A REPORT

ON

**Bayesian Association Rules Mining Algorithm**

**using Hungarian Heart Disease Dataset**

PREPARED FOR

PROF. BASABDATTA BHATTACHARYA

*(DEPARTMENT OF CSIS)*

BY

**GROUP 16**

ADITYA KHILLAN 2020A7PS0154G

ARYAN UDESHI 2019B3A70206G

RISHI JAIN 2020A8PS1772G

RISHI ROY 2020A7PS0135G

SARANSH PRUTHI 2019B4A70718G

SHRUTI RASTOGI 2019B4A70802G

Prepared in partial fulfilment of the requirements of the course

Foundations of Data Science (CS F320)

AT



**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE PILANI, GOA CAMPUS**

**(December, 2022)**

**ACKNOWLEDGMENT**

We would like to express our immense gratitude and appreciation towards Prof. Basabdatta Bhattacharya for providing us with the opportunity to work on this project and for her never-ending support and guidance in pursuing this endeavour. She has been a great source of support and inspiration for us, and this would not have been possible without her. The lectures of the course provided us with the foundation to work on the project and report confidently.

We would also like to thank our family and friends for their constant support whenever we felt overwhelmed and for helping resolve our doubts so effectively. We are truly grateful for their presence.

**ABSTRACT**

**Key Words:** Bayesian Association Rule mining algorithm, Apriori algorithm, Bayesian network

**Abstract:** In this report, we have used the Bayesian Association Rule mining algorithm (BAR) which combines the Apriori association rule mining algorithm and Bayesian networks. Out of all the tested algorithms, BAR outputs the best rules according to the Bayesian Confidence (BC) and Bayesian Lift (BL) values. BC and BL are used to quantify the conditional dependence and independence relationships between variables using the joint probabilities given in a Bayesian network. We have tested the accuracy of BAR using the heart disease dataset from the UCI database. We can see that BAR produces the rules with the highest values of confidence, lift and support.

# TABLE OF CONTENTS

ABSTRACT……………………………………………………………………………………………. 1

ACKNOWLEDGEMENTS……………………………………………………………………………. 2

TABLE OF CONTENTS………………………………………………………………………………. 3

LIST OF FIGURES AND TABLES…………………………………………………………………… 4

[I. INTRODUCTION](#_heading=h.x63on6xokhyj)………………………………………………………………………………….... 5

[II. BASIC CONCEPTS](#_heading=h.4hnraj7gfmqn)………………………………………………………………………………… 7

[III. BAYESIAN ASSOCIATION RULES MINING ALGORITHM](#_heading=h.m80117bze93l)…………………………………. 9

[IV. DATA](#_heading=h.yfpxblfpccqk) PREPROCESSING AND [ANONYMIZATIO](#_heading=h.yfpxblfpccqk)N….……………………………………..... 10

