Published online on the page : https://ejournal.uinbukittinggi.ac.id/index.php/ijokid/

Knowbase : International Journal of Knowledge in Database

| ISSN (Print) 2798-0758 | ISSN (Online) 2797-7501 |

Analysis of A Priori Algorithm in Medical Data for Heart Disease Identification with Association Rule Mining

Davip Sutejo[1], *Villa Indra Yudha Adhi Jaya*[2], *Slamet Kacung*[3,*]

[1,2,3] *Universitas Dr. Soetomo, Surabaya,* Indonesia

Article Information

Article History:

Submitted: December 5, 2024
Revision: December 12, 2024
Accepted: December 20, 2024
Published: December 25, 2024

Keywords

Association Rule Mining,
A Priori Algorithm,
Heart Disease

Correspondence

E-mail: slamet@unitomo.ac.id*

A B S T R A C T

Heart disease is one of the leading causes of death worldwide, so it is important to identify risk factors that can contribute to the development of this disease in order to carry out early prevention. This study aims to identify patterns of association between risk variables and the incidence of heart disease using the Association Rule Mining (ARM) method combined with the A priori algorithm. The data used in this study includes lifestyle information, medical history, and other health parameters, obtained from the UCI Machine Learning repository. The analysis results showed that with a support value between 30% and 70%, the strongest association rule was found between sex (sex = 1) and angina (exang = 1), with a lift value of 1.67, indicating a strong positive relationship towards a positive diagnosis (target = 1). In addition, other moderate association rules were found, such as the combination of cp_1 = 1 and ca_0 = 1, with a lift value of about 0.73, indicating a weaker association. These findings suggest that some attribute combinations have higher predictive power, which can be used to improve prediction models in the medical diagnosis of heart disease. This research also highlights the main challenges faced by the A priori algorithm, such as computational complexity and selecting the right threshold to obtain significant rules.

1. Introduction

Association Rule Mining (ARM) is one of the data mining tasks used to discover hidden patterns or relationships between variables in a dataset. This task has been widely used in various fields, ranging from marketing to healthcare, to extract information that can provide deeper insights into existing data. In healthcare, particularly in the diagnosis and prediction of heart disease, ARM can play an important role in identifying patterns that may not be immediately apparent through conventional analysis [1], [2].

Coronary heart disease is one of the leading causes of death worldwide, accurate heart disease prediction can help in early detection, potentially saving many lives [3]. By identifying key risk factors that contribute to heart disease, these factors can range from physical symptoms such as chest pain, high blood pressure, diabetes, to behavioral patterns such as smoking or physical inactivity. Therefore, it is important to identify the relationship between medical traits that may indicate a person's risk or likelihood of suffering from heart disease [4], [5].

Association Rule Mining is used to explore the relationships that may exist between various risk factors and the presence of heart disease in patients. Using algorithms such as Apriori, ARM can help identify frequent item sets (combinations of medical features that often co-occur) and generate association rules that can provide insight into how those features relate to each other. For example, is there a pattern that suggests that patients with a certain type of chest pain and no angina are more likely to have heart disease [6].

This research aims to apply Association Rule Mining to a heart disease dataset to find association rules that connect various medical features such as the type of chest pain, exercise-induced angina, and the presence of heart disease itself. By discovering these relationships, it is expected to help medical professionals in diagnosing and forecasting the likelihood of heart disease occurrence in patients based on existing data [7].

This method also has the potential to provide support for medical decision-making [8], enabling doctors and cardiologists to more easily understand and utilize patterns emerging from patient data to improve the effectiveness of heart disease prevention, diagnosis, and treatment. Through this research, it is hoped that relevant and significant patterns can be found that can be used as the basis for more appropriate heart disease management strategies [9].

Finding patterns that correspond to certain cardiac problems is the main goal of the application of ARM in the diagnosis of heart disease. For instance, ARM has been effectively utilized to identify correlations between age, exercise-induced angina, the existence of heart disease, and the forms of chest pain in studies using datasets like the Heart Disease dataset. ARM algorithms like Apriori have found common itemsets in these data, such as the combination of exercise-induced angina and the type of chest discomfort, which is suggestive of a higher risk of heart disease .

This literature review explores the application of ARM in heart disease research and its impact on disease prediction, diagnosis, and treatment. Previous research in the 2022 study by Zarei et al., used association rules mined from medical datasets containing cardiovascular risk factors to predict the likelihood of heart disease based on patient data. The results showed that factors such as diabetes, hypertension, and family history significantly increase the risk of developing heart disease [10], [11].

