	0
500000	PAA	6

Table 4. Table of Insurance Premium attribute parameters

Insurance Premium	Parameter
200000	Q1
350000	Q2
500000	Q3

Q1 = -(6/6)*log((6/6),2)-(0/6)*log((0/6),2) = 0.00

Q2 = -(2/8)*log((2/8),2)-(6/8)*log((6/8),2) = 0.81

Q3 = -(0/6)*log((0/6),2)-(6/6)*log((6/6),2) = 0.00

As for calculating the information value of the age attribute, it is as follows:

Table 5. Value Information attribute Age

Age	Type of Insurance	Amount
Young	*PRULife Cover*	2
Young	PAA	7
Old	*PRULife Cover*	6
Old	PAA	5

Table 6. Age attribute parameter table

Age	Parameter
Young	Q1
Old	Q2

Q1 = -(2/9)*log((2/9),2)-(7/9)*log((7/9),2) = 0.76

Q2 = -(6/11)*log((6/11),2)-(5/11)*log((5/11),2) = 0.99

As for calculating the information value of the Income attribute, it is as follows:

Table 7. Information Value of Income attributes

Income	Type of Insurance	Amount
High	*PRULife Cover*	2
High	PAA	7
Low	*PRULife Cover*	6
Low	PAA	5

Table 8. Income attribute parameter

Income	Parameter
High	Q1
Low	Q2

Q1 = -(2/9)*log((2/9),2)-(7/9)*log((7/9),2) = 0.76

*name of corresponding author

Sinkron: Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13444
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Q2 = -(6/11)*log((6/11),2)-(5/11)*log((5/11),2) = 0.99

Table 9. Calculation of Entropy and Gain Values

Node			Case (S)	PAA (S1)	PruLive Cover (S2)	Entropy	Gain
1	Total		20	8	12	0.97	
	Premi Insurance						0.65
		IDR 200,000	6	0	6	0.00	
		IDR 350,000	8	6	2	0.81	
		IDR 500,000	6	6	0	0.00	
	Age						0.08
		Young	9	7	2	0.76	
		Old	11	5	6	0.99	
	Income						0.08
		High	9	7	2	0.76	
		Low	11	5	6	0.99	

RESULT

From the results of the description above, it can be seen that the highest gain is the Premium attribute, which is 0.65, thus Premium can be the root node or initial node. From the results of the calculations above, a temporary decision tree can be drawn as shown in Figure 4. below:

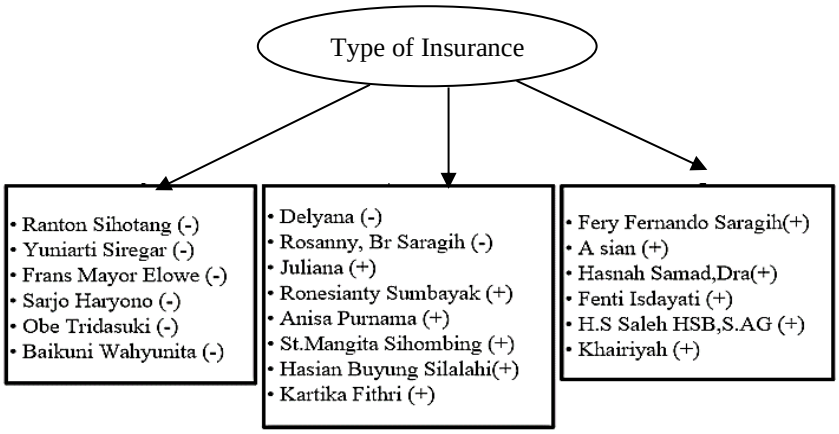

Figure 4. Root = Insurance Premium

Information :
• (-) PRULife Cover product type category
• (+) APA product type category

The following is a clearer tree shape based on Figure 4, namely:

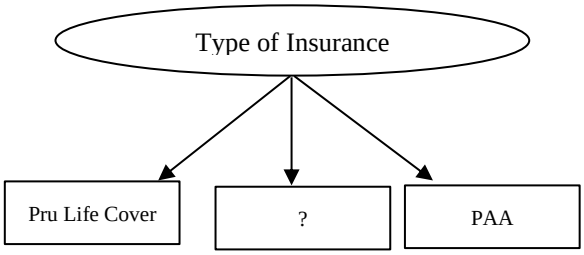

Figure 5. Root = temporary insurance premium

*name of corresponding author

Sinkron: Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 1, January 2024
DOI : https://doi.org/10.33395/sinkron.v9i1.13444
e-ISSN : 2541-2019
p-ISSN : 2541-044X

There are 3 attribute values for insurance premiums, namely IDR 200,000, IDR 350,000 and IDR 500,000. Of the two attribute values, the Insurance Premium attribute values of IDR. 200,000, and IDR. 500,000 have classified each case into 1, namely the PAA decision for a Premium of IDR. 500,000, and for a Premium of IDR. 200,000 to PRULive Cover, so there is no need to calculate further, but for the insurance premium attribute value of IDR. 350,000, further calculations still need to be done. The next node can be selected in the section that has + and - values, in the search results above only the Premium = IDR 350,000 attribute has + and - values, so everything must have an internal node.

DISCUSSIONS

Calculating the number of cases, the number of cases for PAA decisions, the number of cases for PRULife Cover decisions, and the entropy of all cases and cases divided based on the Age and Income attributes which can be the root node of the Insurance Premium attribute = IDR 350,000. After that, the gain calculation is carried out for each attribute. By looking at the decision tree in Figure 3.6, it is known that a decision tree has been formed. And after obtaining the final tree, it is converted into a rule. By carrying out the same process for the age and income categories, the following results are achieved:

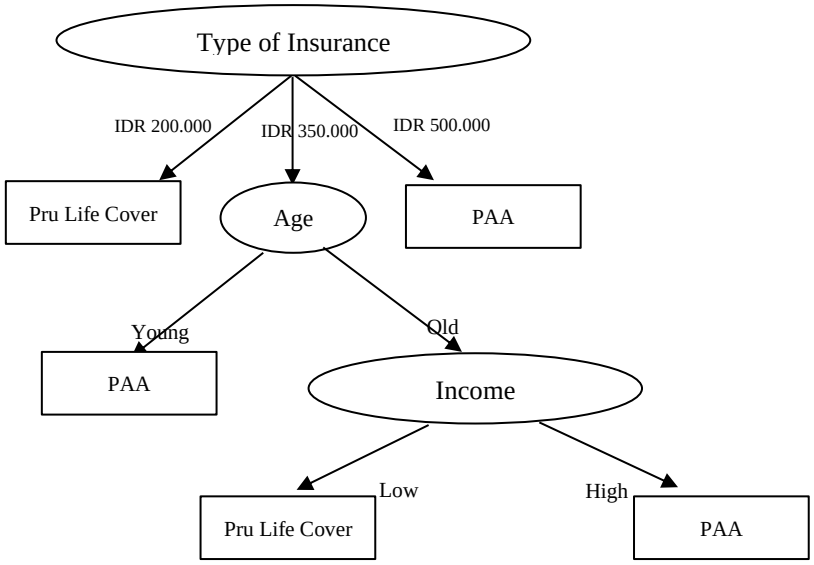

Figure 6. Decision tree final tree results

By looking at the decision tree in Figure 3.6, it is known that a decision tree has been formed. And after obtaining the final tree, it is converted into a rule. The following is a tree form that is converted into a rule:
R1: if Premium = IDR. 200,000 then Type of Insurance Product = PRULive Cover
R2: if Premium = IDR. 500,000 then Insurance Product Type = PAA
R3: if Premium = IDR. 350,000 and Age = Young and then Type of Insurance Product = PAA
R4: if Premium = IDR. 350,000 and Age = Old and income = Height and then Insurance Product Type = PAA
R5: if Premium = IDR. 350,000 and Age = Old and income = Low and then Type of Insurance Product = PRULive Cover

CONCLUSION

From the collection of customer transaction data, there is knowledge that is useful for insurance companies and Underwriting officers who provide policies to customers. From the results of mining transaction data for customers who are elderly and have low incomes for the premium category IDR. 350,000, which is more suitable for taking the type of insurance product in the form of PRULife Cover. And as a consideration for Underwriting analysis, it is necessary to know the income and assets owned by prospective customers. The results of the analysis obtained from data mining Classification Decision Tree rules can help companies, especially agents, in determining the right type of insurance product for prospective customers.