[V. EVALUATION](#_heading=h.35icikjo3vc)…………………………………………………………………………………….. 11

VI. CONCLUS[ION](#_heading=h.35icikjo3vc)……………………………………………………………………………………[1](#_heading=h.35icikjo3vc)2

APPENDIX……….…………………………………………………………………………..……… 13

[REFERENCES](#_heading=h.mjjqer7if34u)……………………………………………………………………………………….. 14

# LIST OF FIGURES AND TABLES

FIG 1 : GRAPH REPRESENTATION OF CAUSAL NETWORK……………………………………5

FIG 2 : GRAPH REPRESENTATION OF BAYESIAN NETWORK WITH ASSOCIATED

CPD…………………………………………………………………………………………………….6

FIG 3 : THE BAYESIAN NETWORK REPRESENTING THE RULE A → B……………………9

FIG 4 : MISSING VALUES IN THE PREPROCESSED DATASET………………………………...12

FIG 5 : DATASET POST REMOVAL OF THE NAN VALUES……………………………………..13

FIG 6 : HISTOGRAMS TO ANALYZE THE DISTRIBUTION OF ATTRIBUTES V/S THEIR FREQUENCY…………………………………………………………………………………………14

FIG 7 : BOX-AND-WHISKER PLOTS OF ATTRIBUTES OF HEART DISEASE DATASET, HUNGARY…………………………………………………………………………………………....16

FIG 8 : HEATMAP OF THE HEART DISEASE ATTRIBUTE SET………………………………...18

FIG 9 : PAIRPLOTS OF THE SELECTED ATTRIBUTE CROSS ALL OTHER

ATTRIBUTES…………………………………………………………………………………………19

TABLE 1 : FEATURE SET FOR HEART DISEASE DATASET, HUNGARY……………………...15

TABLE 2 : TOP 10 TWO-ITEM ASSOCIATION RULES RANKED BY BAYESIAN CONFIDENCE………………………………………………………………………………………..17

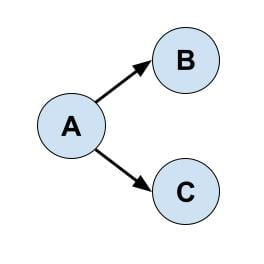
# INTRODUCTION

An Adverse Drug Event (ADE) is any unfavorable and unintended side effect of using a medicinal product whether or not it is directly related to the use of the product. Early ADE detection can help identify any adverse effects of drugs with minimal patient contact and also prevent any harm to the patient as well as billions of dollars worth of social costs.

Two types of associations exist - binary associations and multi-item associations. Association rule mining algorithms have helped to detect new ADE associations. The Apriori association rule mining algorithm has been able to detect multi-item associations more efficiently and accurately. Furthermore, multi-objective evolutionary algorithms have been implemented to find association rules which the Apriori algorithm cannot. In this paper, aim to implement a Bayesian Association Rule mining algorithm (BAR) on select datasets to show its efficiency[5].

**Causal Networks**

Causal networks are a diagrammatic way of representing cause-effect relationships between variables. They can be represented using directed acyclic graphs where the nodes are used to indicate discrete or continuous variables and the arrows indicate direct causal relationships where the arrow points from the cause to the effect.

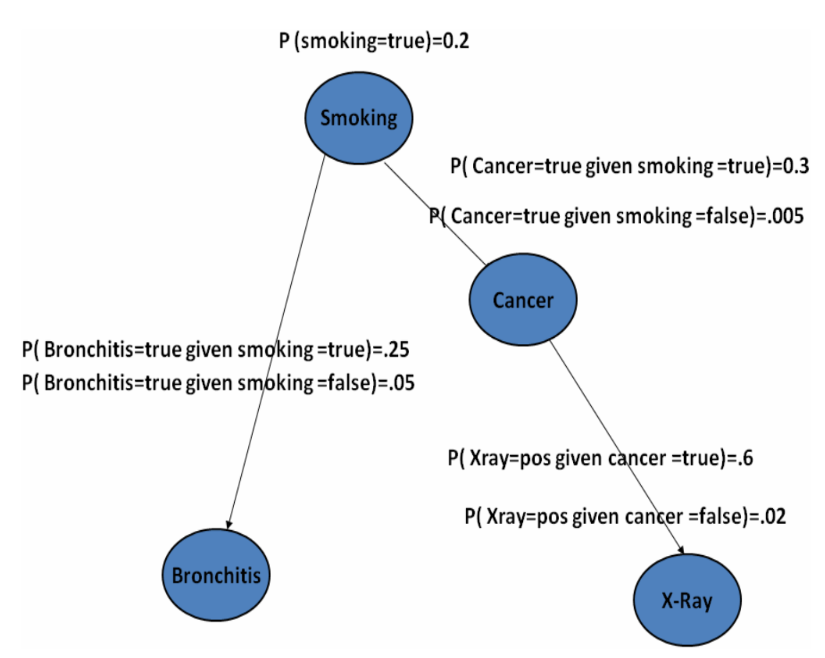


**Fig. 1: Graph representation of a causal network**

Only recently, after years of research have we been able to differentiate between causality and mere correlation using graphical and mathematical models. This has helped us more accurately predict the effects of a change in state of a particular variable[3].

**Bayesian Networks**

Bayesian networks are causal networks with probabilistic relationships between the nodes which are connected by arrows. The combination of the graph and the probabilities represent a unique joint probability distribution for each cause-effect relationship represented by the graph. Each node’s conditional probability distribution(CPD) depends on the state of its parents. The CPD of node Xi is given by P(Xi | Parents(Xi)).



**Fig. 2: Graph representation of bayesian network with associated CPD**

Bayesian Networks are important in many fields including medical where physiological causes can be connected to diagnostic effects such as detecting breast and lung cancer[5].

# BASIC CONCEPTS

**Association Rules**

Assuming A & B are disjoint datasets, we introduce a few terms:

* Association Rule: Implication expression of the form: A⇒B
* Support: 𝑆(𝐴 ∪ 𝐵) of an association rule A⇒B = k/n

{k: number of instances containing all the items of A and B

n: the total number of instances of the dataset}

Support represents the percentage of instances that contain all the items included in

the association rule

* Confidence = 𝑆(𝐴 ∪ 𝐵)/𝑆(𝐴)

Confidence represents how frequently B occurs among all the instances containing A.