In order to find correlations between various clinical characteristics and the risk of acquiring cardiovascular diseases, recent research has used ARM to analyze datasets related to heart disease. Frequent Itemset Mining, for example, has been used to identify common feature combinations that are commonly found among individuals with heart disease, such as age, smoking, and high cholesterol [12].

Furthermore, in therapeutic contexts, the interpretability of the rules produced by ARM may be a challenge. Even if ARM offers valuable insights, healthcare professionals may find it challenging to understand the guidelines without further context. Therefore, to guarantee that the patterns found by ARM are applicable to healthcare decision-making and actionable, efficient visualization approaches and expert collaboration are required [13].

Association rules Mining (ARM) extracted from a transactional dataset consisting of a single set of features $I=\{i_1, \ldots, i_n\}$ of n binary attributes called items and a set $D=\{t_1, \ldots, t_n\}$, $t_k \subseteq I$, of transactions called database. An association rule is a pair of itemsets (X, Y), often denoted by an implication of the form $X \Rightarrow Y$, where X is the antecedent, Y is the consequent. The support of an itemset X determines how often it appears in the transactional database. The support of an association rule $X \Rightarrow Y$ can be defined as the percentage of transactions among the total transactions that contain both itemsets X and Y, which is shown in Eq. 1. [14]

$$Support\ (X \rightarrow Y) = \frac{|(X \cup Y)|}{|D|} \tag{1}$$

How often items in Y occur in transactions that contain X depends on the confidence of an association rule $X \Rightarrow Y$. As determined by Eq. 2, [14] the confidence of a rule is the proportion of transactions that contain itemset X that also contain itemset Y to the total number of records that contain X.

$$Confidence\ (X \rightarrow Y) = \frac{|(X \cup Y)|}{|X|} \tag{2}$$

2. Method

The research method is needed so that the research can be structured so that the results obtained are in accordance with the research objectives. Therefore, this research uses a flow chart to provide a clear picture of the stages of the research carried out as shown in Figure 1.

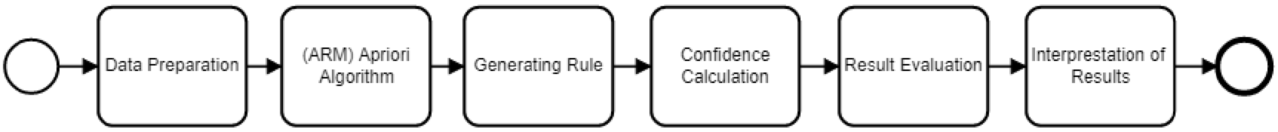

Figure 1. Association Rule Mining Method

Figure 1. shows the research stage that will be carried out by entering the sample data to be analyzed, then the Apriori Algorithm serves to find of patterns / combinations in a set of itemsets

2.1. *Data Preparation*

Data preparation follows a series of steps that start with collecting data that is ready to be used for the data mining process. A dataset containing heart disease information taken from UCI machine learning as determined by table 1,[15] then the data is prepared by converting the features into binary format as determined by table 2.

Tabel 1. Heart Disease Risk Facktor

No.	Feature	Description	Data Type
1	Age	Age of the patient (in years)	Integer
2	Sex	Sex of the patient (1 = male, 0 = female)	Categorical
3	cp (chest pain type)	Type of chest pain (0 = asymptomatic, 1 = mild chest pain, 2 = moderate chest pain, 3 = severe chest pain)	Categorical
4	Trestbps	Resting blood pressure (in mm Hg)	Integer
5	Chol	Serum cholesterol level (in mg/dL)	Integer
6	fbs (fasting blood sugar)	Fasting blood sugar (1 = greater than 120 mg/dL, 0 = less than 120 mg/dL)	Categorical
7	restecg (resting electrocardiogram)	Resting electrocardiogram result (0 = normal, 1 = ST depression, 2 = ST flattening)	Categorical
8	Thalach	Maximum heart rate achieved (beats per minute)	Integer
9	Exang	Exercise-induced angina (1 = Yes, 0 = No)	Categorical
10	Oldpeak	ST depression induced by exercise relative to rest	Integer
11	Slope	Slope of the peak exercise ST segment (0 = upsloping, 1 = flat, 2 = downsloping)	Categorical
12	ca (number of major vessels colored)	Number of major vessels colored by fluoroscopy (0-3)	Integer
13	Thal	Thalassemia (1 = normal, 2 = fixed defect, 3 = reversable defect)	Categorical
14	Class	Heart disease status (1 = positive, 0 = negative)	Integer

This dataset is usually used to predict a target variable, which indicates whether a patient has heart disease or not based on other medical features. The heart disease dataset consists of 14 attributes and 270 rows, a partial sample of the data is shown in Table 2.

