第 49 卷 第 6 期
2022 年 6 月

湖南大学学报（自然科学版）
Journal of Hunan University (Natural Sciences)

Vol. 49 No. 6
June 2022

Open Access Article https://doi.org/10.55463/issn.1674-2974.49.6.17

Data Set Analysis Using Rapid Miner to Predict Cost Insurance Forecast with Data Mining Methods

Johanes Fernandes Andry[1*], Henny Hartono[1], Honni[1], Aziza Chakir[2], Rafael[1]

[1] Information System Department, Universitas Bunda Mulia, Jakarta, Indonesia

[2] Hassan II University of Casablanca, Morocco University, Morocco

Abstract: The insurance protection program cannot be separated from everyday human life because there will always be risks in every human activity. Most people have entered into insurance agreements with state-owned and national private-owned insurance companies. The information system is one of the resources to increase competitive advantage. Information systems can be used to obtain, process, and disseminate information to support day-to-day operations and support strategic decision-making activities. The rapid growth of data accumulation has created data-rich but insufficient information conditions. Data mining is the mining or discovery of new information by looking for specific patterns or rules from large amounts of data expected to overcome these conditions. It is hoped that customer data can accurately produce information about insurance cost predictions. In this analysis, the authors use the RapidMiner Studio version 9.1 software. With the RapidMiner Studio app, authors can analyze the insurance data. A scientific novelty of this research is investigating data set cost insurance with data mining techniques consisting of classification, association, and clustering. Research goals for data mining techniques with classification, association, and clustering case studies implemented are to find all associative rules with high confidence, organize objects into groups whose members are similar, and collect objects between them. The following methods can be used: decision tree for data modeling, FP-Growth for determining which dataset occurs most frequently, and K-Means to classify the data attributes to facilitate the analysis.

Keywords: insurance, information system, data mining, RapidMiner.

使用 快速矿工進行數據集分析，通過數據挖掘方法預測成本保險預測

摘要：保险保障计划与日常生活密不可分，因为人类的每一项活动都存在风险。大多数人都与保险公司签订了保险协议，包括国有和国有私营保险公司。信息系统是可用于提高竞争优势的资源之一。信息系统可用于获取、处理和传播信息，以支持日常运营和战略决策活动。数据积累的快速增长为数据丰富但信息不足创造了条件。数据挖掘是通过从预期克服这些条件的大量数据中寻找特定模式或规则来挖掘或发现新信息。通过利用客户数据，希望它能准确地产生有关保险成本预测的信息。在此分析中，作者使用 快速矿工工作室 9.1 版软件。使用 快速矿工 工作室应用程序，作者可以分析保险数据。这项研究的科学创新是使用由分类、关联和聚类组成的数据挖掘技术来研究数据集成本保险。实施了分类、关联和聚类案例研究的数据挖掘技术的研究目标是找到所有具有高置信度的关联规则，将对象组织成成员相似的组以及它们之间的对象集合。可以使用决策树等方法对数据进行建模，FP-增长用于确定哪个数据集出现频率最高，K-均值将数据属性分组，因此更易于分析。

关键词：保险、信息系统、数据挖掘和 快速矿工.

Received: March 18, 2022 / Revised: April 22, 2022 / Accepted: May 13, 2022 / Published: June 30, 2022

About the authors: Johanes Fernandes Andry, Henny Hartono, Honni, Information System Department, Universitas Bunda Mulia, Jakarta, Indonesia; Aziza Chakir, Hassan II University of Casablanca, Morocco University, Morocco; Rafael, Information System Department, Universitas Bunda Mulia, Jakarta, Indonesia

Corresponding author Johanes Fernandes Andry, jandry@bundamulia.ac.id

1. Introduction

Insurance is a contract (policy) in which an individual or entity receives financial protection or reimbursement against losses from an insurance company. The company pools clients' risks to make payments more affordable for the insured [1]. The portfolio of products offered by insurance providers has diversified over the years. Accumulation of operational data inevitably follows from this growth in the industry. As a result, there is an increasing need to convert their data into a corporate asset to stay ahead and gain a competitive advantage [2].

Data mining is an interdisciplinary field of astronomy, business, computer science, economics, and others to discover new patterns from large datasets. Data mining technology can help insurance firms make crucial business decisions [3].

Insurance data analysis can be considered a way of reducing insurance companies' losses, and data mining may lead to valuable results. Data mining is unknown knowledge and laws discovery process and use of mass data and databases [4]. Data mining techniques have been mainly applied to database marketing by analyzing customer databases. Other applications include analysis and selection of stocks, fraud detection, spending patterns through the study of financial records, and detection of spatial patterns of bit failures in semiconductor memory fabrication [5].

Data mining techniques detect patterns in a large amount of data. As part of the larger concept of knowledge discovery, data mining involves statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and subsequent knowledge from large databases [6]. Data mining is a concrete innovation with great potential to help insurance companies concentrate on the most critical information about the actions of their clients and potential customers in the data they have collected [7].

In this paper, the authors determine the relationship between attributes, look for several sets of items, and classify various attributes based on their similarities using Rapid Miner Studio data mining software with several data mining techniques such as classification, association, and clustering to produce research.