It is used to used to measure the reliability or interestingness of the rule

Confidence is an estimate of 𝑃(𝐵∣𝐴) {

* Lift = Confidence/S(B)

Lift value of an association rule is the ratio of the confidence of the rule to the support of B. Lift is a measure of the deviation of the rule from the statistical independency of A and B

The lift is a value between 0 and infinity:

* a value greater than 1 indicates: A and B appear together more often than expected
* a value less than 1 indicates: A and B appear together less often than expected
* a value close to 1 indicates: A and B appear together almost as often as expected.

Apriori Mines Association Rules:

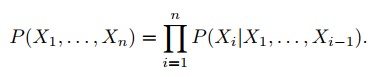
* Generate itemsets and select those itemsets whose supports ≥ the minimum support threshold
* Generate rules from selected itemsets. To generate a rule 𝐴 ⇒ 𝐵 from an itemset, a subset of the itemset forms B and the remaining items forms A.
* Output the rules with the highest confidence and lift.

**Bayesian Networks**

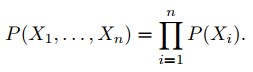
Let X and Y be two disjoint subsets of random variables such that the probability of Y is P(Y)>0. Then, the *conditional probability distribution* (*CPD*) of X given Y is defined as:

CPD: P(X/Y) = P(X,Y)/P(Y)

The joint probability over variables 𝑋1, ..., 𝑋𝑛 is defined as:

****

A set of variables 𝑋 = {𝑋1, . . . , 𝑋𝑛} are independent to one another if and only if:



A *Bayesian Network* G is an:

* Annotated Directed Acyclic Graph (DAG) which represents a joint probability distribution over 𝑋
* The nodes of the graph correspond to the variables 𝑋1, ..., 𝑋𝑛
* The links of the graph correspond to the direct influence from one variable to the other
* If there is a directed link from variable 𝑋𝑖 to variable 𝑋𝑗, 𝑋𝑖 is a parent of 𝑋𝑗.
* Each node is annotated with a CPD that represents 𝑃(𝑋𝑖∣𝑃𝑎(𝑋𝑖)), where 𝑃𝑎(𝑋𝑖) denotes the parents of 𝑋𝑖 in G.

A Bayesian network G represents an unique joint probability distribution 𝑃(𝑋1, . . . , 𝑋𝑛) over 𝑋:

# 

# 

# BAYESIAN ASSOCIATION RULES MINING ALGORITHM

Bayesian Association Rule Mining (BAR Algorithm):

* **Input**: a dataset consisting of instances and attributes
* **Output**: association rules

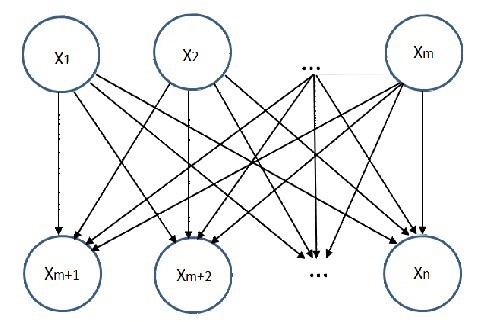
1. Discretize any continuous variables of the input dataset.
2. Generate association rules using Apriori
3. For each rule, construct a Bayesian network
4. Compute the **Bayesian confidence** (BC) and **Bayesian lift** (BL) of each rule.
5. Output those rules with the highest BC and BL

**Bayesian Confidence**

Let 𝐴 and 𝐵 be itemsets such that:

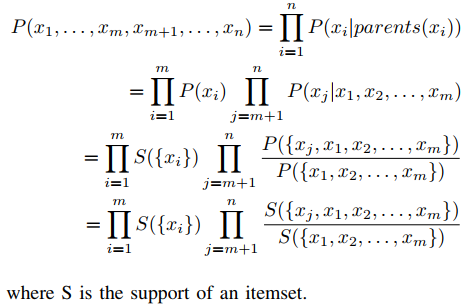
* 𝐴 consists of items 𝐼1, 𝐼2, . . . , 𝐼𝑚
* B consists of items 𝐼𝑚+1, 𝐼𝑚+2 . . . , 𝐼𝑛.

Then, an association rule 𝐴 ⇒ 𝐵 can be represented as the Bayesian network in the figure below, where {𝑥1, . . . , 𝑥𝑚} and 𝐵 are Boolean variables corresponding to{𝐼1, 𝐼2, . . . , 𝐼𝑚} and {𝑥𝑚+1, . . . , 𝑥𝑛} are Boolean variables corresponding to {𝐼𝑚+1, 𝐼𝑚+2 . . . , 𝐼𝑛}.