Tabel 2. Dataset Heart Disease

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	1	145	233	1	2	150	0	2.3	3	0	6	1
67	1	4	160	286	0	2	108	1	1.5	1	3	3	1
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	1	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
56	1	1	120	236	0	2	178	0	0.8	1	0	3	1
62	1	3	140	247	0	2	170	0	0.6	2	0	3	1
57	1	4	120	267	0	2	160	0	1	1	0	3	1
63	1	4	130	272	0	2	174	0	1	1	0	3	1
53	1	2	122	223	0	2	175	0	0	1	0	3	0
54	1	1	130	253	0	2	163	0	0.2	1	0	3	1
61	0	3	140	239	0	2	142	1	2	2	1	7	1
48	1	4	130	236	0	2	160	1	1.1	1	0	7	1
53	0	1	140	236	0	0	185	0	0	1	0	3	0
55	1	3	120	245	0	2	170	0	1.4	1	0	3	1
43	1	2	130	233	0	2	179	0	2.6	1	0	3	0
40	1	2	130	238	0	2	160	1	3.1	1	0	3	0
59	1	4	130	236	0	2	150	1	0.8	1	0	3	1
55	1	3	140	221	0	0	145	0	2	2	0	3	0

The heart disease dataset in Table 2 is data taken from the UCI Machine Learning Repository and must be represented in a tabular (binary format) such as Table 3.

Tabel 3. Feature Binary Format

age	sex	cp_1	cp_2	cp_3	cp_4	trestbps	chol	fbs0	fbs1	ekg	exangno	exang yes	slope_1	slope_2	slope_3	ca_0	ca_1	ca_2	ca_3	thal	target
1	1	0	0	1	0	1	1	1	0	1	0	0	1	0	0	0	1	0	0	1	1
1	1	1	0	0	0	0	1	1	0	1	0	0	1	0	0	0	1	0	0	1	1
1	1	0	0	1	0	1	1	1	0	1	0	0	1	0	0	0	1	0	0	1	1
1	1	0	1	0	0	0	1	1	0	1	0	0	1	0	0	0	1	0	0	1	0
1	1	1	0	0	0	1	1	0	1	1	0	0	1	0	0	0	1	0	0	1	1
1	0	0	1	0	0	1	1	0	1	0	1	0	0	0	1	0	1	0	0	0	1
0	1	1	0	0	0	1	1	1	0	1	1	0	1	0	0	0	0	1	0	0	1
1	0	1	0	0	0	1	1	1	0	0	0	0	1	0	0	0	1	0	0	1	0
1	1	0	1	0	0	0	1	1	0	1	0	0	0	1	0	1	0	0	0	0	1
0	1	0	1	0	0	1	1	1	0	1	0	0	1	0	0	1	0	0	0	1	0
0	1	0	1	0	0	1	1	0	1	1	1	0	0	1	0	1	0	0	0	1	0
1	1	1	0	0	0	1	1	0	1	1	1	0	0	1	0	0	1	0	0	1	1
1	1	0	1	0	0	1	1	1	0	0	0	0	0	0	1	0	1	0	0	1	0

2.2. *Apriori Algorithm*

Association Rule Mining (ARM) is one of the data mining tasks to find association rules between a combination of items, [16] which often appear together (frequent item sets). itemset is one of the set of items in I, and k-itemset is the set of items containing k items. The A priori algorithm for association rule mining can be used in two steps [17]. Finding all of the item sets from the database that appear frequently is the first stage. In this study, the minimum support value provided is 0.3 and confidence 0.7, which is used as the threshold for identifying frequently occurring item sets. In this study use Google Colaboratory, a cloud-based platform provided by Google for Python programming

```
Support and association rule variable:
# Menghitung frequent itemsets dengan minimum support 0.3
frequent_itemsets = apriori(df, min_support=0.3, use_colnames=True)

# Membuat aturan asosiasi
rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.7)
```

```
Apriori Algorithm variable:

Cκ is candidate item sets k
  Lκ is frequent item sets k
Start:
  Find frequent set Lκ-1
  Generate Cκ using CP Lκ-1, i.e Lκ-1 x Lκ-1
  Pruning remove any k-1 size itemset are not frequent
  Return frequent set Lκ-1
END:
```

For example {sex=1, target=1} is a 2-itemset and {sex=1, target=1, thal3=1} is a 3-itemset. Frequent itemset indicates itemsets that have a frequency of occurrence of more than a predetermined minimum value. Here are some examples that show the frequency of occurrence in the following Table 4 shows the candidate 1-itemset of heart disease data, Table 5 shows the candidate 2-itemset, and Table 6 the candidate 3-itemset.