Classification rules may be identified from a part of training data, and they may be tested for the rest of the data. The effectiveness of the classification approach may be evaluated in terms of the rule reliability with the test data set [8]. Association rule mining finds exciting relationships in data. The goal of associative rule data mining is to find all associative rules with high confidence (Strong Rules) in the data set [9]. Clustering is the process of organizing objects into groups whose members are similar in some way. Therefore, a cluster is a collection of objects that are similar to them and are dissimilar to the objects belonging to other clusters [10].

2. Literature Review

2.1. Classification

Classification is the process of finding a model or function that describes a concept or data class that aims to estimate the class of an object whose label is unknown. It can also be called a learning function that maps (classifies) a data element into one of several predetermined classes [11].

The Decision Tree method is a simple representation of the classification technique, which is the learning process of an objective function that maps each set of attributes to one of the previously defined classes. The decision tree can find hidden relationships between several candidate input variables and a target variable. Furthermore, a decision tree can combine data exploration and modeling, which is an excellent first step in the modeling process [12].

The Decision Tree Method has the following advantages [13]:

1) Decision-making areas that were previously complex and very global can be changed to be more straightforward and specific.

2) Eliminating unnecessary calculations occurs when the sample is tested only based on specific criteria or classes using the decision tree method.

3) Flexible in selecting facilities or features from different internal nodes, the selected feature will distinguish one criterion from another in the same node. Because this decision tree method is flexible, it will improve the quality of the resulting decisions compared to conventional one-stage calculation methods.

4) In multivariate analysis, with many criteria and classes, an examiner usually needs to estimate either the distribution of the height dimensions or the specific parameters of the class distribution. The Decision Tree method can avoid the emergence of this problem by using fewer criteria at each internal node without significantly reducing the quality of the resulting decisions, be it high dimensional distribution or specific parameters of the class distribution. The Decision Tree method can avoid this problem by using fewer criteria at each internal node without significantly reducing the quality of the resulting decisions.

2.2. Association

Data mining techniques that can be used and the two above are association data mining or what is commonly referred to as market basket analysis. Market basket analysis is a powerful tool for implementing a cross-selling strategy. This method begins by searching for several frequent item sets and continues with the formation of association rules. The Apriori algorithm and frequent pattern growth (FP-growth) are two prevalent algorithms for finding

frequent item sets from transaction data stored in the database. The task of association in data mining is to find attributes that appear at one time [14].

One of the methods used in this data set is FP-growth. FP-growth is an alternative algorithm that can determine the most frequent itemset in a data set. FP-growth uses a different approach from the paradigm used in the Apriori algorithm [15].

Frequent itemset excavation using the FP-Growth algorithm will generate a data tree structure called FP-Tree. The FP-Growth method can be divided into three main stages, namely [16]:

- The stage of generating the conditional pattern base,
- FP-Tree conditional generation stage, and
- The frequent itemset search stage.

2.3. Clustering

Clustering is a directionless data mining science. Clustering is the process of dividing data into classes based on the level of similarity. In clustering, data with similar characteristics will gather in the same group. Data that have different characteristics will gather in different groups [17].

One of the data mining techniques is the K-means Clustering Algorithm, which partitions data into one or more clusters (groups). Data with different characteristics are grouped into other groups [18].

This method partitions data into clusters/groups. Data with the same characteristics are grouped into the same cluster, and data with different characteristics are grouped into other groups. This clustering data aims to minimize the objective function set in the clustering process, which generally tries to minimize variations within a cluster and maximize variations between clusters [19].

3. Research Method

The CRISP-DM methodology is described as a hierarchical process model, consisting of sets of tasks described at four levels of abstraction (from general to specific): phase, generic task, specialized task, and process instance. CRISP-DM is divided into six phases to be carried out in a DM project, as shown in Figure 1.

Implementation details in each phase are:

1. *Business Understanding:* Understanding the project objectives and requirements from a business perspective and then converting this knowledge into a data mining problem definition. A preliminary plan is designed to achieve objectives [21].
2. *Data Understanding:* This is the second phase of the CRISP-DM process, which focuses on data collection, checking quality, and exploring data to get the insight of data to form hypotheses for confidential information [22].
3. *Data Preparation:* The third phase of the CRISP-DM process focuses on selecting and preparing the final data set. This phase may include many tasks records, tables, and attributes selection and cleaning and transformation of data.

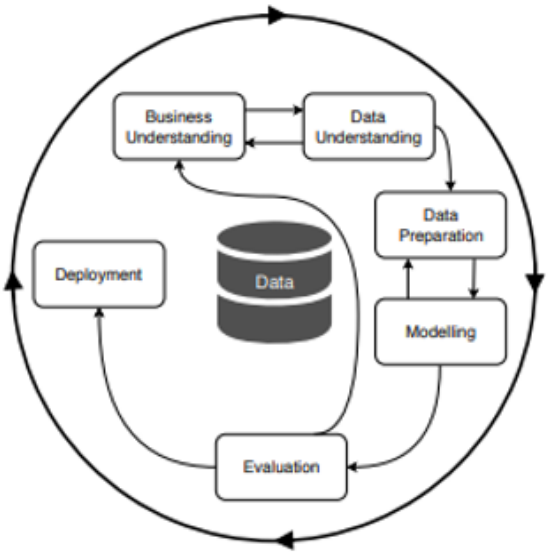

Fig. 1 Phases in CRISP-DM methodology [20]