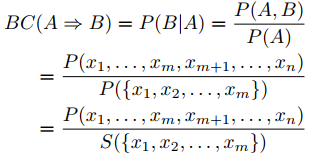


**Fig. 3: THE BAYESIAN NETWORK REPRESENTING THE RULE A → B**

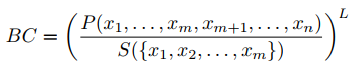
The joint probability distribution represented by the network is the following:



Bayesian confidence (BC) of A⇒B is defined as 𝑃(𝐵∣𝐴) computed using the Bayesian network representing A⇒B:

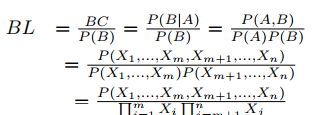


Short rules generalize better than long ones because the shorter rules match more instances than long rules; additionally, short rules are easier to interpret by humans than long ones. To penalize long rules, length L of rule is incorporated:



**Bayesian Lift**

Given a rule A⇒ B, the Bayesian lift (BL) is defined as BC/P(B), computed using the Bayesian network representing A⇒B:



BL is a value between 0 and infinity:

* BL = 1

⇒ P(B|A)/P(B) = 1

⇒ P(A,B)/[P(A).P(B)] = 1

⇒ P(A,B) = [P(A).P(B)]

⇒ A & B are in independent

* BL > 1

⇒ P(B|A)/P(B) > 1

⇒ P(A,B)/[P(A).P(B)] > 1

⇒ P(A,B) > [P(A).P(B)]

⇒ A and B are positively correlated

* BL < 1

⇒ P(B|A)/P(B) < 1

⇒ P(A,B)/[P(A).P(B)] < 1

⇒ P(A,B) < [P(A).P(B)]

⇒ A and B are negatively correlated

# 

# 

# 

# 

# 

# 

# DATA PREPROCESSING AND ANONYMIZATION

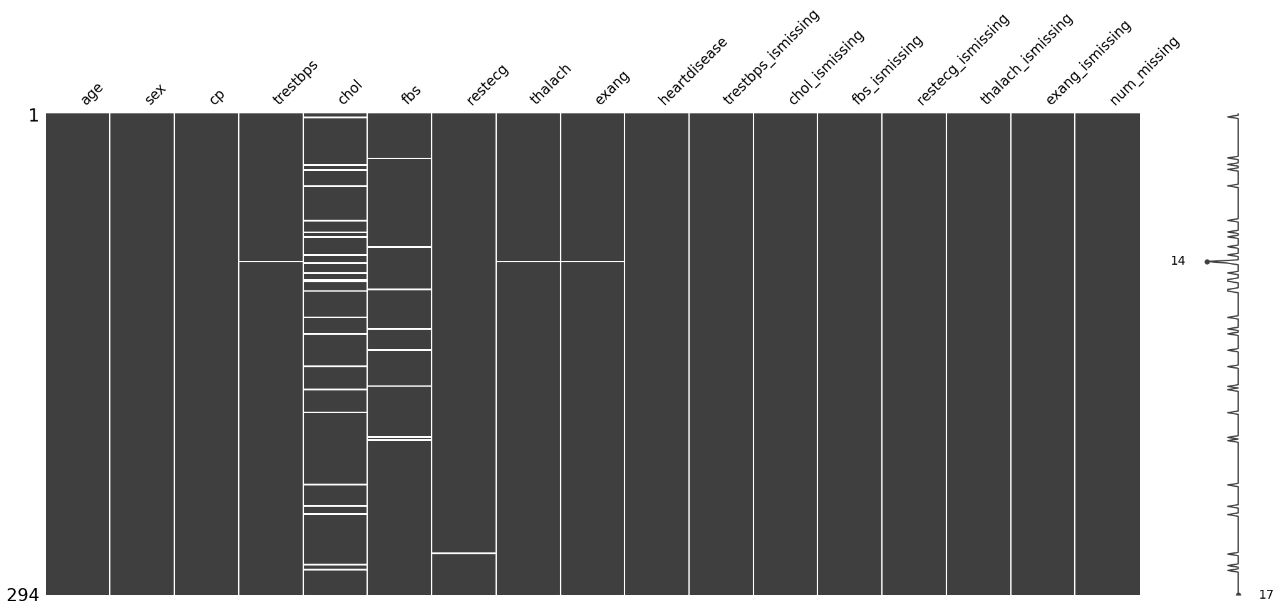
**Data Preprocessing**

Data Preprocessing in Machine Learning (ML) is the process of preparing the raw data before input to any ML model. It is a crucial step while building the model as we may not always retrieve clean and formatted data. Preprocessing the data using specific data transformation steps through pre-learning pipelines removes noises and missing values, thus increasing the accuracy and efficiency of an ML model.[3]

After retrieving Hungary’s Heart disease dataset[4], the attributes for which > 90% of the values were missing were dropped. The data rows with some missing attributes were also dropped while cleaning the data as their frequency was low. For encoding categorical data such as chest pain type attribute (cp) or resting electrocardiographic results (restecg), the required number of integer values were used.