Table 4. Example of 1-Itemset Candidates

Item	Frequency
sex = 1	9
cp_1 = 1	7
cp_2 = 1	4
cp_3 = 1	3
cp_4 = 1	1
fbs = 1	4
restecg_0 = 1	4
restecg_1 = 1	6
restecg_2 = 1	4
exang = 1	6
slope_1 = 1	6
slope_2 = 1	4
slope_3 = 1	2
ca_0 = 1	7
ca_1 = 1	3
ca_2 = 1	2
ca_3 = 1	1
thal_3 = 1	7
thal_6 = 1	5
thal_7 = 1	3
target = 1	9

From Table 4 above, if the threshold value = 5 is set, then : F1 = {sex=1}, {cp_1}, { restecg_1 = 1}, { exang = 1}, { slope_1 = 1}, { ca_0 = 1}, { thal_3 = 1}, { thal_6 = 1}, and { target = 1}, and the same applies to Table 5, and Table 6.

Table 5. Example of 2-Itemset Candidates

Item sets	Frequency
sex = 1, target = 1	7
cp_1 = 1, ca_0 = 1	5
cp_1 = 1, thal_3 = 1	4
restecg_1 = 1, exang = 1	4
cp_2 = 1, restecg_1 = 1	3
slope_1 = 1, exang = 1	4
thal_3 = 1, target = 1	5
ca_0 = 1, thal_3 = 1	4
ca_0 = 1, target = 1	5
sex = 1, exang = 1	4

Some item sets occur more frequently than others, suggesting common combinations of features in patients with the target condition. For example, sex = 1, target = 1 (7 occurrences) indicates a higher frequency of male patients with the target condition.

Table 6. Example of 3-Itemset Candidates

Item Set	Frequency
sex = 1, target = 1, thal_3 = 1	4
cp_1 = 1, ca_0 = 1, thal_3 = 1	3
cp_1 = 1, ca_0 = 1, target = 1	3
sex = 1, exang = 1, target = 1	3
cp_2 = 1, restecg_1 = 1, exang = 1	3
slope_1 = 1, exang = 1, target = 1	3

This set of itemsets shows interesting patterns and associations in the dataset. By analyzing these combinations, we can derive useful insights into how different features like chest pain, exercise-induced angina, ECG results, and vessel blockage interact and contribute to the risk of heart disease (or the target condition).

2.3. *Generating Rules*

Generating association rules usually involves identifying a set of items that frequently appear in a data set and then deriving rules from that set of items. These rules are created based on two main components: Antecedent (the "If" part) and Consequent (the "Then" part), with associated metrics such as Support, and Confidence [18]. As an example seen in Table 7.

Table 7. Example of Generated Rules with Metrics

Antecedent	Consequent	Support	Confidence
(Sex (Male = 1))	(Fasting Blood Sugar (≤120))	0.45	0.9
(Fasting Blood Sugar (≤120))	(Exercise Angina (No = 0))	0.4	0.8
(Fasting Blood Sugar (≤120), Exercise Angina (No = 0))	(Target (Heart Disease - Yes = 1))	0.35	0.87

2.4. *Confidence Calculation*

Confidence is calculated to assess the strength of the relationship between features in an association rule, the confidence of an association rule A→B (i.e., "If A, then B") is calculated as the probability of the consequent (B) occurring, given that the antecedent (A) has occurred. In simpler terms, confidence measures how likely the rule is to hold true when the antecedent is true, which is shown in Table 8 and Table 9.