Table 1 Dataset

Age	Sex	BMI	Smoker	Region	Charges
19	Female	27.9	Yes	Southwest	16.884
18	Male	33.77	No	Southeast	17.255
28	Male	33	No	Southeast	4.449
33	Male	22.71	No	Northwest	2.198.447
32	Male	28.88	No	Northwest	38.668
31	Female	25.74	No	Southeast	37.566
46	Female	33.44	No	Southeast	82.405
37	Female	27.74	No	Northwest	72.815
37	Male	29.83	No	Northeast	64.064
60	Female	25.84	No	Northwest	2.892.313
25	Male	26.22	No	Northeast	27.213
62	Female	26.29	Yes	Southeast	278.087
23	Male	34.4	No	Southwest	1.826
56	Female	39.82	No	Southeast	110.907
27	Male	42.13	Yes	Southeast	396.117

Table 1 gives a Personal Medical Cost data set that will be used for analysis. This data set has six columns:

1. *Age Column:* This column contains information on the age of the insurance customer.
2. *Sex Column:* This column contains information on the gender of the insurance customer.
3. *BMI Column:* This column contains information on body mass index.
4. *Smoker Column:* This column contains a description of whether the patient smokes or does not.
5. *Region Column:* This column contains a description of the area where the insurance customer lives.
6. *Charges Column:* This column contains information on the total cost of the insurance bill.
7. *Modeling:* The fourth phase of CRISP-DM process selection and application of various modeling techniques. Different parameters are set, and different models are built for the same data mining problem. Some techniques have specific requirements in the form of data.
8. *Evaluation:* In this phase, the results are evaluated towards the data mining goals, used models,

and cost-benefit relations. Here, checklists, internal benchmarking, and statistical analysis can help assess the results and create profitable improvement ideas [23].

9. *Deployment:* The final phase of the CRISP-DM process focuses on determining the use of obtaining knowledge and results. This phase also focuses on organizing, reporting, and presenting the gained knowledge when needed.

4. Results and Analysis

In this research, the authors use the process contained in the CRISP-DM method but do not carry out the final process, namely the implementation process.

4.1. Classification

4.1.1. Business Understanding

The business understanding stage focuses on understanding business needs. Then this understanding is transformed into early data mining, a plan designed to achieve a goal. The definition of business refers to the classification of customer data classifications. At this stage, an understanding of the background and objectives of the business process requires the classification of customer data:

1. Determine Business Objective. The business objective of conducting this research is to determine the classification of customers based on the age, BMI, and charges of insurance customers.
2. Determine Data Mining Goals. The purpose of data mining or the purpose of this research is to find knowledge about pattern classification related to the age, BMI, and charges of insurance customers.

4.1.2. Data Understanding

At the stage of understanding the data, the research starts with initialization and supports the data to determine the first insight into the data. Initial data collection was carried out by observation.

4.1.3. Data Preparation

Data preparation tasks will likely be performed multiple times and not in any prescribed order. Tasks include table, record, and attribute selection, data cleaning, new attributes construction, and data transformation for modeling tools.

Table 2 Data set for classification

Age	Sex	BMI	Smoker	Charges
19	Female	27.9	Yes	16.884
18	Male	33.77	No	17.255
28	Male	33	No	4.449
33	Male	22.71	No	2.198.447
32	Male	28.88	No	38.668
31	Female	25.74	No	37.566
46	Female	33.44	No	82.405
37	Female	27.74	No	72.815
37	Male	29.83	No	64.064
60	Female	25.84	No	2.892.313
25	Male	26.22	No	27.213
62	Female	26.29	Yes	278.087
23	Male	34.4	No	1.826
56	Female	39.82	No	110.907
27	Male	42.13	Yes	396.117

The data preparation stage includes all activities that build the final dataset (data included in modeling) from the initial raw data. Data preparation includes all activities to build a data set in the modeling tool from the initial raw data or create a new database for data mining settings. The database is independent or separate from the operational database. The main function is for modeling classification data.

The compilation of data includes all activities to build a data set that will be processed in the modeling process using the C4.5 algorithm to build a decision tree. The final data set can be seen in Table 1.

Table 2 presents a Personal Medical Cost data set that will be used for analysis. This data set has six columns:

1. *Age Column:* This column contains information on the age of the insurance customer.
2. *Sex Column:* This column contains information on the gender of the insurance customer.
3. *BMI Column:* This column contains information on body mass index.
4. *Smoker Column:* This column contains a description of whether the patient smokes or does not.
5. *Charges Column:* This column contains information on the total cost of the insurance bill.

4.1.4. Modeling

The choice of data mining techniques, algorithms, and determining parameters with optimal values. At the modeling stage, there are several things to do: selecting modeling techniques, building models, and assessing models:

1. *Selecting modeling technique:* The data mining technique chosen is the decision tree using the C4.5 algorithm. The decision tree and C4.5 algorithm are very appropriate to achieve the initial objectives of this study, which is to gain knowledge about the classification of insurance customers. Data mining modeling begins with making rules for the formation of decision trees.
2. *The Building Model phase* will be used as a benchmark in classifying current or non-current customers. Customer assessment criteria are benchmarks in classifying insurance customers, namely the age and BMI of each customer.
3. At the *Assessing model phase*, modeling is performed by forming a decision tree using the C.45 algorithm with predetermined rules of the three conditions, which will use 1338 random customer data.