**Visualization of the missing data (‘?’ to NaN)**

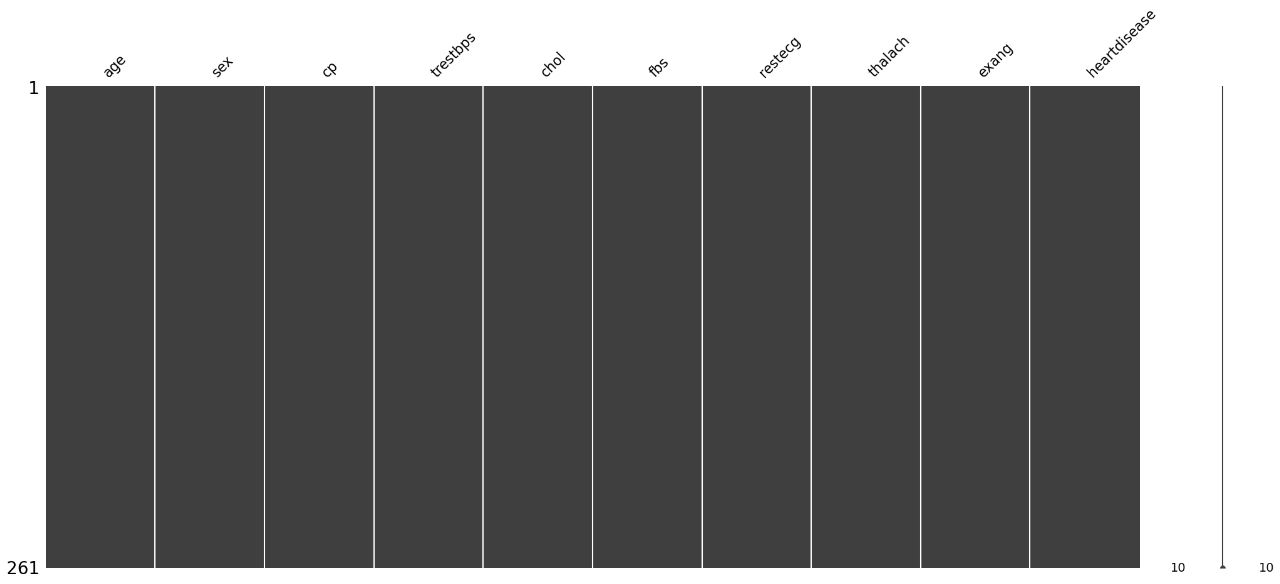
The missing values in the data are visualized by plotting a missing feature set graph along the test case values to analyze the discrepancies in the data as part of data cleaning before further analysis. (Note that processed Hungarian heart disease dataset has been used.)

****

**Fig. 4 : Missing values in the preprocessed dataset**

**Post removal of the NaN values**

It can be observed that there are no missing values and the data has been cleaned.