Tabel 8. Confidence for Frequent 2-itemsets :

Itemsets	Frequency	Support	Confidence (A → B)
sex = 1, target = 1	7	70%	78%
cp_1 = 1, ca_0 = 1	5	50%	71%
cp_1 = 1, thal_3 = 1	4	40%	57%
restecg_1 = 1, exang = 1	4	40%	67%
cp_2 = 1, restecg_1 = 1	3	30%	75%
slope_1 = 1, exang = 1	4	40%	67%
thal_3 = 1, target = 1	5	50%	71%
ca_0 = 1, thal_3 = 1	4	40%	57%
ca_0 = 1, target = 1	5	50%	71%
sex = 1, exang = 1	4	40%	67%

Tabel 9. Confidence for Frequent 3-itemsets :

Item Set	Frequency	Support	Confidence (A → B)
sex = 1, target = 1, thal_3 = 1	4	40%	44%
cp_1 = 1, ca_0 = 1, thal_3 = 1	3	30%	43%
cp_1 = 1, ca_0 = 1, target = 1	3	30%	43%
sex = 1, exang = 1, target = 1	3	30%	43%
cp_2 = 1, restecg_1 = 1, exang = 1	3	30%	43%
slope_1 = 1, exang = 1, target = 1	3	30%	43%

2.5. *Result Evaluation*

To evaluate the results of association rules, this research uses several metrics that help measure whether the rules are significant or not. The main metrics used for evaluation are Support, Confidence, and Lift. The following is a detailed description of these evaluation metric :

Lift Metric : Lift measures the strength of the rule relative to the occurrence of the consequent *B* without the antecedent *A*. A lift greater than 1 indicates a positive association between *A* and *B*. *Lift is calculated as :*[19], [20]

$$lift\ (A \rightarrow B) = \frac{Confidence\ (A \rightarrow B)}{Support(B)} \tag{3}$$

Interpretation :

- Lift > 1 : The occurrence of *A* increases the likelihood of *B* (positif correlation)
- Lift = 1 : *A* and *B* are independent (no correlation)
- Lift < 1 : The occurrence of *A* decreases the likelihood of *B* (negative correlation).

2.6. *Interpretation of Results*

The interpretation of the results from association rule mining involves analyzing the generated association rules based on metrics such as Support, Confidence, and Lift. This section will interpret these rules in the context of the dataset and provide actionable insights based on the association patterns observed. The interpretation helps us understand the relationships between different features (attributes) and how they might be used in decision-making. Relevant findings are interpreted to provide insight into the interplay between features in heart disease.

Interpretation of Association Rules Example. If {cp_1=1}, then {ca_0=1} has support 50%, confidence 71%, and lift 1.42 this indicates the the likelihood of {ca_0 =1} occurring when {cp_1=1} and {ca_0=1} are more likely to appear together than by chance. If {thal_3=1}, then {target =1} has support 50%, confidence 70%, and lift 1.4 this indicates that the presence of {thal_3=1} increase the likelihood of {target=1} is likely to follow. The lift value further emphasizes that this association is statistically significant.

3. Results and Discussion

The data consists of 270 records with 19 relevant attributes. Detailed information regarding the data used in this study is presented in Table 2. The results obtained from the association rule mining of the dataset provide significant insights into the relationships between various attributes, which can be used to make data-driven decisions. This section discusses the key findings from the generated association rules, analyzes the patterns identified, and explores their practical implications [21].

3.1. *Analysis of the Association Rules*

Frequent Item sets

Frequent item sets are those combinations of features that appear together in a significant number of transactions [22]. By identifying these itemsets, we are able to recognize which features commonly co-occur and may have a meaningful relationship. The pseudocode of the algorithm is summarized in Algorithm:

```
Frequent item sets[23]
variable:
  Ci is candidate item sets i
  Li is frequent item sets i
Start:
  L1 ← {large 1-item sets}
  k ← 2
  while Li-1 ≠ 0
  Ck generate (Li-1)
For transaction t € T
    Ct ← subset (Ck,t)
For candidate c € Ct
```

```
    Count {c}← {count{c}+1}
    Lᵢ ← { c € Cᵢ count|c| ≥ €}
    i=i+1
  Return Lᵢ₋₁
END:
```

Formation of Frequent 2-Item sets

Frequent 2-item sets are combinations of two different items that frequently appear together in a dataset. These are a specific type of frequent itemset, where the goal is to find pairs of items that occur together in the dataset more frequently than a given threshold (called support). The results of the combination of the two items are shown in table 10.

Table 10. Frequent 2-Item Sets.