4.1.5. Evaluation

The evaluation in this study is more focused on the model or pattern produced by the C4.5 algorithm. First, the resulting model is analyzed to determine whether the resulting pattern follows the standards of the insurance data set.

4.1.6. Deployment

The resulting model is analyzed to determine whether the resulting pattern follows the standards of the insurance data set. If the resulting pattern is not appropriate, then further analysis of the resulting pattern can result in recommendations for improvements expected in determining whether insurance customers are appropriate or not.

4.1.7. Research Result

A decision tree is formed after testing the C4.5 algorithm method on RapidMiner Studio, as shown in Figure 2.

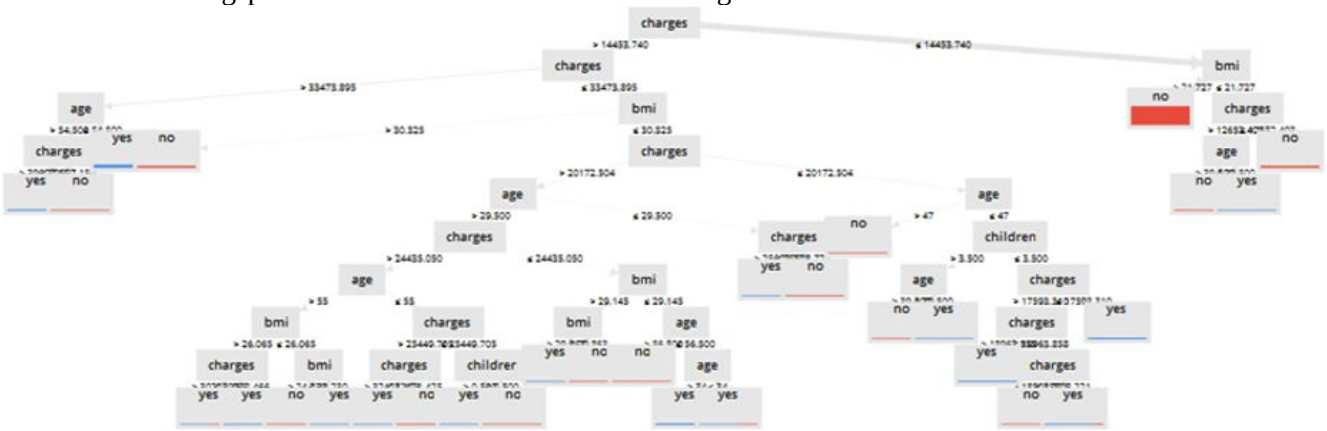

Figure 2. The root of the decision tree

Figure 2 that shows the root of the decision tree is Charges. The decision tree above produces rules that will be implemented in the program. The following is an explanation of the rules formed by modeling the C4.5 algorithm:

1) If charges ≤ 14453.740 then no (4 / 966)
2) If charges > 32328.502 then yes (150 / 4)
3) If bmi ≤ 29.225 and charges ≤ 21981.574 then yes (76 / 8)
4) If bmi > 30.057 then no (0 / 56)
5) If age > 38 and charges ≤ 24435.050 and charges > 22205.276 then yes (16 /0)
6) If age ≤ 55 and charges > 25449.705 then no (1 / 10)
7) If charges > 24516.678 and age > 46.500 then yes (23 / 3)
8) If bmi ≤ 29.670 then no (0 / 13)
9) If age ≤ 27 then yes (2 / 0)
10) If bmi ≤ 29.805 then no (0 / 2)
11) If age ≤ 48 then yes (2 / 0)
12) Else no (0 / 0)

4.2. Association

4.2.1. Business Understanding

The benefits of this transaction data can consider the sales strategy in the following month. In data mining, there is association analysis which helps analyze transaction data. Association means looking for a combination of goods in a transaction that has occurred.

4.2.2. Data Understanding

The stage of data understanding starts with initialization support to determine the first insight into the data. Initial data collection was carried out by observation.

4.2.3. Data Preparation

The data preparation stage includes all activities that build the final dataset (data included in modeling) from the initial raw data. The compilation of data includes all activities to build a data set that will be processed in the modeling process using the Association Rule to build an FP-Growth.

Table 3 Data set for association

Age	Sex	Children	Smoker
19	Female	0	Yes
18	Male	1	No
28	Male	3	No
33	Male	0	No
32	Male	0	No
31	Female	0	No
46	Female	1	No
37	Female	3	No
37	Male	2	No
60	Female	0	No
25	Male	0	No
62	Female	0	Yes
23	Male	0	No
56	Female	0	No
27	Male	0	Yes

Based on Table 2, the Personal Medical Cost data set will be used for analysis. This final data set has six columns:

1. *Age Column:* This column contains information on the age of the insurance customer.
2. *Sex Column:* This column contains information on the gender of the insurance customer.

3. *Children Column:* This column contains a description of the number of children covered.

4. *Smoker Column:* This column contains a description of whether the patient smokes or does not.