****

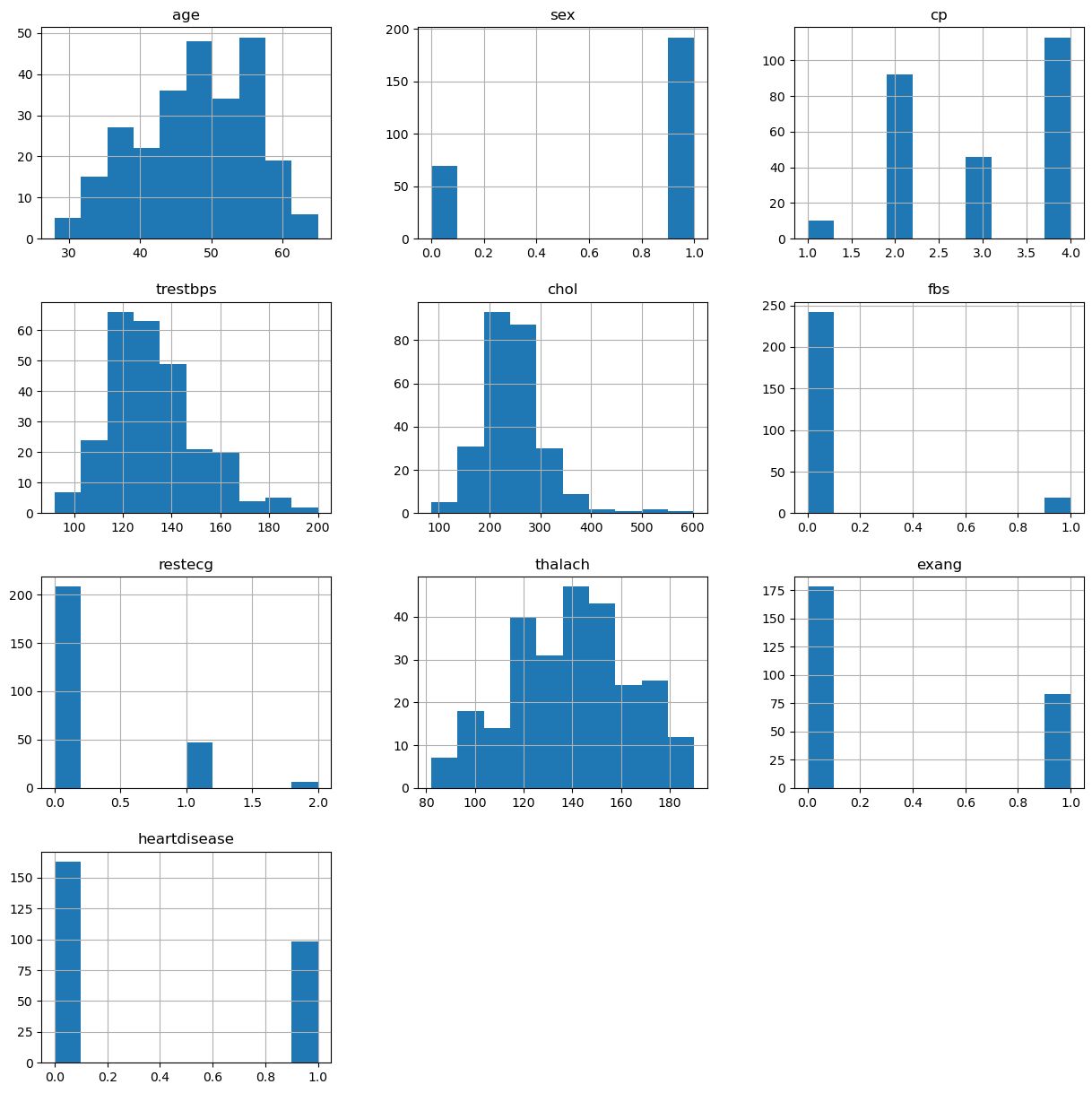
**Fig. 5 : Dataset post removal of the NaN values**

**Data Anonymization**

BAR has been applied to the Heart disease dataset (Hungary) from the UCI Repository. Data Anonymization is an essential part as subject identification from datasets strictly violates privacy laws. L2S data cube approach has been used for data anonymization before inputting to BAR. A subject can be identified by their unique properties in this approach, hence having multiple subjects with identical properties makes the subjects indistinguishable and un-identifiable from one another. Thus, third parties are granted access only to the anonymised data.[5]

The following four methods have been used to perform data anonymization:

1. **Data Discretization**: All continuous variables are represented by discrete values using integers beginning from 0. For example, age and trestbps (resting blood pressure). This is similar to binning the variable based on the pre-decided bins such as, if >5 and <10 then assign 7.
2. **Data Cube Creation**: Datacube is a standard pandas dataframe where the columns are attributes corresponding to dimensions and rows representing the combinations of values on the dimensions, as similarly represented by Figure 5.
3. **Count Perturbation**: In order to hide any insight into a person’s identity, noise (integers) is added to variables in a predefined set of integers e.g. {-1,1} to change the counts of values combinations.
4. **Cell Suppression (Data Anonymization)**: A count threshold has been set so that the cells with counts lower than the threshold value are replaced by value 0 in order to be excluded from further analyses.