REFERENCES

Andarista, R. R., & Jananto, A. (n.d.). *Penerapan Data Mining Algoritma C4 . 5 Untuk Klasifikasi Hasil Pengujian Kendaraan Bermotor*. *16*(2), 29–43.

Arnawisuda Ningsi, B. (2023). Performance Comparison of Data Mining Classification Algorithms on Student Academic Achievement Prediction. *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, *6*(1), 29–39.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Dol, S. M., & Jawandhiya, P. M. (2023). Classification Technique and its Combination with Clustering and Association Rule Mining in Educational Data Mining—A survey. *Engineering Applications of Artificial Intelligence*, *122*, 106071.

Girsang, R., Ginting, E. F., & Hutasuhut, M. (2022). Penerapan Algoritma C4.5 Pada Penentuan Penerima Program Bantuan Pemerintah Daerah. *Jurnal Sistem Informasi Triguna Dharma (JURSI TGD)*, *1*(4), 449. https://doi.org/10.53513/jursi.v1i4.5727

Hartama, D., Perdana Windarto, A., & Wanto, A. (2019). The Application of Data Mining in Determining Patterns of Interest of High School Graduates. *Journal of Physics: Conference Series*, *1339*(1), 012042. https://doi.org/10.1088/1742-6596/1339/1/012042

Heaton, J. (2016). Comparing dataset characteristics that favor the Apriori, Eclat or FP-Growth frequent itemset mining algorithms. *Conference Proceedings - IEEE SOUTHEASTCON*, *2016-July*. https://doi.org/10.1109/SECON.2016.7506659

Ikhwan, A. (2018). A Novelty of Data Mining for FP-Growth Algorithm. *International Journal of Civil Engineering and Technology (IJCIET)*, *9*(7), 1660–1669.

Indini, D. P., Siburian, S. R., & Utomo, D. P. (2022). Implementasi Algoritma DBSCAN untuk Clustering Seleksi Penentuan Mahasiswa yang Berhak Menerima Beasiswa Yayasan. *Prosiding Seminar Nasional Sosial, Humaniora, Dan Teknologi*, 325–331.

Kapil, S., Chawla, M., & Ansari, M. D. (2016). On K-means data clustering algorithm with genetic algorithm. *2016 4th International Conference on Parallel, Distributed and Grid Computing, PDGC 2016*, 202–206. https://doi.org/10.1109/PDGC.2016.7913145

Marlina, D., & Bakri, M. (2021). Penerapan Data Mining Untuk Memprediksi Transaksi Nasabah Dengan Algoritma C4.5. *Jurnal Teknologi Dan Sistem Informasi (JTSI)*, *2*(1), 23–28.

Nasyuha, A. H., Zulham, Z., & Rusydi, I. (2022). Implementation of K-means algorithm in data analysis. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *20*(2), 307. https://doi.org/10.12928/telkomnika.v20i2.21986

Pranata, B. S., & Utomo, D. P. (2020). Penerapan Data Mining Algoritma FP-Growth Untuk Persediaan Sparepart Pada Bengkel Motor (Study Kasus Bengkel Sinar Service). *Bulletin of Information Technology (BIT)*, *1*(2), 83–91.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, *336*(1). https://doi.org/10.1088/1757-899X/336/1/012017

Ubaedi, I., & Djaksana, Y. M. (2022). Optimasi Algoritma C4.5 Menggunakan Metode Forward Selection Dan Stratified Sampling Untuk Prediksi Kelayakan Kredit. *JSiI (Jurnal Sistem Informasi)*, *9*(1), 17–26. https://doi.org/10.30656/jsii.v9i1.3505

Ucha Putri, S., Irawan, E., & Rizky, F. (2021). Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5. *Januari*, *2*(1), 39–46.

Zulham and Asyahri Hadi Nasyuha. (2018). *Penerapan Data Mining Untuk Pengelompokan Wahana*. *17*(1), 92–104.

*name of corresponding author