4.2.4. Modeling

The modeling stage combines selecting modeling techniques and assessing models:

1. *Select Modeling Techniques:* The data mining technique chosen is the FP-Growth using the Association Rule. Fp-growth can be used to achieve the initial objective of this study, namely, to find the relationship between insurance customer data based on age, children, sex, and smoker.

2. *Assess Models:* Modeling is done by forming a frequent item set. The following process is the generation of association rules when the Frequent Itemset has been obtained.

4.2.5. Evaluation

The evaluation is carried out in-depth with the aim that the results at the modeling stage follow the targets to be achieved in this part using the FP-growth association method.

4.2.6. Deployment

The final report regarding the knowledge obtained or relation recognition of data in the data mining process is presented in tables or descriptions that are easy to understand.

4.2.7. Research Analysis

FP-Growth is more beneficial because it only scans the database once or twice, while the Apriori algorithm needs to scan the database repeatedly, so it tends to take longer. In this analysis, the authors try to find Association Rules with two parameters, namely the Frequent Items Set and the preparation of the Rules. The function of FP-Growth, in this case, is to find out the Frequent Item Set, while the Association Rules serve to formulate Rules.

Therefore, it is necessary to determine the parameters of FP-Growth and Association Rules to produce valid data. The results from applying FP-Growth modeling in Rapid Miner are shown in Table 5.

Based on Table 5 can be seen the result of the Frequent Item Sets obtained from the modeling results above. There is a support value and items from each row. For example, support in the value above shows a support value of the several combinations of set items listed.

Table 4 shows the result of preparing the Rules obtained from the Creating Association Rules procedure.

Table 4 Association rules

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
5	smoker	children	0.452	0.569	0.809	-1.138	-0.002	0.996	0.994
6	smoker	age, children	0.452	0.569	0.809	-1.138	-0.002	0.996	0.994
7	age, smoker	children	0.452	0.569	0.809	-1.138	-0.002	0.996	0.994
8	smoker, sex	children	0.220	0.569	0.880	-0.553	-0.001	0.996	0.995
9	smoker, sex	age, children	0.220	0.569	0.880	-0.553	-0.001	0.996	0.995
10	age, smoker, sex	children	0.220	0.569	0.880	-0.553	-0.001	0.996	0.995

4.3. Clustering

4.3.1. Business Understanding

At this stage, an understanding of the substance of the data mining activities will be carried out, and the needs of the prospective business.

Table 5 Frequent item set

Size	Support	Item 1	Item 2	Item 3	Item 4
1	1,000	Age			
1	0,795	Smoker			
1	0,571	Children			
1	0,505	Sex			
2	0,795	Age	Smoker		
2	0,571	Age	Children		
2	0,505	Age	Sex		
2	0,452	Smoker	Children		
2	0,386	Smoker	Sex		
2	0,292	Children	Sex		
3	0,452	Age	Smoker	Children	
3	0,389	Age	Smoker	Sex	
3	0,292	Age	Children	Sex	
3	0,220	Smoker	Children	Sex	
4	0’220	Age	Smoker	Children	Sex

Knowing everything about the business object (research object) is imperative to understanding the data that is then analyzed.

1. *Determine Business Objectives:* The stage of determining business objectives and uncovering the critical factors involved in the planned research ensures that the research does not produce correct answers to wrong questions. The business objective based on the explanation in this analysis is to see the relationship between data on insurance customers.

2. *Determine Data Mining Goals:* The stage of transforming knowledge in the business domain into a data mining problem definition and determining data mining objectives (research). The purpose of data mining or this research is to explore knowledge (discovering knowledge) about patterns of relationships between customer data.

4.3.2. Data Understanding

The stage of data understanding starts with initialization and supports the data to determine the first insight into the data. Initial data collection was carried out by observation.

4.3.3. Data Preparation

Data preparation comprises all activities to build data sets that will be included in the modeling tool from the initial raw data or create a new database for data mining setups. The stage of designing raw data in the data set is used for data mining modeling. Data set

design must adjust to what has been formulated at the business understanding stage, especially in formulating data mining objectives, namely describing patterns (pattern recognition) of relationships between customer data. The final data set in this phase are the same as in Table 1.

4.3.4. Modeling

Selection of data mining techniques, algorithms, and determination of parameters with optimal values is a must for the modeling stage. Therefore, the modeling steps are as follows:

1. *Select Modelling Technique:* The data mining technique chosen is clustering using the K-means algorithm. Clustering and K-means algorithms are very appropriate to achieve the initial goal of this study, namely, to find knowledge about patterns in customer data.
2. *Build Model:* Data mining software is built using the RapidMiner Studio application program. Although many data mining tools provide K-means algorithm features, including Access, Weka, Matlab, and others, by building clustering modeling with the K-means algorithm, it is hoped that who can carry out exploration of the iteration process or the result of modeling.
3. *Assess Model:* Different initiation of cluster centers may result in different final cluster centers. Even though the initiation of the cluster center in the K-Means algorithm is determined randomly, determining the initiation is needed to obtain optimal final cluster center results. Modeling was done several times with different cluster center initiations to assess which modeling resulted in the most optimal final cluster center.
4. *Generate Test Design:* Testing test or learning stage. The clustering technique does not require a learning stage because clustering is unsupervised learning and grouping naturally based on the similarity of its attributes, which is different from other classification techniques. Testing was carried out with ten pairs of different cluster center initiation methods. The most optimal final cluster center will be selected from the different clusters and produce two cluster centers from the different clusters.