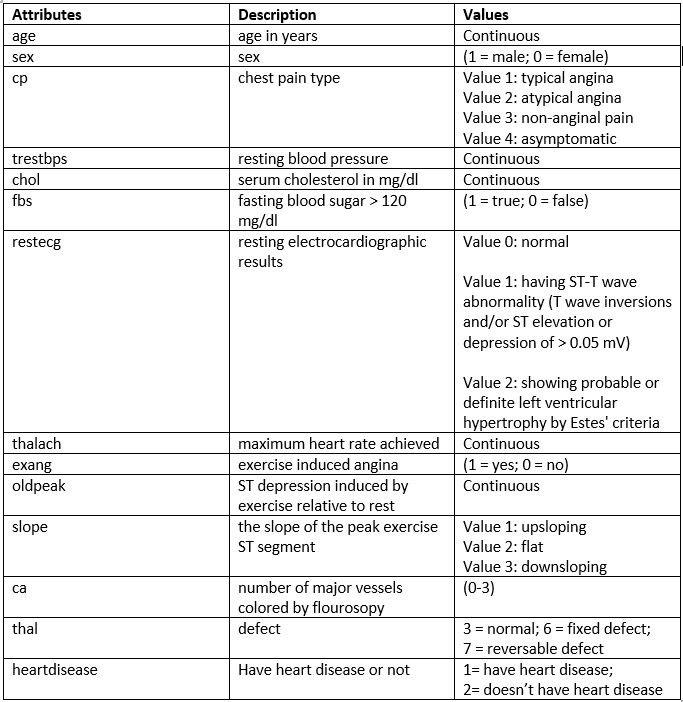
**Fig. 6 : Histograms to analyze the distribution of attributes v/s their frequency. These histograms are a 2-D representation of a data cube representing the attributes and its values**

# EVALUATION

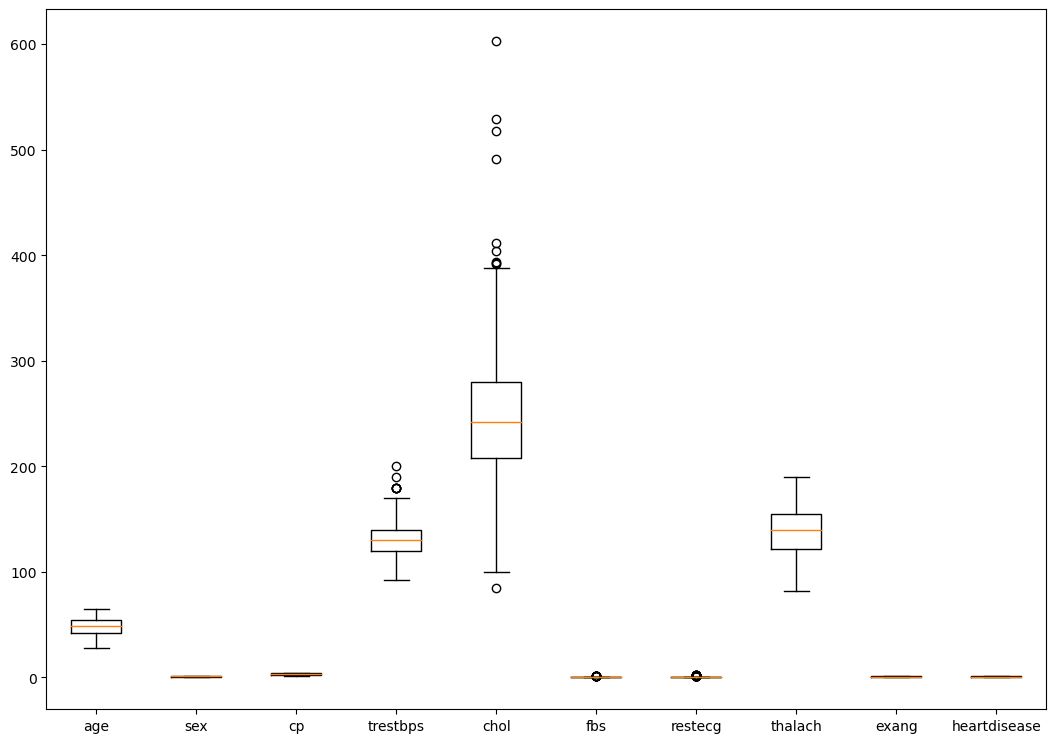
A data matrix is constructed from the anonymized data cube before input to BAR.

1. **Data Preparation**: The heart disease dataset for Hungary consists of 14 main attributes (Table 1) which further have various sub-classifications of attributes. For data discretization, many of the trivially continuous features have been expressed as integer values to demonstrate their severity. This includes, cp : Chest pain type, given an integer value from 1 to 4 based on the severity. In data mining, models obtained from balanced datasets output unbiased predictions and vice versa[6]. A balanced dataset was used with the same number of positive and negative instances of heart disease prediction to achieve class association rules with unbiased classification.