Id	Item Sets	Frequency	Support
DT1	sex = 1, target = 1	7	70%
DT2	cp_1 = 1, ca_0 = 1	5	50%
DT3	cp_1 = 1, thal_3 = 1	4	40%
DT4	restecg_1 = 1, exang = 1	4	40%
DT5	cp_2 = 1, restecg_1 = 1	3	30%
DT6	slope_1 = 1, exang = 1	4	40%
DT7	thal_3 = 1, target = 1	5	50%
DT8	ca_0 = 1, thal_3 = 1	4	40%
DT9	ca_0 = 1, target = 1	5	50%
DT10	sex = 1, exang = 1	4	40%

The most frequent item sets are DT1 (sex = 1, target = 1), This item set appears 7 times, with a support of 70%. its means that 70% of the transactions in the dataset involve male patients who may have heart disease. DT2 (cp_1 = 1, ca_0 = 1) : This item set appears 5 times, with a support of 50%. This indicates that 50% of the transactions involve patients with chest pain type 1 (cp_1 = mild chest pain) and no coronary artery blockage (ca_0 = 1). There are a few notable relationships DT7 and DT9 : DT7 (thal_3 = 1, target = 1) shows that having a normal thalassemia result is associated with the Heart disease condition in 50% of the cases. DT9 (ca_0 = 1, target = 1) shows that not having coronary artery blockage correlates strongly with having the heart disease condition

Table 11. Confidence and Lift Value from Frequent 2-Itemsets

Item Sets	Frequency	Support	Confidence (A → B)	Lift (A → B)
sex = 1, target = 1	7	70%	78%	0.87
cp_1 = 1, ca_0 = 1	5	50%	71%	1.42
cp_1 = 1, thal_3 = 1	4	40%	57%	1.43
restecg_1 = 1, exang = 1	4	40%	67%	1.67
cp_2 = 1, restecg_1 = 1	3	30%	75%	2.5
slope_1 = 1, exang = 1	4	40%	67%	1.67
thal_3 = 1, target = 1	5	50%	71%	1.42
ca_0 = 1, thal_3 = 1	4	40%	57%	1.43
ca_0 = 1, target = 1	5	50%	71%	1.42
sex = 1, exang = 1	4	40%	67%	1.67

{Sex = 1, target = 1}, Support = (70%) and Confidence (78%) : There is a 78% chance that if a patient is male (gender = 1), they will also have potential heart disease (target = 1). with a Lift value (0.87): The lift value is 0.87, which indicates a negative correlation between the two items. This means that the existence of sex = 1 and target = 1 is unlikely to occur, indicating that male patients may not have a chance of developing heart disease.

Formation of Frequent 3-Item sets

Frequent 2-item sets are combinations of three different items that frequently appear together in a dataset. These are a specific type of frequent itemset, where the goal is to find pairs of items that occur together in the dataset more frequently than a given threshold). The results of the combination of the three items are shown in table 11.

Table 12. Frequent 3-Item Sets

Id	Item Set	Frequency	Support
DT1	sex = 1, target = 1, thal_3 = 1	4	40%
DT2	cp_1 = 1, ca_0 = 1, thal_3 = 1	3	30%
DT3	cp_1 = 1, ca_0 = 1, target = 1	3	30%
DT4	sex = 1, exang = 1, target = 1	3	30%
DT5	cp_2 = 1, restecg_1 = 1, exang = 1	3	30%
DT6	slope_1 = 1, exang = 1, target = 1	3	30%

Most Frequent Item Set: DT1 (sex = 1, target = 1, thal_3 = 1) with a frequency of 4 and 40% support is the most frequent 3-item set. This suggests that male patients with the target condition are most likely to also have a normal thalassemia result. Equal Frequency for Other Item Sets: The other item sets have the same frequency (3 occurrences, or 30% support), indicating that these patterns, while not as common as DT1, still represent notable co-occurrences in the dataset.

Table 13. Confidence and Lift Value from Frequent 3-Itemsets

Item Set	Frequency	Support	Confidence (A → B → C)	Confidence (C → A → B)	Lift (A → B → C)	Lift (B → A → C)	Lift (C → A → B)
sex = 1, target = 1, thal_3 = 1	4	40%	44%	0.57	0.63	0.63	0.81
cp_1 = 1, ca_0 = 1, thal_3 = 1	3	30%	43%	0.57	0.73	0.73	0.81
cp_1 = 1, ca_0 = 1, target = 1	3	30%	43%	0.57	0.73	0.73	0.81
sex = 1, exang = 1, target = 1	3	30%	50%	0.67	1.67	1.67	2.33