4.3.5. Evaluation

Evaluation is the interpretation phase of data mining results. The evaluation is carried out in-depth with the aim that the results at the modeling stage follow the targets to be achieved in the business understanding stage using the k-means clustering method.

4.3.6. Deployment

Deployment is the stage of making reports on the results of data mining activities.

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -85.938
Avg. within centroid distance_cluster_0: -84.913
Avg. within centroid distance_cluster_1: -86.965
Davies Bouldin: -0.697
```

Fig. 3 Performance vector result

The final report regarding the knowledge obtained or pattern recognition of data in the data mining process is presented in graphs or descriptions that are easy to understand.

4.3.7. Research and Analysis

Based on the K-Means results that have been searched, the lowest value of Davies Bouldin is obtained from the value of k = 2.

Based on Figure 3, the calculation results of Performance are as follows:

1. *Avg Value:* Within centroid distance of each cluster is not close to zero, which means that each cluster member is not nearby.
2. *The K-Means (Davies Bouldin) value* shows a value of 0.697; this number means that each object in the cluster is reasonably good because it is close to zero.

A plot view from K-means clustering is formed like Figure 4.

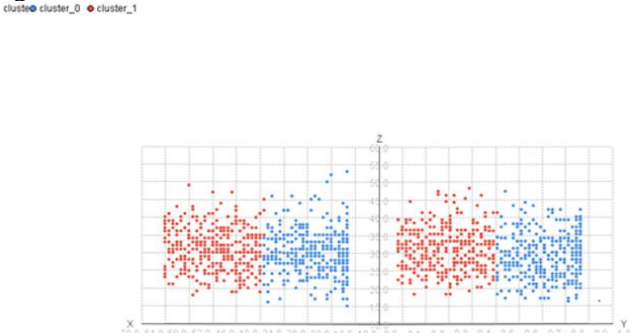

Fig. 4 Scatter plot

Figure 4 shows a Scatter Diagram of the Clustering results with the x-Axis attribute age, y-Axis attribute sex, and z-Axis attribute BMI. Again, two colors can distinguish, where the blue color is cluster 1 while the red color is Cluster 2.

5. Conclusion

The classification method, the C4.5 algorithm, is applied to the insurance data set by calculating the information resulting in a value of 95.81% +/- 1.30% and C4.5 decision tree modeling.

The Association method implements the insurance data set using the Fp-Growth algorithm and the Association Rule. Search Association rules are carried out using two parameters, Frequent Item Sets, and Rule compilation; hence, FP-Growth functions can view Frequent Item Sets and their parameters. The result of

the Frequent Item Sets is that there are support values and items from each existing line. The support value is shown as the supporting value of the listed set of items. Meanwhile, in the Association Rules, a confidence value is used to measure an association's validity. When the trust value is equal to or greater than the predetermined minimum value, the data can be said to be a good association rule.

In the clustering method, the insurance data set is implemented using the K-means algorithm, then the data set is modified because the clustering method can only accept integer data. From the results of the K-Means algorithm that has been searched, the lowest value of Davies Bouldin is the value of k = 2. Then, the calculation results of the Performance Vector get the value of Average within the centroid distance of each cluster does not say the number 0, and the K-Means value shows a value of 0.697.

6. Suggestion and Limitation

The limitation of the study is the use of the C4.5 algorithm to perform classification; it is necessary to select the right variables so that the results from the decision tree are more accurate or detailed. However, this research can be developed by combining or comparing other data mining methods for better prediction results.

References

[1] WISEMAN V, THABRANY H, ASANTE A. et al. An evaluation of health systems equity in Indonesia: study protocol. *International Journal for Equity in Health*, 2018, 17(1): 138. https://doi.org/10.1186/s12939-018-0822-0

[2] ELTAHIR O A B. The Effect of Information Technology on The Cooperative Insurance Industry Case Study: Shiekan Insurance and Reinsurance Company – Sudan (Empirical Study). *International Journal of Economics, Business and Accounting Research*, 2020, 4(1): 27–37, https://jurnal.stie-aas.ac.id/index.php/IJEBAR