**Table 1 : Feature Set for Heart Disease Dataset, Hungary**



1. **Data Cleaning**: Inconsistent values of the attributes such as several missing values, earlier distinguished by value of -9.0 or character ‘?’, were removed. Outliers within the attribute values were plotted as shown in Figure 6 and were also treated as missing values. Post cleaning of data, the data cube was generated and association rules were mined from the data cube using BAR.



**Fig. 7 : Box-and-whisker plots of attributes of Heart Disease Dataset, Hungary: each box shows the distribution of values where the bottom line of the box represents the 1st quartile, the middle line represents the median and the top line of the box represents the 3rd quartile; the points outside the whiskers and extending from the box are the outliers.**

1. **Two-items Associations**: For the purpose of this report, two-item association rules have been mined and ranked in order of the Bayesian Confidence (BC) value. The values of support and Bayesian Lift (BL) have also been found. As shown in Table 2, the top 10 mined rules according to BC value also have the maximum BL values amongst all of the mined rules, indicating that the rules with highest confidence also have the highest lift. Thus, this is in agreement with the findings of the assigned paper for our project.

**Table 2 : The top 10 Two-item Association Rules from Heart Disease Dataset, Hungary ranked by Bayesian Confidence**

| **Rule** | **Support** | **Confidence** | **Lift** |
| --- | --- | --- | --- |
| sex → age | 0.735632 | 1 | 1 |
| cp → age | 1 | 1 | 1 |
| age → cp | 1 | 1 | 1 |
| trestbps → age | 1 | 1 | 1 |
| age → trestbps | 1 | 1 | 1 |
| chol → age | 1 | 1 | 1 |
| age → chol | 1 | 1 | 1 |
| fbs → age | 0.072797 | 1 | 1 |
| restecg → age | 0.203065 | 1 | 1 |
| thalach → age | 1 | 1 | 1 |

1. **Class Association Rules**: In the end, class association rules are mined from the 14-dimensional data cube using BAR. All of the top 10 class association rules have very high confidence, and the highest confidence amongst each of these mined rules is 0.56 value corresponding to the rule: exang = 1 => heart disease = 1.

# CONCLUSION

As can be observed, the best association rules are having the highest BC and BL values, as well as very high support, confidence and lift. The relationship of statistical conditional dependence and independence between the apriori association rules has been used as promising quality measures. This is justified as BC and BL use joint probability distributions of the association rules, in which a high joint probability value is indicative of strong conditional dependence and a lower value would indicate an independent relations between the items of the association rule.[5]  
  
Thus, this methodology can be used to mine the unknown associations between heart disease and patient information and symptoms which can help doctors and medical specialists take necessary actions well in time as well as provide them with additional support during their patient’s analysis.  
  
Attached below is the correlation graph (heatmap) between the heart disease attribute set depicting how different attributes have a *strong* effect on heart disease determination by analysing the correlation level between them. (Refer Appendix A for correlation pair plot.)

# 

**Fig. 8 : Correlation Heatmap of the Hungarian heart disease attribute set**

\*\*\*\*\*\*\*

# APPENDIX

# Appendix A

# 

**Fig. 9 : Pairplots of the selected attribute (in row) cross all other attributes (columns) taking in a unique attribute in each from the attribute set used**

# REFERENCES

1. Code Sources:-

* Apriori Algorithm (AprioriAlgo.py): JavaPoint. Implementing Apriori Algorithm in Python

(<https://www.javatpoint.com/implementing-apriori-algorithm-in-python>)

* Bayesian Networks (Bayesian.py): Dr. Thyagaraju G.S. Python Bayesian Network.

(<https://github.com/profthyagu/Python-Bayesian-Network>)

1. Gadewadikar, Jyotirmay & Kuljaca, Ognjen & Agyepong, Kwabena & Sarigul, Erol & Zheng, Yufeng & Zhang, Ping. (2010). Exploring Bayesian Networks for medical decision support in breast cancer detection. African Journal of Mathematics and Computer Science Research. 3.
2. C. V. Gonzalez Zelaya, "Towards Explaining the Effects of Data Preprocessing on Machine Learning," 2019 IEEE 35th International Conference on Data Engineering (ICDE), 2019, pp. 2086-2090, doi: 10.1109/ICDE.2019.00245.
3. Dataset used: Hungarian Heart Disease Dataset, UCI Repository.

(<https://archive.ics.uci.edu/ml/datasets/Heart+Disease>)

1. D. Tian et al., "A Bayesian Association Rule Mining Algorithm," 2013 IEEE International Conference on Systems, Man, and Cybernetics, 2013, pp. 3258-3264, doi: 10.1109/SMC.2013.555.
2. J. Han and M. Kamber, Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann, 2011