Table 13 show results from association rule mining for frequent 3-itemsets, likely from a dataset concerning factors such as "sex," "target," "thal," "cp," and others, which could be related to clinical health data (e.g., heart disease diagnosis). This rule shows a much stronger association than the previous ones. The combination of sex = 1, exang = 1, and target = 1 occurs in 30% of the records, and the confidence of 50% indicates that given sex = male and exang = 1, there is a 50% chance that target = heart disease diagnosis will occur. The lift value of 1.67 suggests a stronger-than-expected relationship, meaning these conditions are more likely

3.2. Association Rules Formation

Table 14. Confidence and Lift Value from Frequent 2-Itemsets

Rules	Support	Confidence (A → B)	Lift (A → B)
IF sex=1 **THEN** target = 1	70%	78%	0.87
IF cp_1 = 1 **THEN** ca_0 = 1	50%	71%	1.42
IF cp_1 = 1 **THEN** thal_3 = 1	40%	57%	1.43
IF restecg_1 = 1 **THEN** exang = 1	40%	67%	1.67
IF cp_2 = 1 **THEN** restecg_1 = 1	30%	75%	2.5
IF slope_1 **THEN** exang = 1	40%	67%	1.67
IF thal_3 = 1 **THEN** target=1	50%	71%	1.42
IF ca_0 = 1 **THEN** thal_3 = 1	40%	57%	1.43
IF ca_0 = 1 **THEN** target=1	50%	71%	1.42
IF sex=1 **THEN** exang = 1	40%	67%	1.67

The table 14 above results in three associations, Strong Positive Associations : The rules with Lift > 1 suggest positive relationships. Rules like cp_2 = 1 → restecg_1 = 1 (Lift = 2.5) and restecg_1 = 1 → exang = 1 (Lift = 1.67) suggest that these associations are stronger than expected.

Moderate Associations: Rules like sex = 1 → target = 1 (Lift = 0.87) and ca_0 = 1 → target = 1 (Lift = 1.42) indicate positive associations, but the lift values suggest that the associations are less pronounced compared to others.

Confidence Values: The confidence values range from 57% to 78%, showing that these rules have a moderate to high likelihood of the consequent (B) occurring when the antecedent (A) is true. For instance, the rule IF sex = 1 THEN target = 1 has a confidence of 78%, suggesting a relatively strong association.

Confidence Values: The confidence values range from 57% to 78%, showing that these rules have a moderate to high likelihood of the consequent (B) occurring when the antecedent (A) is true. For instance, the rule IF sex = 1 THEN target = 1 has a confidence of 78%, suggesting a relatively strong association.

Table 11. Confidence and Lift Value from Frequent 3-Itemsets

Rules	Support	Confidence (A → B → C)	Confidence (C → A → B)	Lift (A → B → C)
IF sex = 1 AND target = 1, **THEN** thal_3 = 1	40%	44%	57%	0.63
IF cp_1 = 1 AND ca_0 = 1, **THEN** thal_3 = 1	30%	43%	57%	0.73
IF cp_1 = 1 AND ca_0 = 1, **THEN** target = 1	30%	43%	57%	0.73
IF sex = 1 AND exang = 1, **THEN** target = 1	30%	50%	67%	1.67

Rule 4 (sex = 1 AND exang = 1 → target = 1) has the highest lift (1.67), indicating a strong positive association between these attributes. Rules with lift values below 1 (like Rules 1, 2, and 3) suggest weaker associations than expected.

This could imply that while these itemsets are frequent in the dataset, they are not as strongly related as other itemsets. Confidence values are relatively high (ranging from 44% to 67%), indicating moderate-to-strong probabilities for the consequent (C) occurring when the antecedents (A → B) are true.

4. Conclusion

Based on the analysis of the association rules derived from the dataset, we can draw several important conclusions regarding the relationships between various attributes. Knowledge are valuable for understanding how different factors are interconnected and how they influence each other. Referring to the results of the research conducted, it can be concluded that the application of A Priori in detecting heart disease provides significant results. The findings show that the rule with the highest support (30%), confidence (67%), and highest Lift (1,67) is: "If the patient is male and Exang=Yes, it is likely that the patient will be diagnosed with heart disease. For further development, it is necessary to use other algorithms in using association rule mining such as FP-Growth, this is because A Priori has limitations that require a lot of calculations in managing large item sets.