[3] HIWASE V A, and AGRAWA A J. Review on Application of Data Mining in Life Insurance. *International Journal of Engineering & Technology*, 2018, 7(45):159-162, http://dx.doi.org/10.14419/ijet.v7i4.5.20035.

[4] KARAMIZADEH F, and ZOLFAGHARIFAR S A. Using the Clustering Algorithms and Rule-based of Data Mining to Identify Affecting Factors in the Profit and Loss of Third-Party Insurance, Insurance Company Auto. *Indian Journal of Science and Technology*. 2016, 9(7): 1-9. https://doi.org/10.17485/ijst/2016/v9i7/87846.

[5] TEMBHURNE D S, ADHIKARI J, and BABU R. Implementation of Data Mining Techniques in CRM of Pharmaceutical Industry. *Proceedings of the 2019 International Conference on Innovation & Research in Engineering, Science & Technology*, 2019: 7-12.

[6] SAADATDOOST R, SIM A T H, JAFARKARIMI H, and HEE J M. Knowledge Discovery for Large Databases in Education Institutes. In *Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications* 2018, Chapter 10: 158-245. https://doi.org/10.4018/978-1-5225-5191-1.Ch010

[7] SENOUSY Y, HANNA W K, SHEHAB A, RIAD A M, EL-BAKRY H M, and ELKHAMISY N. Egyptian social insurance big data mining using supervised learning algorithms. *Revue d'Intelligence Artificielle*, 2019, 33(5): 349–357. https://doi.org/10.18280/ria.330504.

[8] JIJO B T, and ABDULAZEEZ A M. Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, 2021, 2(1): 20-28.

[9] GUPTA N, NARAYAN R and CHAUDHARI A. Implementation of Meteorological Data Analysis Using Techniques for Weather Prediction. *International Journal of Engineering Applied Sciences and Technology*, 2018, 2(12): 28-31.

[10] BALA A. A Comparative Study of Various Clustering Algorithms in Data Mining. *International Journal of Advanced Research in Science and Engineering*, 2018, 7(7): 1182-1191.

[11] ANDRY J F, GUNADI J, REMBULAN G D, and TANNADY H. Big Data Implementation in Tesla Using Classification with Rapid Miner. *International Journal of Nonlinear Analysis and Applications*, 2021, 12(Special Issue): 2057-2066. http://dx.doi.org/10.22075/ijnaa.2021.6016

[12]. SAFRI Y F, ARIFUDIN R, and MUSLIM M A. K-nearest neighbor and naive Bayes classifier algorithm in determining the classification of healthy card Indonesia giving to the poor. *Scientific Journal of Inf*ormatics, 2018, 5(1): 9-18. https://doi.org/10.15294/sji.v5i1.12057

[13] MADYATMADJA E D, JORDAN S I, and ANDRY J. F. Big Data Analysis Using RapidMiner Studio to Predict Suicide Rate in Several Countries, *ICIC Express Letters Part B: Applications*, 2021, 12(8): 757-764, https://doi.org/10.24507/icicelb.12.08.757.

[14] MADYATMADJA E D, MARVEL, ANDRY J F, TANNADY H. and CHAKIR A. Implementation of Big Data in Hospital using Cluster Analytics. *Proceedings of the 2021 International Conference on Information Management and Technology (ICIMTech):* 496-500.

[15] ANDRY J F, REYNALDO S A, CHRISTIANTO K, et al. Algorithm of Trending Videos on YouTube Analysis using Classification, Association and Clustering. *Proceedings of the 2021 International Conference on Data and Software Engineering (ICoDSE), IEEE Catalog Number: CFP21AWL-USB.*

[16] MADYATMADJA E D, SEMBIRING D J. M, PERANGIN ANGIN S M, FERDY D, and ANDRY J. F. Big Data in Educational Institutions using RapidMiner to Predict Learning Effectiveness, *Journal of Computer Science*, 2021, 17(4): 403-413, https://doi.org/10.3844/jcssp.2021.403.413

[17] SILALAHI R M P, ANDRY J F, BERNANDA D Y, TANNADY H, and ENIRIANTI. Big Data Analytics in Library to Classification Book Publishers, *Journal of Positive School Psychology*, 2022, 6(2): 4303 – 4310,

[18] ANDRY J F, TANNADY H, LIMAWAL I I, REMBULAN G D, and MARTA R F. Big Data Analysis on YouTube with Tableau. *Journal of Theoretical and Applied Information Technology*, 2021, 99(22): 5460-5469.

[19] MADYATMADJA E D, RIANTO A, ANDRY J F, TANNADY H, and CHAKIR A. Analysis of Big Data in Healthcare Using Decision Tree Algorithm, *Proceedings of 2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)*, IEEE Part Number: CFP19H83-ART

[20] MARTÍNEZ-PLUMED F, OCHANDO L. C, FERRI C, et al. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 2021, 33(8): 3048-3061, https://doi.org/10.1109/tkde.2019.2962680.
[21] SCHRÖER C, KRUSE F, and GÓMEZ J M. A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, 2021, 181: 526-534.
[22] PLOTNIKOVA V, DUMAS M, and MILANI F. Adaptations of data mining methodologies: a systematic literature review, *PeerJ Computer Science*, 2020; 6: e267, https://doi.org/10.7717/peerj-cs.267.
[23] SCHÄFER F, ZEISELMAIR C, and BECKER J. Synthesizing CRISP-DM and Quality Management: A Data Mining Approach for Production Processes *2020 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*, 2018: 190-195, https://doi.org/10.1109/ITMC.2018.8691266.

参考文:

[1] WISEMAN V、THABRANY H、ASANTE A. 等。印度尼西亚卫生系统公平性评估：研究方案。国际健康公平杂志，2018，17(1)：138. https://doi.org/10.1186/s12939-018-0822-0
[2] ELTAHIR O A B. 信息技术对合作保险业的影响案例研究：舍感觉保险公司和再保险公司– 苏丹（实证研究）。国际经济、商业和会计研究杂志，2020，4(1)：27-37，https://jurnal.stie-aas.ac.id/index.php/IJEBAR
[3] HIWASE V A 和 AGRWA A J. 数据挖掘在人寿保险中的应用回顾。国际工程与技术杂志，2018，7(45):159-162，http://dx.doi.org/10.14419/ijet.v7i4.5.20035。
[4] KARAMIZADEH F 和 ZOLFAGHARIFAR S A. 使用聚类算法和基于规则的数据挖掘来识别第三方保险损益的影响因素，保险公司汽车。印度科学技术杂志。 2016，9（7）：1-9。https://doi.org/10.17485/ijst/2016/v9i7/87846。
[5] TEMBHURNE D S、ADHIKARI J 和 BABU R. 制药行业客户关系管理中数据挖掘技术的实施。2019 年工程、科学与技术创新与研究国际会议论文集，2019：7-12。
[6] SAADATDOOST R、SIM A T H、JAFARKARIMI H 和 HEE J M. 教育机构大型数据库的知识发现。在信息检索和管理：概念、方法、工具和应用程序 2018 年，第 10 章：158-245。https://doi.org/10.4018/978-1-5225-5191-1.Ch010
[7] SENOUSY Y、HANNA W K、SHHEHAB A、RIAD A M、EL-BAKRY H M 和 ELKHAMISY N. 使用监督学习算法进行埃及社会保险大数据挖掘。人工智能评论，2019，33(5)：349–357。https://doi.org/10.18280/ria.330504。
[8] JIJO B T，和 ABDULAZEEZ A M. 基于机器学习决策树算法的分类。应用科学与技术趋势杂志，2021，2(1)：20-28.
[9] GUPTA N、NARAYAN R 和 CHAUDHARI A. 使用天气预报技术实施气象数据分析。国际工程应用科学与技术学报，2018，2(12)：28-31.
[10] BALA A. 数据挖掘中各种聚类算法的比较研究。国际科学与工程高级研究杂志，2018，7(7)：1182-1191.
[11] ANDRY J F、GUNADI J、REMBULAN G D 和 TANNADY H. 使用 快速矿工分类的特斯拉大数据实施。国际非线性分析与应用杂志，2021，12（特刊）：2057-2066。http://dx.doi.org/10.22075/ijnaa.2021.6016
[12] SAFRI Y F、ARIFUDIN R 和 MUSLIM M A. K-最近邻和朴素贝叶斯分类算法在确定印度尼西亚向穷人提供健康卡的分类中。信息学科学杂志，2018，5(1)：9-18. https://doi.org/10.15294/sji.v5i1.12057
[13] MADYATMADJA E D、JORDAN S I 和 ANDRY J. F. 大数据分析使用 快速矿工工作室预测几个国家的自杀率，ICIC 快报 B 部分：应用，2021，12(8)：757-764，https://doi .org/10.24507/icicelb.12.08.757。
[14] MADYATMADJA E D、MARVEL、ANDRY J F、TANNADY H. 和 CHAKIR A. 使用集群分析在医院实施大数据。2021 年信息管理与技术国际会议(ICIM 科技)论文集：496-500。
[15] ANDRY J F, REYNALDO S A, CHRISTIANTO K, 等。使用分类、关联和聚类分析 YouTube 上热门视频的算法。2021 年数据与软件工程国际会议(ICoDSE)论文集，IEEE 目录号：CFP21AWL-USB。
[16] MADYATMADJA E D, SEMBIRING D J. M, PERANGIN ANGIN S M, FERDY D 和 ANDRY J. F. 教育机构中的大数据使用快速矿工预测学习效果，计算机科学杂志，2021，17(4)：403-413，https://doi.org/10.3844/jcssp.2021.403.413
[17] SILALAHI R M P、ANDRY J F、BERNANDA D Y、TANNADY H 和 ENIRIANTI。图书馆大数据分析给分类图书出版商，《积极学校心理学杂志》，2022，6(2)：4303 – 4310，
[18] ANDRY J F、TANNADY H、LIMAWAL I I、REMBULAN G D 和 MARTA R F。使用画面在 YouTube 上进行大数据分析。理论与应用信息技术学报，2021，99(22)：5460-5469.
[19] MADYATMADJA E D、RIANTO A、ANDRY J F、TANNADY H 和 CHAKIR A. 使用决策树算法分析医疗保健中的大数据，2021 年第一届计算机科学与人工智能国际会议(ICCSAI)论文集，IEEE 部件号：CFP19H83-ART
[20] MARTÍNEZ-PLUMED F、OCHANDO L. C、FERRI C 等。脆-DM 二十年后：从数据挖掘过程到数据科学轨迹。IEEE 知识和数据工程交易，2021，33(8)：3048-3061，https://doi.org/10.1109/tkde.2019.2962680。
[21] SCHRÖER C、KRUSE F 和 GÓMEZ J M。关于应用脆-DM 过程模型的系统文献综述。普罗西迪亚计算机科学，2021，181：526-534。
[22] PLOTNIKOVA V、DUMAS M 和 MILANI F. 数据挖掘方法的改编：系统文献综述，PeerJ 计算机科学，2020；6：e267，https://doi.org/10.7717/peerj-cs.267。
[23] SCHÄFER F、ZEISELMAIR C 和 BECKER J. 综合脆-DM 和质量管理：生产过程的数据挖掘方法 2020 IEEE 技术管理、运营和决策国际会议，2018：190-195，https ://doi.org/10.1109/ITMC.2018.8691266。