References

[1] M. M. A. H. Zarei, S. M. M. Fakhraei, "A novel association rule mining-based algorithm for healthcare data analysis," *J. Healthc. Eng.*, pp. 1–10, 2021.
[2] P. S. S. Singh, M. M. S. Soni, "Data mining techniques and tools for healthcare data," *Int. J. Comput. Appl.*, vol. 174, no. 3, pp. 10–17, 2021.
[3] A. Yazdani, K. D. Varathan, Y. K. Chiam, A. W. Malik, and W. A. Wan Ahmad, "A novel approach for heart disease prediction using strength scores with significant predictors," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, pp. 1–16, 2021.
[4] E. Yunianto, "Prediksi penyakit kardiovaskular berbasis asosiation rule," 2017.
[5] and H. Z. M. Aljarah, S. M. M. Fakhraei, "Predicting cardiovascular disease using machine learning: A systematic review," *Healthcare*, vol. 8, no. 3, pp. 1–16, 2022.
[6] and F. T. V. C. A. Rojas, J. C. Peña, "Association rule mining and its application in medical diagnosis: A review," *J. Appl. Res. Technol.*, vol. 19, no. 2, pp. 113–124, 2021.
[7] and D. N. R. A. R. Singh, K. R. Singh, "Analysis of cardiovascular diseases using association rule mining," *Int. J. Data Sci. Anal.*, vol. 3, no. 4, pp. 200–210, 2021.
[8] K. Rameshkumar, M. Sambath, and S. Ravi, "Relevant association rule mining from medical dataset using new irrelevant rule elimination technique," *2013 Int. Conf. Inf. Commun. Embed. Syst. ICICES 2013*, pp. 300–304, 2013.
[9] R. B. Tiwari and A. G. Sharma, "Predictive modeling of heart disease using association rule mining and neural

network," *Int. J. Eng. Technol.*, vol. 13, no. 3, pp. 1124–1131, 2023.

[10] S. M. M. F. H. Zarei, M. Aljarah, "Predicting cardiovascular disease using machine learning: A systematic review," *Healthcare*, vol. 8, no. 3, pp. 1–16, 2022.

[11] S. Darrab, D. Broneske, and G. Saake, "Exploring the predictive factors of heart disease using rare association rule mining," *Sci. Rep.*, vol. 14, no. 1, pp. 1–26, 2024.

[12] M. Kaushik, R. Sharma, I. Fister, and D. Draheim, "Numerical Association Rule Mining: A Systematic Literature Review," vol. 1, no. 1, 2023.

[13] H. Haryanto, H. Winarto, and C. Juliane, "Application of Data Mining Techniques in Healthcare: Identifying Inter-Disease Relationships through Association Rule Mining," *J. World Sci.*, vol. 3, no. 5, pp. 520–529, 2024.

[14] X. Zhang and J. Zhang, "Analysis and research on library user behavior based on apriori algorithm," *Meas. Sensors*, vol. 27, no. April, p. 100802, 2023.

[15] UCI Machine Learning Repository, "Heart Disease." [Online]. Available: https://archive.ics.uci.edu/dataset/45/heart+disease. [Accessed: 30-Dec-2024].

[16] Kusrini, Luthfi T, E., *Algoritma Data Mining. Buku Algoritma Data Mining*. Andi Offset, Yogyakarta, 2009.

[17] M. H. Santoso, "Application of Association Rule Method Using Apriori Algorithm to Find Sales Patterns Case Study of Indomaret Tanjung Anom," vol. 1, no. 2, pp. 54–66, 2021.

[18] C. Mihaescu, "Mining Frequent Itemsets – Apriori Algorithm," *Lab. Modul. 8*, 2012.

[19] E. Hikmawati, N. U. Maulidevi, and K. Surendro, "Minimum threshold determination method based on dataset characteristics in association rule mining," *J. Big Data*, vol. 8, no. 1, 2021.

[20] B. Mudumba and M. F. Kabir, "Mine-first association rule mining: An integration of independent frequent patterns in distributed environments," *Decis. Anal. J.*, vol. 10, no. February, p. 100434, 2024.

[21] J. Tempaiboolkul, "Mining rare association rules in a distributed environment using multiple minimum supports," *2013 IEEE/ACIS 12th Int. Conf. Comput. Inf. Sci. ICIS 2013 - Proc.*, pp. 295–299, 2013.

[22] M. I. Habibi, A. Nazir, E. Haerani, and E. Budianita, "Application of Data Mining for Ceramic Sales Data Association Using Apriori Algorithm," vol. 05, no. 02, pp. 1–11, 2024.

[23] F. A. Rohmah, S. Kacung, and E. Prihartono, "Association Rule Mining To Enhance Sata Bottle Sales," vol. 04, no. 01, pp. 28–37, 2024.